nature neuroscience

Article

https://doi.org/10.1038/s41593-022-01225-z

Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning

Received: 21 September 2021

Accepted: 28 October 2022

Published online: 12 January 2023

Check for updates

Shanshan Qin[1,2], Shiva Farashahi[3,6], David Lipshutz[3,6], Anirvan M. Sengupta[3,4], Dmitri B. Chklovskii[3,5] & Cengiz Pehlevan[1,2]

Recent experiments have revealed that neural population codes in many brain areas continuously change even when animals have fully learned and stably perform their tasks. This representational 'drift' naturally leads to questions about its causes, dynamics and functions. Here we explore the hypothesis that neural representations optimize a representational objective with a degenerate solution space, and noisy synaptic updates drive the network to explore this (near-)optimal space causing representational drift. We illustrate this idea and explore its consequences in simple, biologically plausible Hebbian/anti-Hebbian network models of representation learning. We find that the drifting receptive fields of individual neurons can be characterized by a coordinated random walk, with effective diffusion constants depending on various parameters such as learning rate, noise amplitude and input statistics. Despite such drift, the representational similarity of population codes is stable over time. Our model recapitulates experimental observations in the hippocampus and posterior parietal cortex and makes testable predictions that can be probed in future experiments.

Memories and learned behavior can be stable for a long time. We can recall vividly the memory of events that happened years ago. Motor skills, such as riding a bike, once learned, can last life-long even without further practice. Learning and memory lead to structural changes in the brain's neural networks establishing associations between external stimuli and internal neural population activities or representations. A natural question is then whether stable task performance and memories are related to stable neural representations.

Recent technical advances in electrophysiology and optical imaging have enabled researchers to address this question by studying the long-term dynamics of neural population activity in awake-behaving animals[1–6]. A number of these experiments have shown that neural activities in cortical areas that are essential for specific tasks undergo continuous reorganization even after the animals have fully learned their tasks, a phenomenon termed 'representational drift'[7,8]. For instance, in sensorimotor tasks, neural representations in the posterior parietal cortex (PPC) of mice change across days while the performance of animals remain stable and high[9]. Place fields of individual place cells in the CA1 region of the hippocampus drift over days and weeks even when the animals remain in the same familiar environment[1,10,11]. Individual neurons in the primary motor cortex and supplementary motor cortex show unstable tuning while animals perform highly stereotyped motor tasks[12] (but see refs. 13–15). The neural ensemble representation of conditional stimuli in mouse amygdala shows continuous change over days after fear conditioning[16]. Representational drift has been observed even in primary sensory cortices, such as the

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]Center for Brain Science, Harvard University, Cambridge, MA, USA. [3]Center for Computational Neuroscience, Flatiron Institute, New York, NY, USA. [4]Department of Physics and Astronomy, Rutgers University, New Brunswick, NJ, USA. [5]NYU Langone Medical Center, New York, NY, USA. [6]These authors contributed equally: Shiva Farashahi, David Lipshutz. e-mail: cpehlevan@seas.harvard.edu

https://doi.org/10.1038/s41593-022-01225-z

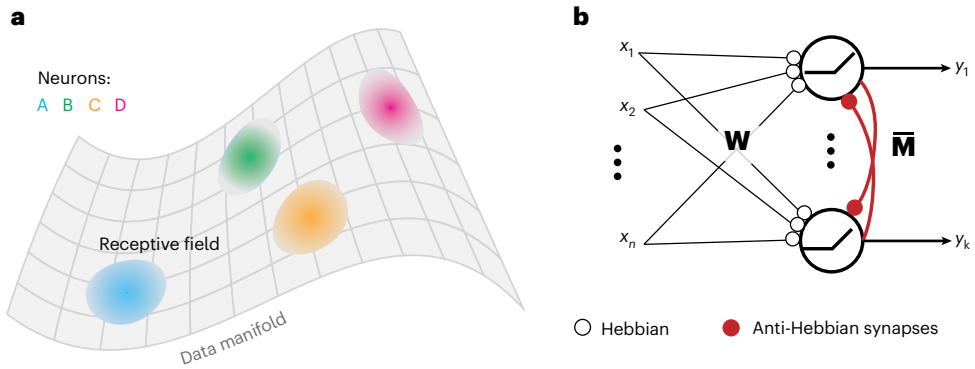


Fig. 1 | Learning localized RFs in Hebbian/anti-Hebbian networks. a, Illustration of localized RFs that tile the data manifold. **b**, A Hebbian/anti-Hebbian network with non-negative neural activity can learn localized RFs.

mouse visual cortex[17,18] and piriform cortex[3]. The ubiquity of representational drift raises several important questions: what are the underlying mechanisms of such drift? How can neural circuits generate stable coding in the presence of continuous drift? What are the dynamics of representational drift?

We put forward that representational drift can be accounted by a learning process with noisy synaptic dynamics and a degeneracy of possible learned representations. Indeed, synapses in the cortex are highly dynamic and have a relatively short lifetime with respect to task memories[19–21]. Thus, synaptic configurations in the brain might continuously evolve even if task performance is stable. Representational degeneracy can arise when learning representations of sensory inputs by optimizing a representational objective. Many different normative accounts of sensory representations have been proposed in neuroscience[22–30]. If the representational objective has many solutions, meaning there are many optimal neural representations of the input stimuli, noisy synaptic updates during learning will drive the network to explore the synaptic weight space that corresponds to (near-)optimal neural representations. In other words, the neural representation will drift within the space of optimal representations.

We illustrate this idea and explore its consequences by focusing on representational drift in brain areas where neurons have localized receptive fields (RFs), such as the hippocampus and PPC[1,9,10,31,32]. In these systems, populations of neurons with 'bump' RFs tile the parameter space they encode (Fig. 1a); however, these bumps may move, or drift, on the parameter space over time[1,9,10]. By tracking the dynamics of these bumps over their respective parameter spaces, recent experimental studies measured and quantified the drift of localized RFs in hippocampal place cells[1,10] and PPC[9] and provided datasets that allow testing of our hypothesis.

To further constrain our hypothesis and generate testable predictions for drift of localized RFs in such datasets, we introduce a minimal neural network model that exhibits all the mentioned ingredients—learning, degeneracy and noisy synaptic dynamics—for representational drift. Localized RFs can be modeled by a well-studied class of biologically plausible networks for representation learning: Hebbian/anti-Hebbian networks[33,34] (Fig. 1b). These networks optimize similarity-matching objectives which exhibit a degeneracy of optimal representational solutions, all of which share the same representational similarity matrix[34–36]. As we will show, representational similarity is also preserved in experiments we consider, providing further motivation for our choice. Through mathematical analysis of the optimal solutions of certain similarity-matching objectives, Hebbian/anti-Hebbian networks with rectifying nonlinearities have been shown to learn population codes that tile the domain of latent variables of input data in a way that individual neurons have local 'bump' tuning curves[37]. Hence, these networks satisfy the learning and representational degeneracy criteria of our hypothesis. To complete the ingredients, in this paper, we introduced noisy synaptic updates to Hebbian/anti-Hebbian networks and explored the properties of the representational drift that arises. We then tested the predictions of our model qualitatively and quantitatively against experimental data, comparing to relevant null models, including an independent random walk model of representational drift.

By numerical and analytical methods, we find that while the RFs of individual neurons change significantly over time, the representational similarity of population codes remains stable. We show that the drift dynamics of individual RFs can be largely captured by a random walk on the input data manifold, with the effective diffusion constant depending on noise amplitude, learning rate and other model parameters. However, at the population level, the drifting RFs are coordinated—unlike a null model where RFs perform independent random walks from each other on the data manifold—in a way that preserves representational similarity. We verify this key prediction by analyzing experimental data. Our model accounts for many other recent experimental observations in the hippocampus and PPC and makes further testable predictions.

Overall, our results show how optimal representation learning and noise can lead to representational drift while maintaining representational similarity. While our modeling effort focuses on specific models and brain areas, our observation that drift of RFs is coordinated may be a general property of noisy representation learning with degenerate optimal network configurations.

Results

Drift dynamics in linear Hebbian/anti-Hebbian networks

We first study drift in linear Hebbian/anti-Hebbian networks, which compress inputs into a lower dimensional principal subspace[29]. While the resulting RFs are not localized, it is still instructive to study how learned representations evolve with noisy synaptic updates in this analytically tractable model. The insights we build will carry over to nonlinear network models.

The network we consider minimizes a similarity-matching cost function for its learned representations[29]. Here, the similarity between two vectors is defined as their dot product. Let $\mathbf{x}_t \in \mathbb{R}^n, t = 1, \cdots, T$ be a set of network inputs (or sensory stimuli) and $\mathbf{y}_t \in \mathbb{R}^k, k < n$ be the corresponding outputs constituting a neural representation, that is, firing rate vector. Similarity matching minimizes the mismatch between the similarity of pairs of inputs and corresponding pairs of outputs

$$\min_{\forall t \in \{1,\ldots,T\}:\mathbf{y}_t} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\mathbf{x}_t^\top \mathbf{x}_{t'} - \mathbf{y}_t^\top \mathbf{y}_{t'}\right)^2. \quad (1)$$

https://doi.org/10.1038/s41593-022-01225-z

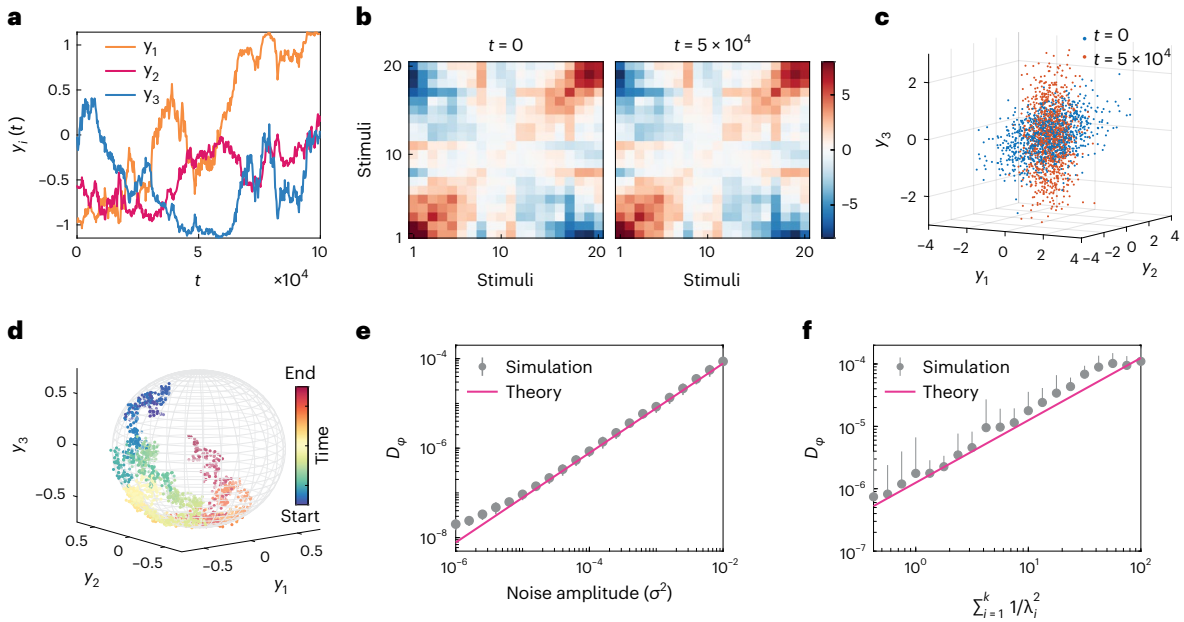


Fig. 2 | Drift dynamics in the PSP task. a–d, We present a simulation where each input $\mathbf{x} \in \mathbb{R}^{10}$ is drawn independently from a correlated Gaussian distribution $N(\mathbf{0}, \mathbf{C})$. The first three eigenvalues of the covariance matrix are 4.5, 3.5, 1 and the rest are 0.01. Eigenvectors are random orthogonal vectors. A Hebbian/anti-Hebbian network learns to project this input to a subspace with $k = 3$ dimensions. **a**, The learned representation of an example input continuously changes due to noisy updates. Shown are the three components of the output to the same input through learning, denoted by $\mathbf{y}(t)$. **b**, Pairwise dot-product similarities between the learned representations are stable over time, as shown by the almost identical similarity matrices at $t = 0$ (left) and $t = 5 \times 10^4$ (right). To calculate these similarity matrices, we froze the weights at $t = 0$ and $t = 5 \times 10^4$ and ran the network for each input to find the corresponding output. **c**, The ensemble of outputs $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ at two distinct time points, calculated as in **b**. The data clouds have an ellipsoid shape. **d**, Drifting representation as a random walk on a sphere. We show the representation of a single sample $\mathbf{y}_t$ over time. Color codes different time steps. **e**, Relationship between D_φ and noise amplitude σ^2 (mean ± s.d., $n = 40$ independent simulations). Symbols with error bars denote numerical simulations, and the solid line is our theory (Eq. 4). **f**, Dependence of D_φ on the eigenspectrum $\{\lambda_i\}$ of the input covariance matrix (mean ± s.d., $n = 2{,}000$ simulations are aggerated into 20 bins based on their $\sum_{i=1}^{k} 1/\lambda_i^2$ in the log scale). In **e**,**f**, only upper error bars are shown to avoid cluttering. In all the figures, $t = 0$ marks a starting point after the representation is learned. See Supplementary Note Section 5 for details of simulations in **e**,**f**.

Importantly, there is a continuum of equally optimal solutions to this problem which are given by the infinitely different ways of projecting the inputs to their principal subspace[29]. This degeneracy can be seen from the rotational symmetry of the cost function, Eq. (1). For any set of outputs $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, the sets $\{\mathbf{R}\mathbf{y}_1, \cdots, \mathbf{R}\mathbf{y}_T\}$ have the same cost, where $\mathbf{R}$ is an orthogonal matrix.

Previous work showed that optimal representations for this cost function can be learned by a neural network in an online manner, where each input $\mathbf{x}_t$ is presented sequentially and an output $\mathbf{y}_t$ is produced by running the following neural dynamics until convergence[29] (Methods):

$$\dot{\mathbf{y}}_t = \mathbf{W}\mathbf{x}_t - \mathbf{M}\mathbf{y}_t. \tag{2}$$

Here, $\mathbf{W}$ holds the feedforward synaptic weights and $\mathbf{M}$ the lateral synaptic weights.

After each presentation of an input and convergence of the neural dynamics, the weights $\mathbf{W}$ and $\mathbf{M}$ are updated with a Hebbian and an anti-Hebbian rule, respectively:

$$\Delta\mathbf{W} = \eta\left(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}\right), \quad \Delta\mathbf{M} = \eta\left(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}\right). \tag{3}$$

These learning rules are local in the sense that synaptic updates only depend on activities of presynaptic and postsynaptic neurons. The update of $\mathbf{M}$ is anti-Hebbian due to the negation in Eq. (2). These weight updates constitute a gradient-based optimization algorithm for the similarity-matching cost function given in Eq. (1)[29,36]. As the number of inputs increases, these weights converge to a configuration where the network outputs are projections of the input to a principal subspace, minimizing the similarity-matching cost function[29,36].

We now turn to representational drift under noisy synaptic plasticity. We introduce an additive, independent and identically distributed Gaussian noise for each synaptic update (Eq. 12 in Methods). The network still learns the principal subspace and maintains a stable performance in its ability to project its inputs to their principal subspace (Extended Data Fig. 1a,b). However, due to the synaptic noise, network weights do not settle down to fixed points but roam around in the subspace that gives equally good solutions to the similarity-matching problem. Consequently, the representation of a given stimulus $\mathbf{y}_t$ drifts over time (Fig. 2a). However, the similarity between any two outputs $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ remains stable which we show by plotting the output similarity matrix $\mathbf{Y}^\top\mathbf{Y}$, where $\mathbf{Y} \equiv [\mathbf{y}_1, \cdots, \mathbf{y}_T]$, at two different times (Fig. 2b and Extended Data Fig. 1c). The drift of the representation ensemble $\{\mathbf{y}_1, \cdots, \mathbf{y}_T\}$ behaves like a randomly rotating rigid body consisting of a cloud of points (Fig. 2c). Such drift does not change the length of each output vector, $\mathbf{y}_t$, which undergoes a random walk on a spherical surface (Fig. 2d).

These observations motivate us to quantify the drift speed by a rotational diffusion constant D_φ (Extended Data Fig. 1d)[38,39]. We can derive an approximate analytical formula for D_φ from a linear stability analysis (Methods and Supplementary Note Section 1):

$$D_\varphi \approx \frac{1}{4}\eta\sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i^2}, \tag{4}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ are the top k ordered eigenvalues of the input covariance matrix. Equation (4) indicates that D_φ is proportional to the noise amplitude. Further, the drift amplitude along each eigenvector is proportional to $1/\lambda_i^2$. This is analogous to the rotation of an ellipsoid

https://doi.org/10.1038/s41593-022-01225-z

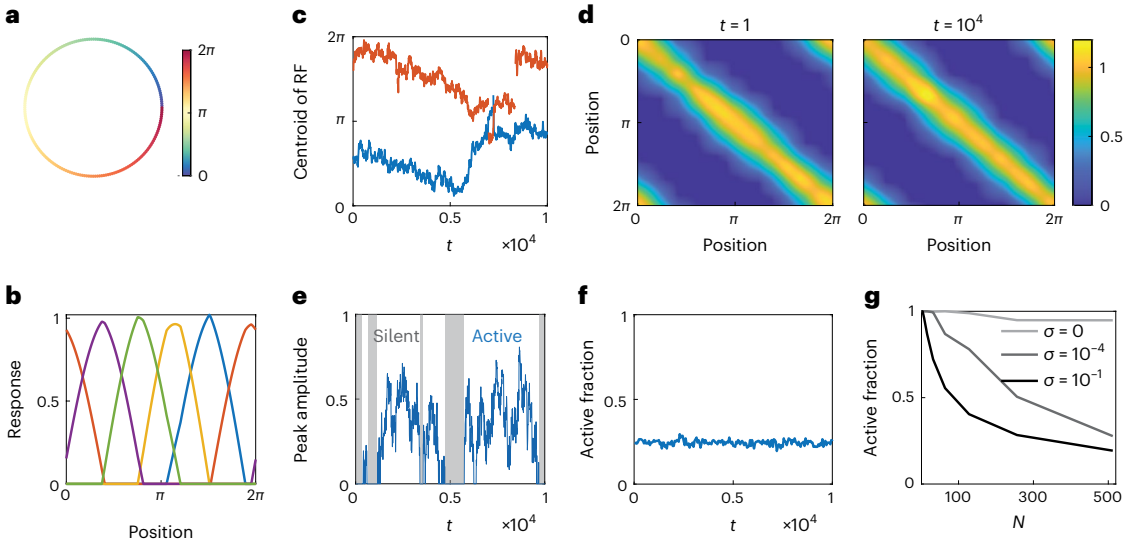


Fig. 3 | Drift of manifold-tiling localized RFs in nonlinear Hebbian/anti-Hebbian networks. a, A ring in two-dimensions as input dataset: $\mathbf{x}(\theta) = [\cos(\theta), \sin(\theta)]^\top, \theta \in [0,2\pi)$. **b**, Learned localized RFs tile the input ring data manifold. Colors represent RFs of five example neurons. **c**, Evolution of the RF centroids of two example neurons due to synaptic noise. **d**, The representational similarity matrix $\mathbf{Y}^\top\mathbf{Y}$ is approximately circulant and stable over time. **e**, When there are a large number of neurons, each neuron has active and silent (shaded region) periods. **f**, At the population level, the fraction of neurons with active RFs are constant. **g**, The fraction of neurons that have active RFs decreases with the total number of output neurons, as well as the noise amplitude.

rigid body under torque, where rotations around axes with smaller moment of inertia are faster. Predictions of Eq. (4) agree well with simulations (Fig. 2e,f).

The above simulations and analyses demonstrate that learned representations drift over time in linear Hebbian/anti-Hebbian networks, while the similarity of representations is preserved. This is due to a coordinated random walk in the representational space in the form of a rigid body rotation. This coordinated drift explores equally optimal representations.

Next, we consider nonlinear networks and show that these results carry over.

Drift dynamics in nonlinear Hebbian/anti-Hebbian networks

RFs of neurons in many brain areas are localized in the parameter space which they represent. For example, simple cells in the primary visual cortex (V1) are tuned to orientations of gratings[40]. Neurons in the owl's external nucleus of the inferior colliculus (ICX) are tuned to different horizontal and vertical positions, forming an auditory spatial map[41]. Place cells in the hippocampus are active when an animal is at a particular location of an environment[31].

If a rectifying neuronal nonlinearity is introduced, our Hebbian/anti-Hebbian network can capture these localized RF properties (Fig. 1b; Methods). This network minimizes a non-negative similarity-matching (NSM) cost function:[29,33,37]

$$\min_{\forall t\in\{1,\ldots,T\}:\mathbf{y}_t\geq 0} \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(\mathbf{x}_t^\top\mathbf{x}_{t'} - \mathbf{y}_t^\top\mathbf{y}_{t'} - \alpha^2\right)^2, \tag{5}$$

where $\mathbf{x}_t$ is the input and $\mathbf{y}_t$ in the non-negative output firing rate vector, and α^2 sets the threshold of similarity to be preserved in the output representation. The number of output neurons can be larger than, equal to, or smaller than the input dimensions. With non-negative neural activity, the above NSM objective function strives to preserve similarity for similar pairs of input samples but orthogonalizes the outputs corresponding to dissimilar input pairs. Compared with the linear case, non-negativity breaks the rotational symmetry of the solution, but the permutation symmetry is still preserved, that is, exchanging identities of neurons does not change the objective function. One can further control the behavior of learned representations by introducing regularizers to $\mathbf{y}_t$ in Eq. (5), for example, an l_1 norm of $\mathbf{y}_t$ leads to a sparser representation (Methods).

Like the previous linear Hebbian/anti-Hebbian network, this network also operates in an online fashion with a similar local learning rule (Methods). Again, we focus on how the learned representations evolve in the presence of noise in synaptic updates. Specifically, we first consider stimuli living on a ring (Fig. 3a), like the direction of a drifting grating used in experimental studies of visual systems. The location of the stimulus on the ring is parameterized by an angular variable $\theta \in [0,2\pi)$ (Methods and Fig. 3a).

In the case of a single output neuron and without synaptic noise, the learned RF can be shown to be a truncated cosine curve centered around a random angle. Derivation of this result is given in Methods and Supplementary Note Section 2. With synaptic noise during learning, the centroid of the RF drifts on the ring like a random walk (Extended Data Fig. 2a). We quantified the speed of drift with an effective diffusion constant, D, on the ring (Methods and Supplementary Note Section 2). For a single neuron, when $\alpha = 0$, the dependence of D on the learning rate η and noise amplitude σ^2 can be analytically approximated as $D \approx \eta^2/2 + 8\eta\sigma^2$ (Methods and Supplementary Note Section 2). The first term is due to the sampling noise, that is, the fact that the network sees one random stimulus at a time, and the second term is due to the explicit synaptic noise. This result indicates that faster learning and larger synaptic noise lead to more rapid drift of the RF as verified by our numerical simulations (Extended Data Fig. 2b,c).

When there is a population of output neurons, the nonlinear Hebbian/anti-Hebbian network learns multiple localized RFs that tile the ring manifold with overlaps (Fig. 3b), consistent with previous analytical accounts of simplified versions of such networks[37]. With synaptic noise, we expect each RF to drift by a similar diffusion process as in the single neuron case, but with interactions between the neurons affecting the dynamics (Fig. 3c). In particular, the structure of neural population activity, as indicated by output similarity matrix, is stable across time (Fig. 3d). Further, a neuron's response to the same stimulus is intermittent, having active and silent periods (Fig. 3e). At the population level,

https://doi.org/10.1038/s41593-022-01225-z

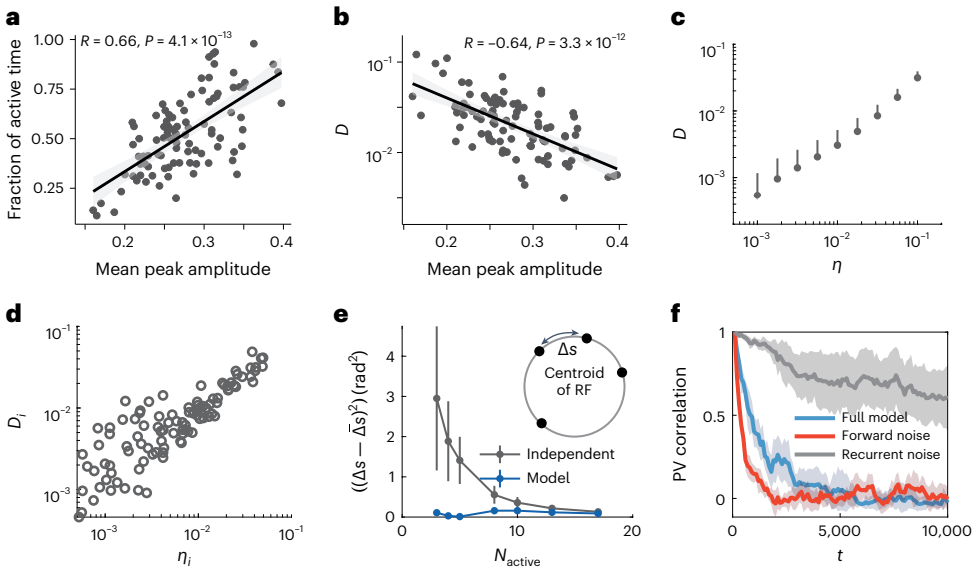


Fig. 4 | Predictions of the nonlinear model. a,**b**, Neurons with stronger RFs have longer active time (**a**) and are more stable as quantified by smaller effective diffusion constants D (**b**) (Pearson's correlation, two sided Student's t-test; shaded region is the 95% confidence interval). **c**, The average diffusion constant D of different RFs is proportional to the homogeneous learning rate η (mean ± s.d., $n = 40$ independent simulations; only upper error bars are shown to reduce cluttering). **d**, When each output neuron has its own learning rate of η_i drawn uniformly in the log scale, diffusion constant D_i of individual RF also increases with η_i. **e**, At the population level, the drift of RFs is coordinated such that their centroids are more uniformly distributed over the ring compared to that of a collection of independent random walkers. Shown are the variances of distances between adjacent centroids (mean ± s.d., $n = 40$ simulations). **f**, Synaptic noise in recurrent connections (gray) has a smaller influence on representational drift compared to synaptic noise in feedforward connections (red) as indicated by the much slower decay of the average autocorrelation coefficient of population vectors (Methods) (±s.d, $n = 200$ input angles).

the fraction of neurons that have active RFs at any given time is constant (Fig. 3f), and it decreases with total number of output neurons as well as the noise amplitude (Fig. 3g). Thus, in a large population of neurons, only a small fraction of them will be active at a given time, forming a sparse population code.

Different neurons show different levels of drift. We observed that neurons with stronger tuning (characterized by the peak amplitude of the RF) tend to be active more often (Fig. 4a) but have slower drift (Fig. 4b). Faster learning rate is correlated with more rapid drift for both homogeneous learning rates and heterogeneous learning rates across synapses in the population (Fig. 4c,d). These observations are consistent with the single neuron case (Extended Data Fig. 2d).

As in the linear case, we find that the drift of RFs is coordinated at the population level. To illustrate this point, we compared our model to a null model of independent random walkers: we simulated N_{active} RFs undergoing independent random walks on the ring. The step size of the independent random walks was drawn from the same distribution as the distribution of RF centroid shifts between two adjacent time steps obtained from our model (Methods). We observed that centroids of RFs in our model tile the ring manifold more uniformly than those of independent random walkers, as indicated by the smaller variance of distances between two adjacent centroids on the ring (Fig. 4e).

We also explored how different sources of synaptic noise in the network contribute to the representational drift. To this end, we examined the drift speed of population responses to stimuli (population vector or PV) by including synaptic noise either in the forward connections or in the recurrent connections. We found that the feedforward synaptic noise has much stronger influence on drift than the recurrent synaptic noise as indicated by a faster decay and lower asymptotic PV correlation coefficient (Fig. 4f and Extended Data Fig. 3).

Having gained better understanding of drifting dynamics in the above simple model, we now discuss models of representational drift in the hippocampus CA1 region and PPC. The observations made in Figs. 3 and 4 will conceptually carry over, providing explanations for previous experimental observations.

A Hebbian/anti-Hebbian Network model for drifting place fields in the hippocampal CA1 region

CA1 place cells in the hippocampus play a crucial role in spatial memory and navigation. Recent long-term recording experiments show that place coding by the population of CA1 pyramidal cells are dynamic even when the animal is in the same familiar environment. In a time course of several weeks, some neurons lose their place fields while other previously non-place coding cells gain place fields. Despite the drift, the spatial information is preserved[1,10,11].

One possible mechanism of place field formation is that hippocampal CA1 place cells receive both direct external input from grid cells in the entorhinal cortex and lateral competition from other place cells via local inhibitory interneurons[42–45]. Synapses between these neurons are plastic[46,47]. This motivated us to use the Hebbian/anti-Hebbian network to learn a place cell representation of a two-dimensional square environment. Each position on the plane is represented by the firing rate vector of a population of grid cells with different grid spacing, phases and offsets (Fig. 5a; Methods), which serves as the input $\mathbf{x}_t$ to the network. For simplicity, we modeled the inhibitory-neuron-mediated feedback inhibition as mutual inhibition between place cells (Fig. 5a). A more detailed network with inhibitory neurons generates quantitatively very similar results (Supplementary Note Section 3 and Extended Data Fig. 4).

After learning, many output neurons develop localized RFs (or place fields, Fig. 5b). This can be visualized by arranging each row of response matrix $\mathbf{Y}$ into a square matrix (Fig. 5b, left). These active place cells tile the two-dimensional environment, as indicated by the uniform distribution of centroids of place fields (Fig. 5b, right). Due to the noise in the weight update, place fields continuously drift over time (Fig. 5c). Despite the drift, representational similarity of positions in the two-dimensional environment is stable (Extended Data Fig. 5a).

https://doi.org/10.1038/s41593-022-01225-z

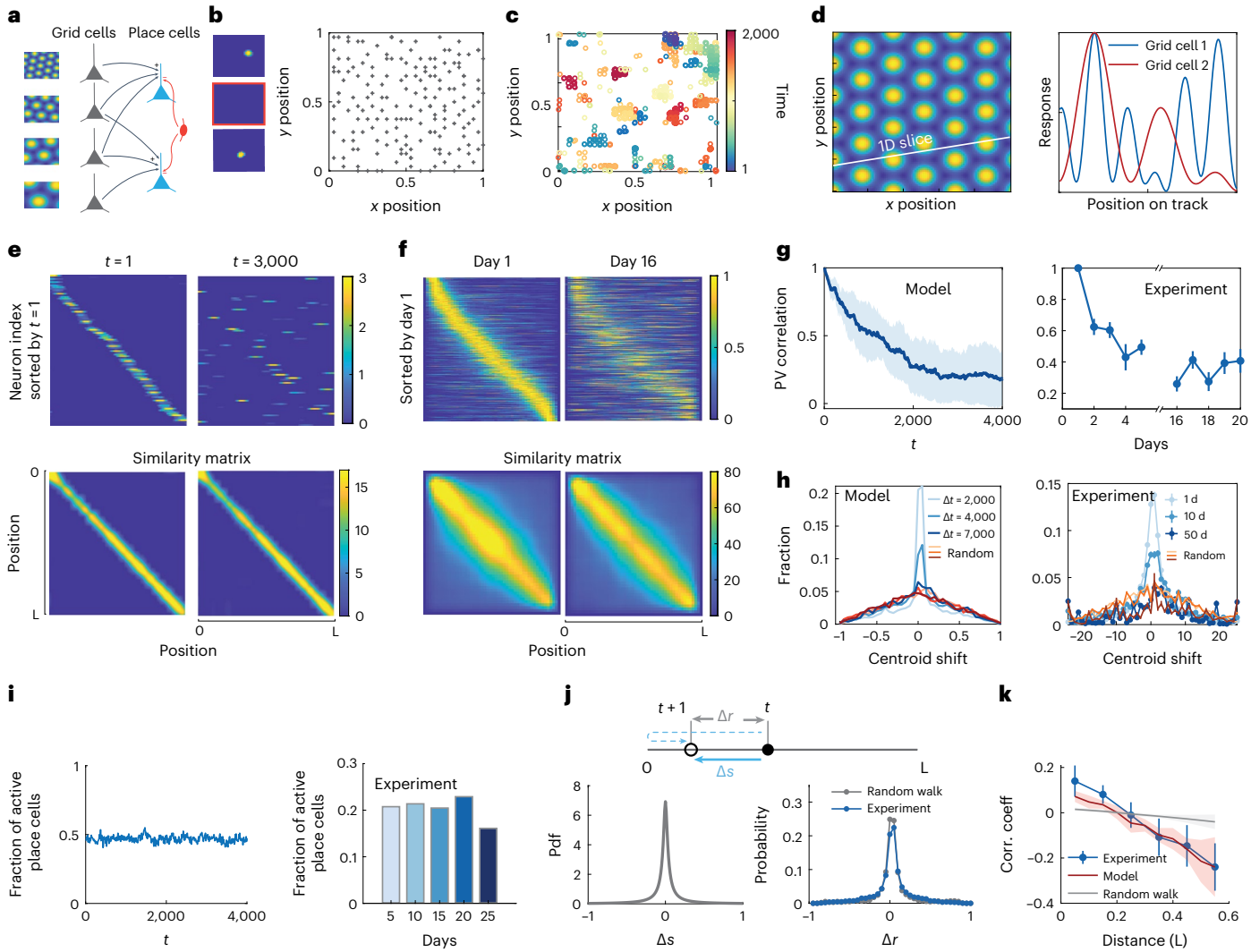


Fig. 5 | Drift of place fields. a, Place cells receive input from different grid cells and compete via local inhibitory neurons. **b**, Left: learned place fields tile the entire two-dimensional plane with red square highlighting a silent neuron. Right: each dot represents a centroid of a place field. **c**, Drift of place field of an exemplar place cell over time. **d**, Left: slice through a two-dimensional grid field. Right: responses of neurons across this slice. **e**, Upper: learned place fields tile a one-dimensional linear track when sorted by their centroid positions (left), but continuously change over time (right). Lower: representational similarity matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ of position is stable over time. **f**, Experimental results corresponding to (**e**). Place fields of a group of CA1 place cells pooled from several mice when exploring the same familiar one-dimensional linear track. **g**, The average autocorrelation coefficient of population vectors representing each spatial position in the model (left, shading, mean ± s.d.; $n = 200$ input positions) and experiment (right, mean ± s.d.; $n = 48$) decay over time. **h**, Probability distribution of centroid drifts of place fields at three different time intervals. Random distributions of centroid shifts are obtained by randomly shuffling the place fields of neurons at the end of the time interval and calculating the centroid differences. Left: model, right: experimental result. **i**, The fraction of active place fields is stable over time in the model (left) and experiment (right). **j**, Two different random walks (blue lines) under reflecting boundary conditions (upper panel). To make a fair comparison with an independent random walk, we sample step sizes Δs of the random walk from a distribution $p(\Delta s)$ (lower left panel) that produces a similar centroid shift distribution to the experiment $p(\Delta r)$ (lower right). **k**, Drifts of RFs show distance-dependent correlations, quantified by average Pearson's correlation coefficients. The model (mean ± s.d., $n = 20$ repeats) can recapitulate the behavior observed in the experiment (mean ± s.d., $n = 13$ animals), while correlation coefficient in a random walk null model is always around 0 (shading, mean ± s.d.; $n = 20$ repeats). Experimental results in **f**–**k** are plotted using data from[10].

We also observed that a place cell may lose its place field for some time and gain a new place field later on (Extended Data Fig. 5b). The intermittence of RFs is due to both the competition between RFs and the fluctuations of synaptic weights. For example, once the forward input to a neuron is smaller than the lateral inhibition at the centroid of an RF, it becomes silent. The durations of these silent periods follow approximately an exponential distribution (Extended Data Fig. 5c), suggesting that the transition between active and silent state of RFs is random and memoryless. However, the fraction of neurons with active place fields at any given time remains constant (Extended Data Fig. 5d).

The drift speed can again be quantified by an effective diffusion constant D, measuring how place field centroids drift over the arena. The dependence of D on the number of neurons N is nonmonotonic (Extended Data Fig. 5e). Neurons whose RFs have stronger tuning (larger peak amplitude of the RF) tend to be active more often and have smaller drift (Extended Data Fig. 5f,g).

While the above predictions could be compared with long-term recording experiments for animals in a two-dimensional environment, existing long-term recording experiments are limited to one-dimensional environments (typically linear tracks). To compare

our model with these experimental results, we simulated our model in a one-dimensional environment, where grid cell responses are modeled as one-dimensional slices through the two-dimensional grid fields, as observed in experiments[48] (Fig. 5d; Methods). The model generates qualitatively similar results as the above two-dimensional place cell model. The learned place fields tile the linear track but drift over time due to ongoing noisy weight updates, yet the representational similarity is stable over time (Fig. 5e). This is also observed in an experiment[10], where CA1 pyramidal cells were recorded when mice were in the same familiar environment for several months (Fig. 5f). Due to drifting place fields, the autocorrelation coefficients of neural population vectors in both our model and the experiment decay over time (Fig. 5g). The shift of centroids of place fields increases with time, with a distribution eventually approaching the case wherein the place fields are randomly permuted. Such behavior closely resembles that of the experiment[10] (Fig. 5h). Despite the continuous reconfiguration of the neural assemblies representing the position, the fraction of active place cells is stable over time in both our model and the experiment (Fig. 5i).

To further explore the underlying structure of centroid shifts and test the main prediction of our model that the drift of RFs is coordinated, we set out to compare the experiment and our simulation results to a null hypothesis: the shifts of RFs behave like independent random walks. To make a fair comparison, for the null hypothesis, we assume that each centroid takes a step size Δs that is drawn from a distribution $p(\Delta s)$ with a reflecting boundary condition (upper panel of Fig. 5j). The distribution $p(\Delta s)$ was chosen such that the resulting centroid shift Δr closely matches that of the experiment (Fig. 5j; Methods). Experimental centroid shifts show clear distance-dependent correlations, that is, two RFs that are very close to each other are more likely to drift in the same direction on the next day, while RFs that are far apart are more likely to drift in opposite directions (blue line, Fig. 5k). This is in stark contrast with the independent random walk picture (gray horizontal line, Fig. 5k) but can be recapitulated by our model (red line, Fig. 5k), suggesting that the drift of RFs is coordinated at the population level, possibly to preserve representational similarity.

In our model, synaptic weights are updated only when there is task-relevant sensory input. The time scale (or inverse learning rate η) involved with such updates is chosen to capture the magnitude of representational drift which is significant even across a few experimental sessions, each of which is on the order of hundred trials. However, the fact that animals can retain neural representations during rest between sessions and task memories even after weeks without training suggests the existence of other and longer time scales involved in synaptic plasticity. Indeed, if only the time scale involved with (relatively) fast task learning was present and if our model's synapses were updated during intersession periods in a way independent of task-relevant sensory variables, then task-relevant RFs would be rapidly forgotten. We addressed these issues by introducing longer synaptic forgetting time constants in our model (Supplementary Note Section 4 and Extended Data Fig. 6). We found that this model can exhibit both representational drift and retention of representations during arbitrarily long periods without any stimulus access.

A Hebbian/anti-Hebbian network model for drifting RFs of neurons in the PPC

We next study another sensorimotor task in which mice were trained to navigate a virtual T-maze[9,49]. In this experiment, at the first half of the T-stem, mice saw one of two alternative visual scenes and associated them with a left turn or a right turn at the T-junction to receive a reward at the end of the track (Fig. 6a). After learning, a subpopulation of neurons in the PPC developed localized RFs, that is, they fired when a mouse is at a specific position along the T-maze and their RFs tiled the T-maze[9,49]. While mice stably performed the task after learning, the neural population activities in the PPC continuously drifted over weeks[9]. Despite such drift, the task information could be stably encoded by the activities of a subpopulation of PPC neurons[9].

We modeled this system using a Hebbian/anti-Hebbian network with noisy weight updates. This choice is consistent with recent connectome data which show that excitatory neurons in the PPC strongly inhibit each other via local inhibitory neurons[50]. Here, we implement these inhibitory interactions through mutual inhibition between principal neurons. A more detailed network with both excitatory and inhibitory neurons generates very similar result (Extended Data Fig. 7). For simplicity, the input is represented by a vector $\mathbf{x}_{R/L}(\theta) = [\cos(\theta), \sin(\theta), \pm 1]^\top, \theta \in [0, \pi)$, with the last entry indicating a right-turn (1) or a left-turn task (−1).

After learning, the population of output neurons in the model develops positional tuning to the T-maze, that is, for either left-turn input $\mathbf{x}_L$ or right-turn input $\mathbf{x}_R$, there is a subpopulation of neurons that fire most strongly when the animal is at specific positions of the track, forming RFs that tile the maze (Fig. 6b). These tuning properties are consistent with what is observed during experiments[9,49].

To see how the RFs of neurons evolve over time, we first sort neurons with significant RFs based on the centroid positions of their RFs at a reference time point. We find that RFs of neurons drift over time, that is, neurons rarely have the same or similar RFs at two long-separated time points. However, the population representation of animal position and task context information is stable across time. Thus, at any given time, we can identify a subset of neurons with significant RFs that tile the positions of the T-maze for both left-turn and right-turn tasks (Fig. 6c, upper and middle panels). Despite the drift, representational similarities of both left-turn and right-turn tasks are stable over time (Fig. 6c, lower panel).

We can make finer level comparisons to data from PPC experiments[9]. The drift of an RF accumulates over time such that the probability of a centroid shift larger than a certain distance increases with time (left, Fig. 6d). Neurons also gain or lose their tuning to task choices. For example, a group of neurons that are tuned to the left-turn or right-turn tasks may lose such tuning later, and vice versa (left, Fig. 6e). Overall, the fraction of neurons that have positional tuning at any time is constant (left, Fig. 6f). All these behaviors are consistent with data from PPC experiments[9] (right panels of Fig. 6d–f). Together, these comparisons shows that our simple model can explain many characteristics of representational drift in the PPC.

Discussion

In this paper, we explored the hypothesis that representational drift is due to the existence of many (possibly infinite) population codes that achieve a representational objective. Noise in learning drives the network to explore this space, causing the drift of population activity. We explored the experimental consequences of this idea by mathematically modeling representational drift in the hippocampus CA1 and PPC where neurons have drifting localized RFs[1,9,10,31,32] using a well-studied class of models for biologically plausible representation learning: Hebbian/anti-Hebbian networks. These networks are ideal because they optimize similarity-based representational objectives with degenerate optima[34] and are known to learn localized RFs[37]. Further, while simple, Hebbian/anti-Hebbian networks capture the essential properties of RF formation in the studied areas, RFs are shaped by input from upstream and effective lateral inhibition/competition within the layer which can be biologically implemented by inhibitory neurons (Supplementary Note Section 3 and Extended Data Figs. 4 and 7). Through simulations and analytical arguments, we showed that these networks exhibit representational drift, and drifting RFs in these networks are coordinated such that the representational similarity is stable across time.

Our model reproduces key drift phenomena at the population level observed in the hippocampal CA1 place cells and neurons in the PPC. First, a constant fraction of active neurons represents task variables at a given day. Second, neurons drop in and out of this assembly over days. Third, the autocorrelation coefficient of population vectors decay over

https://doi.org/10.1038/s41593-022-01225-z

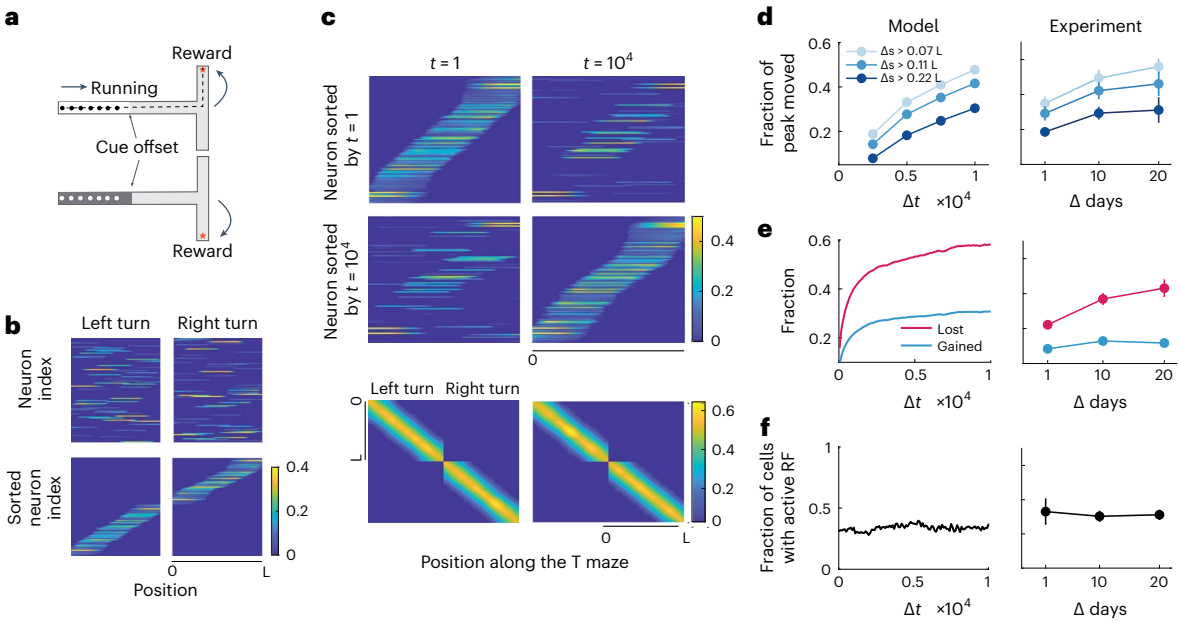


Fig. 6 | Representational drift in the PPC. a, Schematic of the visual-cue-guided T-maze sensorimotor task as in ref. 9. The linear length of the track from the beginning to the end (dashed line) is L. **b**, Population activity for the left-turn and right-turn task before (upper) and after (lower) sorting based on the centroids of neural RFs. Only neurons that have active RFs at the given time point are shown. **c**, Population activity drifts but representational similarity is stable over time. Activity of neurons with active RFs (either tuned to left turn or right turn) in the sorted time (upper and middle). Representational similarity matrix is stable for both left-turn and right-turn task (lower panels). **d**, Shift of RFs for neurons with a significant peak between time t and $t + \Delta t$. Smaller shifts happen more often than larger shifts. **e**, Left: For a group of neurons that have active RFs, the fraction of them that lose tuning at a later time (magenta), and, for a group of neurons that are inactive, the fraction of them that gain tuning at a later time (cyan). **f**, The fraction of the neurons with active RFs is stable across time. In **d**–**f**, left panels are simulation results of our model, right panels are corresponding experimental results plotted using data from ref. 9. Error bars: mean ± s.e.m., $n = 5$, 5 and 4 mice for 1, 10 and 20 d, respectively.

time. Fourth, drift at the population level preserves representational similarity (Figs. 5 and 6).

Besides reproducing already described experimental phenomena, our model makes several testable predictions. First, our model predicts that neurons whose synapses have faster turnover dynamics are more likely to drift more rapidly. For example, the lifetime of spines of pyramidal neurons in the hippocampus is about 1–2 weeks, much shorter than that of neocortex neurons[20]. This suggests that a representational drift should be more prominent in the hippocampus than in the neocortex. Furthermore, the lifetime of synapses can be perturbed by blocking receptors such as NMDA[51], which will alter the stability of RFs. A definitive examination of this prediction requires experiments that both measure the lifetime of synapses and the long-term neural activity in brain regions that represent learned stereotyped behavior under unperturbed and perturbed states. While challenging, this is nonetheless becoming within reach with new experimental techniques. Second, our model predicts that neurons with strongly tuned RFs should be more stable. This prediction can be tested by examining the tuning curve amplitudes of individual neurons and their stability in long-term recording experiments. Furthermore, RF strengths can be perturbed by optogenetic tools to examine how they affect the RF stability. Third, our model predicts that RF drifts are coordinated in a specific way. This coordination arises from the fact that the neural population as a whole is optimizing a similarity-matching objective, and this process enforces RF drifts to be coordinated in a way to preserve representational similarity. We quantified the various distinct ways in which the drift in our models differs from independent RF random walks (Fig. 4e) and verified our prediction in the hippocampal CA1 data (Fig. 5j,k).

Optimization of representational objective functions, such as variants of efficient coding, have been successfully used to describe neural representations especially in early sensory areas[22–30]. A difference of our work is that we not only consider the optimal representations but also the drifting representations encountered during the process of noisy learning. An interesting question is whether this consideration of representational drift can provide evidence and information about the existence and nature of such objective functions. If a neuronal population as a whole is optimizing a representational objective function with degenerate solutions, we expect this process to lead to coordinated drift of RFs in a way that keeps neural representations near the optimum of the representational objective. Then, existence of a representational objective function could be falsified by, for example, observing that the drift of individual RFs is statistically independent, as in the random walk null model we simulated in Fig. 5k. Further, one may be able to gather information about the representational objective from the particular way the drift is coordinated. In our case, preservation of representational similarity during drift was a consequence of the particular representational objective we considered.

Our model can be extended in several ways to study representational changes and drift in other contexts. First, while our focus was on synaptic noise, other sources of noise can also cause representational drift with potentially different statistics. However, we expect the drift to be strongly affected by the degeneracy of the solution space of the objective function. For example, in a feedforward network performing online principal component analysis, which has no degeneracy as the principal subspace projection (PSP) task, we found stabilized representations in the presence of noise (Extended Data Fig. 8). Second, networks optimizing other objective functions than the ones we considered, such as those minimizing a supervised readout error through the biologically implausible backpropagation algorithm, can also show representational drift when learning with noise and redundancy in optimal network weight configurations[12,52]. Third, our model explored drift of localized RFs arising from a particular competitive mechanism for RF formation. Other mechanisms may lead to different

drift statistics. For example, place fields of CA3 also show drift[53] but may be more stable than CA1[54]. This may be due to the strong collateral excitatory synapses between CA3 pyramidal cells[55,56], which are not captured in our model. In another example, sequential activity in the PPC was modeled by training a general recurrent neural network with biologically implausible learning rules[57]. It will be interesting to see whether neurons in such network models with noisy weight updates also show representational drift. Fourth, there can be other representational changes than the ones we considered, for example, changes due to the process of learning near-optimal representations starting from suboptimal ones, as opposed to drifting among near-optimal representations.

Representational drift contradicts the hypothesis that stable neural population activity is the substrate of stable behavior. However, there needs to be stable aspects of representations which provide a substrate for stable downstream decoding and readout[7,58]. Representational similarity can be one such substrate for multiple reasons. First, our modeling shows that achieving stable representational similarity despite the drift of population activity is biologically plausible. Second, stable representational similarity may be a general internal structure of drifting neural population activity. For example, mouse visual cortices show strong representational drift yet the relation between population activities that represent different inputs remains stable and stereotyped[17]. The conserved and stable internal structure of neural activity has also been discovered in the hippocampus and prefrontal cortex in free-behaving mice[59]. A counterexample also exists; representational similarity is not preserved in the mouse piriform cortex during drift, yet the animal can still retain fear-conditioning memory for at least 2 weeks[3]. Third, experimental evidence is consistent with stable representational similarity being a foundation for robust downstream decoding. Studies in monkey motor cortices have shown that stable geometry of latent population dynamics underlies stereotyped reaching tasks[14] despite the inherently variable single neuron activities[12] (see however[4,13]). Interestingly, a recent experiment has shown that the spatial code of different environments in the hippocampus is random in individual rodents but shares the same geometry across different animals[60]. Finally, preserving pairwise similarity of representations may provide some computational benefits. Recent unsupervised learning algorithms for image recognition, such as contrastive representational learning[61] and 'Barlow Twins'[62], are based on objectives that maximize representational similarity between a sample and its distorted/augmented versions. Such algorithms can achieve comparable performance to supervised learning algorithms. From a theoretical point of view, the representational similarity matrix (or kernel) determines the number of sampled stimuli required to learn an accurate linear readout from a population code, indicating that performance need not suffer as long as the representational kernel is preserved[63].

A hypothesis for achieving stable readout despite time-varying neural activity is that the variation happens in the 'coding null space'[64,65]. Representations in our model exhibit drift in all dimensions, precluding the existence of such a space. Similarly, a closer scrutiny of the response of PPC neurons in the T-maze task showed that drift is not confined to a coding null space[66]. Hence, an adaptive readout mechanism which involves synaptic plasticity to track and compensate the drift is required to achieve stable behavior[66,67]. Whether and how such a mechanism is implemented in the brain remains an open question.

The ubiquity of representational drift raises the question of whether it serves a function or it is an inevitable consequence of noise in the brain that needs to be compensated for[7,68]. Representational drift may indeed be desirable or a byproduct of another desirable feature under certain circumstances[8]. For example, in a model of the bird song learning system, variation in the neural representation of the stereotyped behavior enables the system to adapt quickly to a shift of target song and to reduce error due to loss of neurons[69]. Drift can accommodate new learning with minimal inference by modifying existing memories[8]. Drift can also arise in a fast representation learning system which continuously tracks changing environment statistics[3]. Other authors proposed that noisy synaptic plasticity and spine motility enable cortical networks of neurons to carry out probabilistic inference by sampling from a posterior distribution of network configurations[70]. Such sampling could lead to a representational drift as a byproduct.

Overall, our study presents mathematical models that provide parsimonious and mechanistic views of representational drift. Our models capture essential features of drift observed in experiments and make testable predictions. Further, because our mechanistic models are derived from optimization principles, they provide a link between normative accounts of neuronal representations and statistics of representational drift.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01225-z.

References

1. Ziv, Y. et al. Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264 (2013).
2. Li, M. et al. Long-term two-photon imaging in awake macaque monkey. *Neuron* **93**, 1049–1057 (2017).
3. Schoonover, C. E. et al. Representational drift in primary olfactory cortex. *Nature* **594**, 541–546 (2021).
4. Katlowitz, K. A., Picardo, M. A. & Long, M. A. Stable sequential activity underlying the maintenance of a precisely executed skilled behavior. *Neuron* **98**, 1133–1140 (2018).
5. Ulivi, A. F. et al. Longitudinal two-photon imaging of dorsal hippocampal CA1 in live mice. *J. Vis. Exp.* **148**, e59598 (2019).
6. Luo, T. Z. et al. An approach for long-term, multi-probe Neuropixels recordings in unrestrained rats. *eLife* **9**, e59716 (2020).
7. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Curr. Opin. Neurobiol.* **58**, 141–147 (2019).
8. Mau, W., Hasselmo, M. E. & Cai, D. J. The brain in motion: how ensemble fluidity drives memory-updating and flexibility. *eLife* **9**, e63550 (2020).
9. Driscoll, L. N. et al. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986–999 (2017).
10. Gonzalez, W. G. et al. Persistence of neuronal representations through time and damage in the hippocampus. *Science* **365**, 821–825 (2019).
11. Lee, J. S. et al. The statistical structure of the hippocampal code for space as a function of time, context, and value. *Cell* **183**, 620–635 (2020).
12. Rokni, U. et al. Motor learning with unstable neural representations. *Neuron* **54**, 653–666 (2007).
13. Chestek, C. A. et al. Single-neuron stability during repeated reaching in macaque premotor cortex. *J. Neurosci.* **27**, 10742–10750 (2007).
14. Gallego, J. A. et al. Long-term stability of cortical population dynamics underlying consistent behavior. *Nat. Neurosci.* **23**, 260–270 (2020).
15. Redman, W. T. et al. Long-term transverse imaging of the hippocampus with glass microperiscopes. *eLife* **11**, e75391 (2022).
16. Grewe, B. F. et al. Neural ensemble dynamics underlying a long-term associative memory. *Nature* **543**, 670–675 (2017).
17. Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. *Curr. Biol.* **31**, 4327–4339 (2021).
18. Marks, T. D. & Goard, M. J. Stimulus-dependent representational drift in primary visual cortex. *Nat. Commun.* **12**, 5169 (2021).

19. Rumpel, S. & Triesch, J. The dynamic connectome. *Neuroforum* **22.3**, 48–53 (2016).
20. Attardo, A., Fitzgerald, J. E. & Schnitzer, M. J. Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature* **523**, 592–596 (2015).
21. Hazan, L. & Ziv, N. E. Activity dependent and independent determinants of synaptic size diversity. *J. Neurosci.* **40**, 2828–2848 (2020).
22. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954).
23. H. Barlow. *Sensory Communication* (*MIT Press*, 1961).
24. Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes?. *Neural Comput.* **4**, 196–210 (1992).
25. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B Biol. Sci.* **216**, 427–459 (1982).
26. van Hateren, J. H. A theory of maximizing sensory information. *Biol. Cybern.* **68**, 23–29 (1992).
27. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
28. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by V1?. *Vis. Res.* **37**, 3311–3325 (1997).
29. Pehlevan, C., Hu, T. & Chklovskii, D. B. A Hebbian/anti-Hebbian neural network for linear subspace learning: a derivation from multidimensional scaling of streaming data. *Neural Comput.* **27**, 1461–1495 (2015).
30. Chalk, M., Marre, O. & Tkacik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl Acad. Sci. USA* **115**, 186–191 (2018).
31. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
32. Nieh, E. H. et al. Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).
33. Földiak, P. Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* **64**, 165–170 (1990).
34. Pehlevan, C. & Chklovskii, D. B. Neuroscience-inspired online unsupervised learning algorithms: artificial neural networks. *IEEE Signal Process Mag.* **36**, 88–96 (2019).
35. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
36. Pehlevan, C., Sengupta, A. M. & Chklovskii, D. B. Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural Comput.* **30**, 84–124 (2018).
37. Sengupta, A. M. et al. Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. In: *Advances in Neural Information Processing Systems* 7080–7090 (2018).
38. Kämmerer, S., Kob, W. & Schilling, R. Dynamics of the rotational degrees of freedom in a supercooled liquid of diatomic molecules. *Phys. Rev. E* **56**, 5450 (1997).
39. Mazza, M. G. et al. Relation between rotational and translational dynamic heterogeneities in water. *Phys. Rev. Lett.* **96**, 057803 (2006).
40. Hubel, D. H. *Eye, Brain, and Vision (Scientific American Library)* (1995).
41. Peña, J. L. & Konishi, M. Auditory spatial receptive fields created by multiplication. *Science* **292**, 249–252 (2001).
42. Solstad, T., Moser, E. I. & Einevoll, G. T. From grid cells to place cells: a mathematical model. *Hippocampus* **16**, 1026–1031 (2006).
43. Savelli, F. & Knierim, J. J. Hebbian analysis of the transformation of medial entorhinal grid-cell inputs to hippocampal place fields. *J. Neurophysiol.* **103**, 3167–3183 (2010).
44. Bezaire, M. J. & van Soltesz, I. Quantitative assessment of CA1 local circuits: knowledge base for interneuron-pyramidal cell connectivity. *Hippocampus* **23**, 751–785 (2013).
45. Rolotti, S. V. et al. Local feedback inhibition tightly controls rapid formation of hippocampal place fields. *Neuron* **110**, 783–794 (2022).
46. Udakis, M. et al. Interneuron-specific plasticity at parvalbumin and somatostatin inhibitory synapses onto CA1 pyramidal neurons shapes hippocampal output. *Nat. Commun.* **11**, 4395 (2020).
47. Basu, J. & Siegelbaum, S. A. The corticohippocampal circuit, synaptic plasticity, and memory. *Cold Spring Harb. Perspect. Biol.* **7**, a021733 (2015).
48. Yoon, K. J. et al. Grid cell responses in 1D environments assessed as slices through a 2D lattice. *Neuron* **89**, 1086–1099 (2016).
49. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
50. Kuan, A. T. et al. Synaptic wiring motifs in posterior parietal cortex support decision-making. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.13.488176 (2022).
51. Zuo, Y. et al. Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* **436**, 261–265 (2005).
52. Aitken, K., Garrett, M., Olsen, S. & Mihalas, S. The geometry of representational drift in natural and artificial neural networks. *PLoS Comput. Biol.* **18**, e1010716 (2022).
53. Hainmueller, T. & Bartos, M. Parallel emergence of stable and dynamic memory engrams in the hippocampus. *Nature* **558**, 292–296 (2018).
54. Mankin, E. A. et al. Neuronal code for extended time in the hippocampus. *Proc. Natl Acad. Sci. USA* **109**, 19462–19467 (2012).
55. Amaral, D. G. & Witter, M. P. The three-dimensional organization of the hippocampal formation: a review of anatomical data. *Neuroscience* **31**, 571–591 (1989).
56. Rolls, E. T. An attractor network in the hippocampus: theory and neurophysiology. *Learn. Mem.* **14**, 714–731 (2007).
57. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent network models of sequence generation and memory. *Neuron* **90**, 128–142 (2016).
58. Xia, J. et al. Stable representation of a naturalistic movie emerges from episodic activity with gain variability. *Nat. Commun.* **12**, 5170 (2021).
59. Rubin, A. et al. Revealing neural correlates of behavior without behavioral measurements. *Nat. Commun.* **10**, 4745 (2019).
60. Kinsky, N. R. et al. Hippocampal place fields maintain a coherent and flexible map across long timescales. *Curr. Biol.* **28**, 3578–3588 (2018).
61. Chen, T. et al. A simple framework for contrastive learning of visual representations. In *Proc. of the 37th International Conference on Machine Learning*. PMLR, 1597–1607 (2020).
62. Zbontar, J. et al. Barlow twins: self-supervised learning via redundancy reduction. In *Proc. of the 38th International Conference on Machine Learning*. PMLR, 12310–12320 (2021).
63. Bordelon, B. & Pehlevan, C. Population codes enable learning from few examples by shaping inductive bias. *eLife* **11**, e78606 (2022).
64. Druckmann, S. & Chklovskii, D. B. Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol.* **22**, 2095–2103 (2012).
65. Kaufman, M. T. et al. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).
66. Rule, M. E. et al. Stable task information from an unstable neural population. *eLife* **9**, e51121 (2020).

67. Rule, M. E. & O'Leary, T. Self-healing codes: how stable neural populations can track continually reconfiguring neural representations. *Proc. Natl Acad. Sci. USA* **119**, e2106692119 (2022).
68. Masset, P., Qin, S. & Zavatone-Veth, J. A. Drifting neuronal representations: bug or feature?. *Biol. Cybern.* **116**, 253–266 (2022).
69. Duffy, A. et al. Variation in sequence dynamics improves maintenance of stereotyped behavior in an example from bird song. *Proc. Natl Acad. Sci. USA* **116**, 9592–9597 (2019).
70. Kappel, D. et al. Network plasticity as Bayesian inference. *PLoS Comput. Biol.* **11**, e1004485 (2015).

Methods

Similarity matching and the linear Hebbian/anti-Hebbian network

The linear Hebbian/anti-Hebbian network can be derived from Eq. (1). The detailed derivation can be found in[29,36], we sketch the main steps here. Starting from the cross term in Eq. (1), by introducing a new matrix variable $\mathbf{W} \in \mathbb{R}^{k\times n}$, we obtain

$$-\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{y}_t^\top\mathbf{y}_{t'}\mathbf{x}_t^\top\mathbf{x}_{t'} = \min_{\mathbf{W}\in\mathbb{R}^{k\times n}} -\frac{2}{T}\sum_{t=1}^{T}\mathbf{y}_t^\top\mathbf{W}\mathbf{x}_t + \mathrm{Tr}\left(\mathbf{W}^\top\mathbf{W}\right). \tag{6}$$

Similarly, we can introduce another matrix variable $\mathbf{M}$ for the quartic $\mathbf{y}_t$ term in Eq. (1):

$$-\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{y}_t^\top\mathbf{y}_{t'}\mathbf{y}_t^\top\mathbf{y}_{t'} = \max_{\mathbf{M}\in\mathbb{R}^{k\times k}} \frac{2}{T}\sum_{t=1}^{T}\mathbf{y}_t^\top\mathbf{M}\mathbf{y}_t - \mathrm{Tr}\left(\mathbf{M}^\top\mathbf{M}\right). \tag{7}$$

By substituting Eqs. (6) and (7) into Eq. (1) and changing orders of optimization[36], we get:

$$\min_{\mathbf{W}\in\mathbb{R}^{k\times n}}\max_{\mathbf{M}\in\mathbb{R}^{k\times k}}\frac{1}{T}\sum_{t=1}^{T}\left[2\mathrm{Tr}\left(\mathbf{W}^\top\mathbf{W}\right) - \mathrm{Tr}\left(\mathbf{M}^\top\mathbf{M}\right) + \min_{\mathbf{y}_t\in\mathbf{R}^k} l_t\left(\mathbf{W},\mathbf{M},\mathbf{y}_t\right)\right], \tag{8}$$

where

$$l_t\left(\mathbf{W},\mathbf{M},\mathbf{y}_t\right) = -4\mathbf{x}_t^\top\mathbf{W}\mathbf{y}_t + 2\mathbf{y}_t^\top\mathbf{M}\mathbf{y}_t. \tag{9}$$

The minimax problem (8) can be solved in an online fashion by the following two-step algorithm. For each input, first, we minimize (9) while keeping $\mathbf{W}$ and $\mathbf{M}$ fixed, which is done by running the following dynamics of the output variable $\mathbf{y}_t$ until convergence

$$\dot{\mathbf{y}}_t = \mathbf{W}\mathbf{x}_t - \mathbf{M}\mathbf{y}_t. \tag{10}$$

Here the time derivative is with respect to the parameter that governs neural dynamics, not the online time step t. Second, after the convergence of $\mathbf{y}_t$, we keep it fixed and update $\mathbf{W}$ and $\mathbf{M}$ by gradient descent and gradient ascent steps on (8), respectively:

$$W_{ij} \leftarrow W_{ij} + \eta\left(y_i x_j - W_{ij}\right), \quad M_{ij} \leftarrow M_{ij} + \eta\left(y_i y_j - M_{ij}\right). \tag{11}$$

The above learning algorithm defined by Eqs. (10) and (11) can be naturally mapped onto a single-layer biologically plausible neural network, the linear Hebbian/anti-Hebbian network. Here $\mathbf{y}_t$ is the neural activity of the output, $\mathbf{W}$ and $\mathbf{M}$ are synaptic matrices of the forward and lateral connections, respectively. The synaptic update rule Eq. (11) is local since the change of a synapse only depends on the activity of presynaptic and postsynaptic neurons.

To achieve drifting representations, we introduce noise to synaptic updates

$$\Delta\mathbf{W}_t = \eta\left(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}_t\right) + \boldsymbol{\xi}_t^W, \quad \Delta\mathbf{M}_t = \eta\left(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}_t\right) + \boldsymbol{\xi}_t^M, \tag{12}$$

where the noise terms $\xi_{ij,t}^W, \xi_{ij,t}^M$ are independent Gaussian noises with the following statistics: $\left\langle\xi_{ij,t}^W\right\rangle = \left\langle\xi_{ij,t}^M\right\rangle = 0$ and $\left\langle\xi_{ij,t}^W\xi_{kl,t'}^W\right\rangle = \eta\sigma_1^2\delta_{ik}\delta_{jl}\delta_{tt'}$, $\left\langle\xi_{ij,t}^M\xi_{kl,t'}^M\right\rangle = \eta\sigma_2^2\delta_{ik}\delta_{jl}\delta_{tt'}$. For simplicity, we set $\sigma_1 = \sigma_2 = \sigma$ in all our numerical simulations.

Calculation of the rotational diffusion constant

An analytical calculation of the rotational diffusion constant, defined by refs. 38, 39, 71,

$$D_\varphi \equiv \lim_{t\to\infty}\frac{1}{2(k-1)t}\left\langle|\vec{\varphi}(t) - \vec{\varphi}(0)|^2\right\rangle, \tag{13}$$

where $\vec{\varphi}(t)$ is the sum of single-step angular displacements of the data cloud (see Supplementary Note Section 1) and k is the number of dimensions in which rotation occurs, is difficult. However, we were able to obtain an approximation that matches numerical experiments very well, as shown in Fig. 2e,f. We present the details of this derivation in Supplementary Note Section 1. Our approximation assumes that (1) angular displacements of the representation vectors after different time steps are not correlated, and (2) the network weights stay close to the optimal representation manifold. Under these assumptions, D_φ can be approximated by the mean squared angular displacement (MSAD),

$$D_\varphi \approx \frac{1}{2(k-1)}\left\langle|\Delta\vec{\varphi}|^2\right\rangle, \tag{14}$$

where $\Delta\vec{\varphi}$ arises from a noisy synaptic update to the network with an optimal set of synapses. We calculate MSAD analytically (Supplementary Note Section 1) to arrive at Eq. (4).

To numerically estimate D_φ from the trajectory of $\mathbf{y}_t$ with a total length of T time steps, we first calculate each simulation step and then estimate $\vec{\varphi}(t)$ by cumulatively summing $\delta\vec{\varphi}$ up to time step t. Next, we estimate the MSAD of interval τ, measured in discrete time steps, using all the pairs of $\vec{\varphi}(t+\tau)$ and $\vec{\varphi}(t)$, which gives $\left\langle|\Delta\vec{\varphi}|^2\right\rangle = \left\langle|\vec{\varphi}(t+\tau) - \vec{\varphi}(t)|^2\right\rangle$. Last, we plot $|\Delta\vec{\varphi}|^2$ as a function of τ and fit a line that pass the origin to the data. The slope of the best fit is then $4D_\varphi$.

NSM and the nonlinear Hebbian/anti-Hebbian network

The nonlinear Hebbian/anti-Hebbian network (Eqs. 16 and 17) can be derived from the NSM problem[34,72]. Denoting the input data as a set of vectors $\mathbf{x}_{t=1,\cdots,T} \in \mathbb{R}^n$ and the corresponding outputs as vectors $\mathbf{y}_{t=1,\cdots,T} \in \mathbb{R}^k$, the NSM objective is defined as

$$\min_{\forall t\in\{1,\ldots T\}:\mathbf{y}_t\geq 0}\frac{1}{2T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}(\mathbf{x}_t^\top\mathbf{x}_{t'} - \mathbf{y}_t^\top\mathbf{y}_{t'} - \alpha^2)^2 + \frac{1}{T}\sum_{t=1}^{T}(2\beta_1||\mathbf{y}_t||_1 + \beta_2||\mathbf{y}_t||_2^2), \tag{15}$$

where α^2 sets the threshold of similarity to be preserved in the output representation, and the other two regularizers β_1, β_2 control the sparsity and amplitude of the output. When $\beta_1 = \beta_2 = 0$, this objective function reduces to Eq. (5) up to an overall factor of ½ which does not affect the solutions. Compared to Eq. (1), the non-negativity of $\mathbf{y}_t$ breaks the rotational symmetry of the solution but keeps its permutation symmetry. To see this more clearly, the sum in the first term in Eq. (15) can be written in terms of input–output Gram matrices: $||\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y} - \alpha^2\mathbf{E}||_F^2$, where $\mathbf{X} \in \mathbb{R}^{n\times T}$, $\mathbf{Y} \in \mathbb{R}^{k\times T}$, and $\mathbf{E} \in \mathbb{R}^{T\times T}$ is the matrix with all entries set to 1. Thus, if $\mathbf{Y}$ is a solution, then $\mathbf{PY}$ is also a solution for any permutation matrix $\mathbf{P}$. The regularizer terms are also invariant under such permutation.

A characteristic feature of the NSM objective, Eq. (15), is that it leads to localized RFs[37]. To build an intuition to why this happens, we can consider the simpler case where $\beta_1 = \beta_2 = 0$ and a single pair of inputs. If two inputs are similar, that is, $\mathbf{x}_1 \cdot \mathbf{x}_2 > \alpha^2$, then the corresponding outputs $\mathbf{y}_1$ and $\mathbf{y}_2$ would prefer $\mathbf{y}_1 \cdot \mathbf{y}_2 = \mathbf{x}_1 \cdot \mathbf{x}_2 - \alpha^2$, that is, they are also similar. In contrast, if two inputs are less similar, that is, $\mathbf{x}_1 \cdot \mathbf{x}_2 < \alpha^2$, due to the non-negativity of outputs, $\mathbf{y}_1, \mathbf{y}_2$ they tend to be orthogonal: $\mathbf{y}_1 \cdot \mathbf{y}_2 = 0$. To achieve this, dissimilar inputs must activate nonoverlapping sets of neurons. Thus, as in manifold learning, Eq. (15) preserves the local geometric structure of inputs. A detailed explanation of why localized RFs are learned in a simplified version of Eq. (15) is provided in ref. 37.

As in the linear case, an online optimization of Eq. (15) can be interpreted as a neural network algorithm, see refs. 72, 73 for detailed derivations. When $\beta_1 = \beta_2 = 0$, the network takes an input $\mathbf{x}_t$ and generates an output $\mathbf{y}_t$ by running the following neural dynamics until it converges:[72]

$$\begin{aligned}\dot{u}_i &= -u_i + [\mathbf{W}\mathbf{x}_t]_i - \alpha b_i - [\overline{\mathbf{M}}\mathbf{y}_t]_i,\\ y_i &= \max\{u_i/M_{ii}, 0\},\end{aligned} \tag{16}$$

where u_i and y_i represent the membrane potential and firing rate of neuron i, and b_i is the bias term. The forward weight matrix $\mathbf{W} \in \mathbb{R}^{k\times n}$ and recurrent weight matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ (we have defined $\overline{\mathbf{M}} = \mathbf{M} - \mathrm{diag}(\mathbf{M})$) update according to the following noisy learning rule[72]:

$$\Delta\mathbf{W} = \eta\left(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}\right) + \boldsymbol{\xi}_t^W, \quad \Delta\mathbf{M} = \eta\left(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}\right) + \boldsymbol{\xi}_t^M, \\ \Delta\mathbf{b} = \eta\left(\alpha\mathbf{y}_t - \mathbf{b}\right), \tag{17}$$

where η is the learning rate, and ξ_t^W and ξ_t^M are Gaussian white noise terms: $\left\langle \xi_{ij,t}^W \right\rangle = \left\langle \xi_{ij,t}^M \right\rangle = 0$ and $\left\langle \xi_{ij,t}^W \xi_{kl,t'}^W \right\rangle = \eta\sigma_1^2\delta_{ik}\delta_{jl}\delta_{tt'}$, $\left\langle \xi_{ij,t}^M \xi_{kl,t'}^M \right\rangle = \eta\sigma_2^2\delta_{ik}\delta_{jl}\delta_{tt'}$.

The properties of the above learning rule without noise have been studied previously[33,37,72–74]. With the regularization of $\mathbf{y}_t$, that is, $\beta_1 \neq 0, \beta_2 \neq 0$, the neural dynamics derived from Eq. (15) differs from Eq. (16) only by the transfer function

$$y_i = \max\left\{\left(u_i - \beta_1\right)/\left(\beta_2 + M_{ii}\right), 0\right\}. \tag{18}$$

Derivation of the diffusion constant of the ring model

We define a diffusion constant of the centroid by the conventional relation: $\left\langle \left(\phi\left(t+\Delta t\right) - \phi\left(t\right)\right)^2 \right\rangle = 2D\Delta t$, where $\phi(t)$ is the centroid position of the RF at time t. Here Δt corresponds to an arbitrary number of time steps. We sketch the derivation of diffusion constant in the single neuron scenario here, more details are provided in Supplementary Note Section 2. We again consider the approximation that the diffusion constant can be approximated by the mean squared displacement around a fixed point by a noisy synaptic update.

Consider a single output neuron that learns an RF from inputs that are on a ring manifold (Fig. 3a). The response of the output neuron to an input $\mathbf{x} = [\cos\theta, \sin\theta]^\top$ is

$$y\left(\theta\right) = \frac{1}{m+\beta}\left[w_1\cos\theta + w_2\sin\theta - \alpha b\right]_+, \tag{19}$$

where $[\ldots]_+$ denotes the rectified linear function and β is the l_2 regularizer (we have set $\beta_1 = 0$). The stationary state parameters $\{w_1^*, w_2^*, m^*, b^*\}$ satisfy the following conditions:

$$w_1^* = \left\langle y(\theta)\cos\theta \right\rangle_\theta, \quad w_2^* = \left\langle y(\theta)\sin\theta \right\rangle_\theta, \\ m^* = \left\langle y^2(\theta) \right\rangle_\theta, \quad b^* = \alpha\left\langle y\left(\theta\right) \right\rangle_\theta, \tag{20}$$

where $\langle\ldots\rangle_\theta$ denotes an average over the ring. These equations can be solved self-consistently by assuming an ansatz of the form:

$$y_\phi\left(\theta\right) = \mu[\cos\left(\theta - \phi\right) - \cos\left(\psi\right)]_+, \tag{21}$$

where ϕ is the centroid. This gives the dependence of μ and ψ on α, β

$$\mu^2 = \frac{2\psi - \sin 2\psi - 4\beta\pi}{4\psi + 2\psi\cos 2\psi - 3\sin 2\psi}, \quad \alpha^2 = \frac{\cos\psi(2\psi - \sin 2\psi)}{4(\sin\psi - \psi\cos\psi)}, \tag{22}$$

from which all parameters can be recovered (Supplementary Note Section 2). Then, noting the fact that $dy(\theta)/d\theta = 0$ at $\theta = \phi$ implies $\tan\phi = w_2/w_1$, one can approximate the change in the centroid due to small weight changes by

$$\Delta\phi = \frac{1}{\hat{\mu}}\left(\Delta w_2\cos\phi - \Delta w_1\sin\phi\right), \tag{23}$$

where $\hat{\mu} = \sqrt{w_1^{*2} + w_2^{*2}}$ is the norm of weight vector. Using the noisy update rule (17) and (21), the shift of the centroid due to one-step update becomes

$$\Delta\phi = \frac{1}{\hat{\mu}}\{\eta\mu[\cos(\theta - \phi) - \cos\psi]_+\sin\left(\theta - \phi\right) + \left(\xi_2\cos\phi - \xi_1\sin\phi\right)\}. \tag{24}$$

Finally, using the relation $\left\langle (\Delta\phi)^2 \right\rangle \approx 2D$ for a single-step update, we have

$$D \approx \frac{1}{2}\left(\gamma\eta^2 + \frac{\eta\sigma^2}{\hat{\mu}^2}\right), \tag{25}$$

where

$$\gamma \equiv \frac{\mu^2}{\hat{\mu}^2}\left\langle\left([\cos(\theta - \phi) - \cos\psi]_+^2\sin^2(\theta - \phi)\right)\right\rangle_\theta. \tag{26}$$

See Supplementary Note Section 2 for full expressions. When $\alpha = \beta = 0$, we have $\gamma = 1$ and $\hat{\mu} = 1/4$, (25) reduces to

$$D \approx \eta^2/2 + 8\eta\sigma^2. \tag{27}$$

Numerical simulation of two-dimensional place cells

We considered a 32×32 grid plane as the environment, each position (x, y) is represented by a group of grid cells with different grid spacings, orientations and offsets as observed in experiment[75]. The hexagonal firing fields of grid cells are modeled as a summation of three two-dimensional sinusoidal functions as in[42,76,77]

$$G\left(\mathbf{r}\right) = \frac{2}{3}\left(\frac{1}{3}\sum_{i=1}^{3}\cos\left(\frac{4\pi}{\sqrt{3}l}\mathbf{e}_i\cdot\left(\mathbf{r} - \mathbf{r}_0\right)\right) + \frac{1}{2}\right), \tag{28}$$

where $\mathbf{r} = [x, y]^\top$ is the location on the plane, $\mathbf{r}_0 = [x_0, y_0]^\top$ is the phase offset, l is the grid spacing, and $\mathbf{e}_i = \left(\cos\left(\frac{2\pi i}{3} + \theta\right), \sin\left(\frac{2\pi i}{3} + \theta\right)\right), i = 1, 2, 3$ is the unit vector in the direction $2\pi i/3 + \theta$ with θ being the grid orientation. In the simulation, grid cells have 5 modules, that is, $N_l = 5$. The value of l increases as geometric series with a ratio 1.42 that is consistent with experiments[75]. For example, if the smallest spacing is $0.2L$ with L being the linear length of the plane, then the rest of the spacings would be $0.2 \times 1.42L, \cdots, 0.2 \times 1.42^{N_l-1}L$. In each module, the number of orientations θ is $N_\theta = 6$, which are drawn uniformly in the range $[0, \pi/3)$. Similarly, the number of grid phase offsets x_0, y_0 are N_x and N_y, which are drawn uniformly in the range $[0, l)$. As a result, the total number of grid cells is $N_g = N_l N_\theta N_x N_y$.

Numerical simulation of one-dimensional place cells

We consider a linear track with length L. Tuning curves of grid cells on the linear track are slices through two-dimensional grid fields described above with $N_x = N_y = 3$ and smallest grid spacing $0.25L$. The orientations of the slices are the same and randomly selected in the range $[0, \pi/3]$.

Autocorrelation coefficient of the population vector

In all the cases, the autocorrelation coefficient ρ of population vector is defined as Pearson's correlation coefficient between $\mathbf{y}_t$ and $\mathbf{y}_0$ to the same input:

$$\rho\left(t\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{0,i} - \bar{y}_0}{\sigma_{y,0}}\right)\left(\frac{y_{t,i} - \bar{y}_t}{\sigma_{y,t}}\right), \tag{29}$$

where $\bar{y}_0, \bar{y}_t$ are the means of $y_{0,i}$ and $y_{t,i}$; and $\sigma_{y,0}, \sigma_{y,t}$ are the standard deviations of $y_{0,i}$ and $y_{t,i}$.

Step size in simulations where place fields perform independent random walks

In Fig. 3c, the step size of each independent random walker was drawn from the distribution $p\left(\Delta r\right)$ of one-step centroid shift in our model, $\Delta r = r(t+1) - r(t)$.

In Fig. 5j,k, the step size of independent random walks was drawn from a distribution $p(\Delta s)$ closely matching that of experiment.

To determine this distribution, we first calculated the distribution of centroid shifts between two adjacent days in the experiment: $p(\Delta r)$ with $\Delta r = r(t+1) - r(t)$. For a random walk whose centroid is at position $\hat{r}_t$, its position at the next time step is $\hat{r}_{t+1} = \hat{r}_t + \Delta s$ with Δs randomly sampled from $p(\Delta s)$. To constrain $\hat{r}_{t+1}$ in the range of the track $[0, L]$ with L being the length of the track, we assumed a reflecting boundary condition, which gives

$$\hat{r}_{t+1} = \begin{cases} |\hat{r}_t + \Delta s|, & \hat{r}_t + \Delta s < 0, \\ 2L - (\hat{r}_t + \Delta s) & \hat{r}_t + \Delta s > L \\ \hat{r}_t + \Delta s, & \text{otherwise} \end{cases}. \quad (30)$$

The shift of centroid in the random walk model is then determined by $\Delta\hat{r} = \hat{r}_{t+1} - \hat{r}_t$ according to the above equation. Our aim is to find a distribution $p(\Delta s)$, such that $p(\Delta\hat{r})$ is close to that of experiment $p(r)$. Based on the shape of experimentally measured $p(r)$, we searched $p(\Delta s)$ from a family of Levy's alpha stable distribution[78] by minimizing the Kullback–Leibler divergence between $p(r)$ and $p(\hat{r})$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

No new experimental data were generated in this study.

Experimental data presented in Fig. 5 are originally described in ref. 10. We used the processed data and MATLAB code, which are available at the Caltech Research Data Repository (https://doi.org/10.22002/d1.1229) to produce these plots.

Experimental data presented in Fig. 6 is extracted from Fig. 2c and d of ref. 9. The data are freely available in ref. 79.

Code availability

Codes for numerical experiments were written in MATLAB (R2020b). Analysis and figures were made using MATLAB (R2020b) except Fig. 4a,b, which is made by R (version 4.2.0). All codes are available in the GitHub repository https://github.com/Pehlevan-Group/representation-drift.

References

71. Hunter, G. L. et al. Tracking rotational diffusion of colloidal clusters. *Opt. Express* **19**, 17189–17202 (2011).
72. Pehlevan, C. A spiking neural network with local learning rules derived from nonnegative similarity matching. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 7958–7962 (2019).
73. Pehlevan, C. & Chklovskii, D. B. A Hebbian/anti-Hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features. In *Proc. of 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 769–775 (2014).
74. Pehlevan, C., Mohan, S. & Chklovskii, D. B. Blind nonnegative source separation using biological neural networks. *Neural Comput.* **29**, 2925–2954 (2017).
75. Stensola, H. et al. The entorhinal grid map is discretized. *Nature* **492**, 72–78 (2012).
76. Kropff, E. & Treves, A. The emergence of grid cells: Intelligent design or just adaptation? *Hippocampus* **18**, 1256–1269 (2008).
77. Lian, Y. & Burkitt, A. N. Learning an efficient hippocampal place map from entorhinal inputs using Non-Negative sparse coding. *eNeuro* **8**, ENEURO.0557-20.2021 (2021).
78. Samorodnitsky, G. & Taqqu, M. S. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance: Stochastic Modeling* (Routledge, 2017).
79. Driscoll, L. N. et al. *Data From: Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex Dataset* (Dryad, 2020).
80. Sanger, T. D. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Netw.* **2**, 459 (1989).

Acknowledgements

This work was supported by the NIH grant 1UF1NS111697-01 (C.P. and S.Q.), the Intel Corporation through Intel Neuromorphic Research Community (C.P.) and a Google Faculty Research Award (C.P.).

We thank W. Gonzalez, H. Zhang, A. Harutyunyan and C. Lois from California Institute of Technology for sharing the data on place cell recordings. We thank L. Driscoll, N. Pettit, M. Minderer, S. Chettih and C. Harvey for making the T-maze experimental data available. We are grateful to members of Pehlevan group for helpful discussions, and W. Gonzalez, L. Driscoll and C. Harvey for comments on the manuscript.

Author contributions

This work resulted from the merging of two independent projects at their initial stages: one by S.Q. and C.P. and the other by S.F., D.L., A.M.S and D.B.C. For the current manuscript, S.Q. and C.P. conceived and designed the study with input from S.F., D.L., A.M.S and D.B.C. S.Q. performed the numerical simulations and analytical calculations. S.Q. and C.P. analyzed and interpreted the data with input from S.F., D.L., A.M.S. and D.B.C. S.Q. and C.P. wrote the manuscript with comments from S.F., D.L., A.M.S. and D.B.C.

Competing interests

The authors declare no competing interests.

Additional information

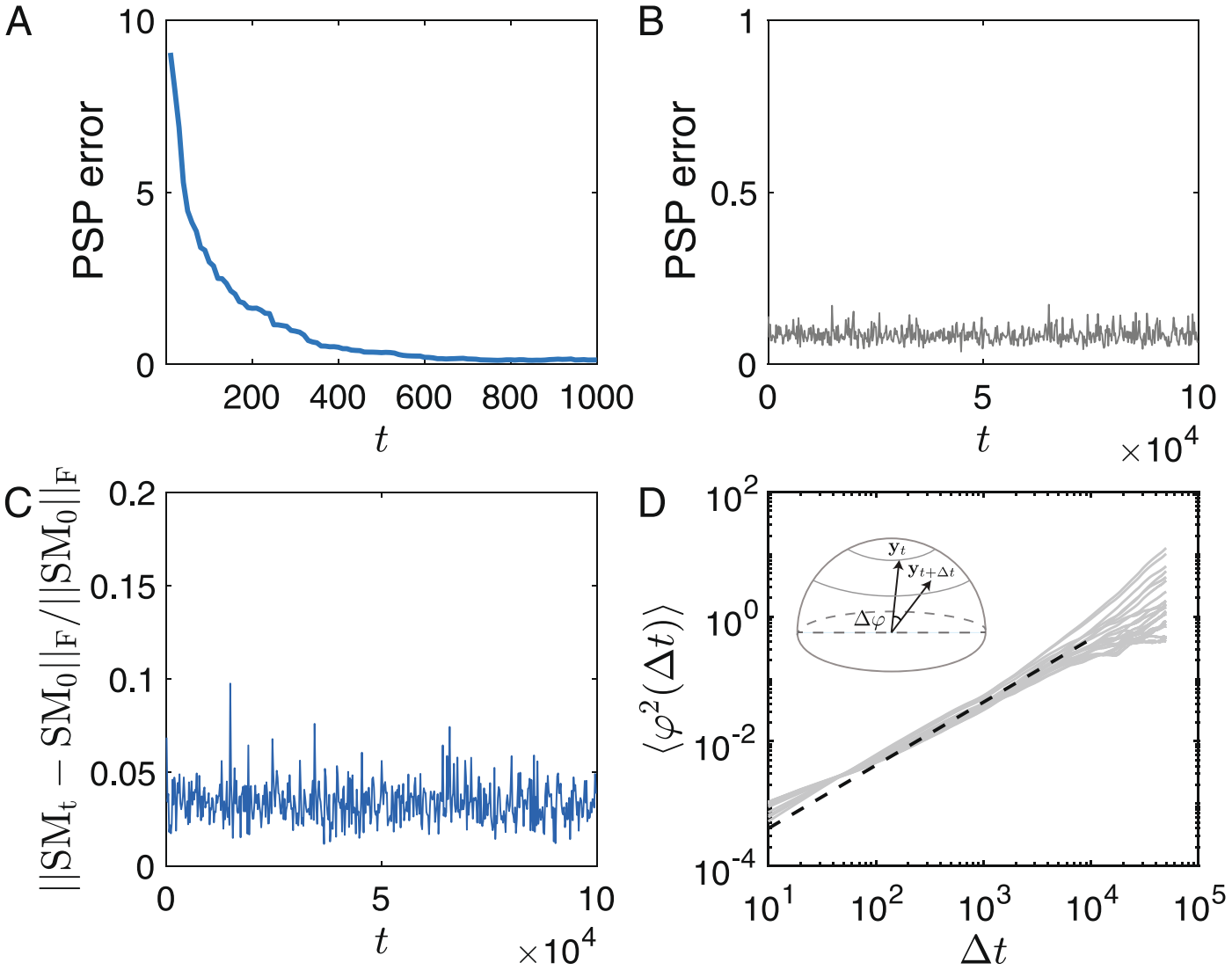
Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01225-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01225-z.

Correspondence and requests for materials should be addressed to Cengiz Pehlevan.

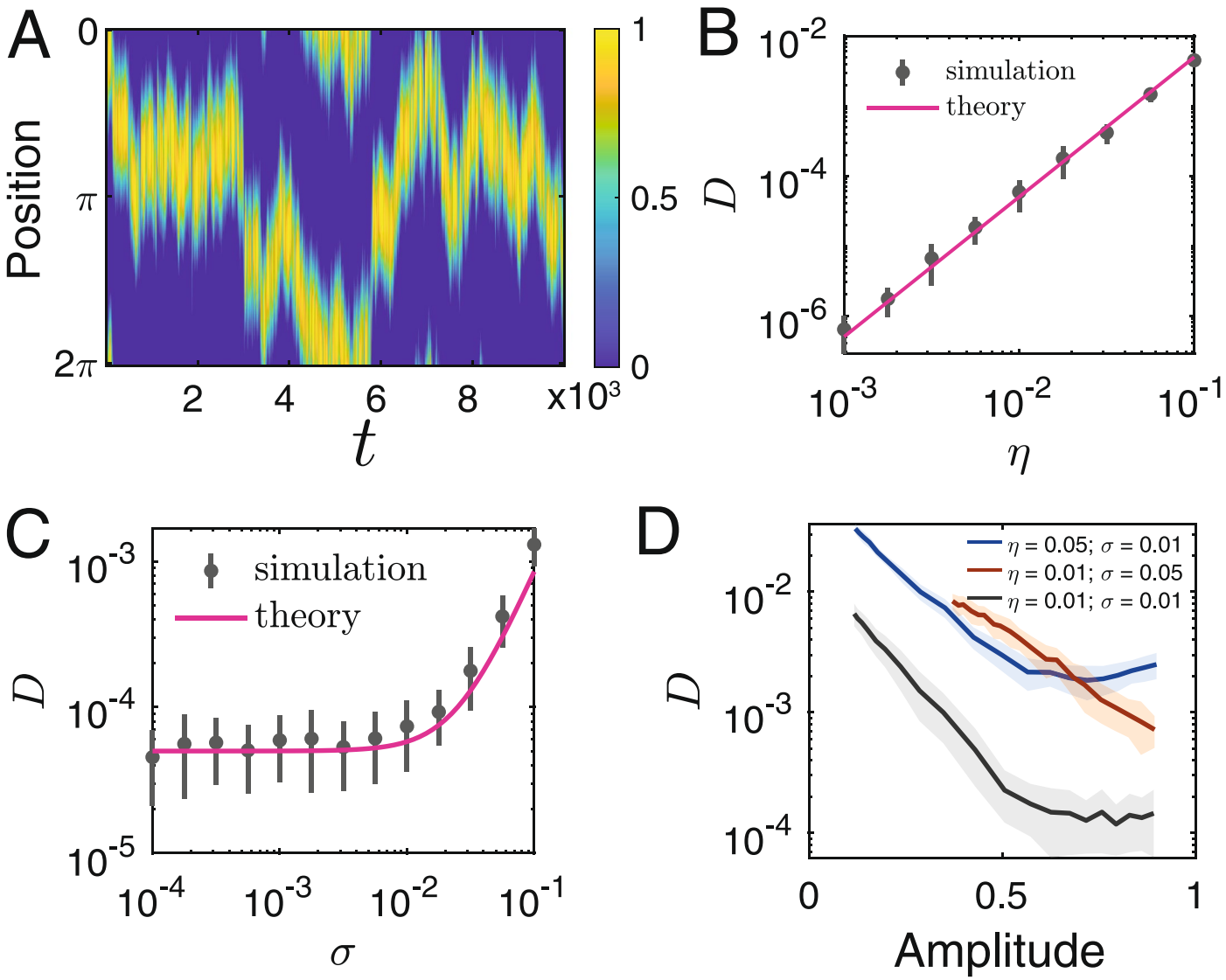
Peer review information *Nature Neuroscience* thanks Adrienne Fairhall, Timothy O'Leary and Alessandro Treves for their contribution to the peer review of this work.

https://doi.org/10.1038/s41593-022-01225-z



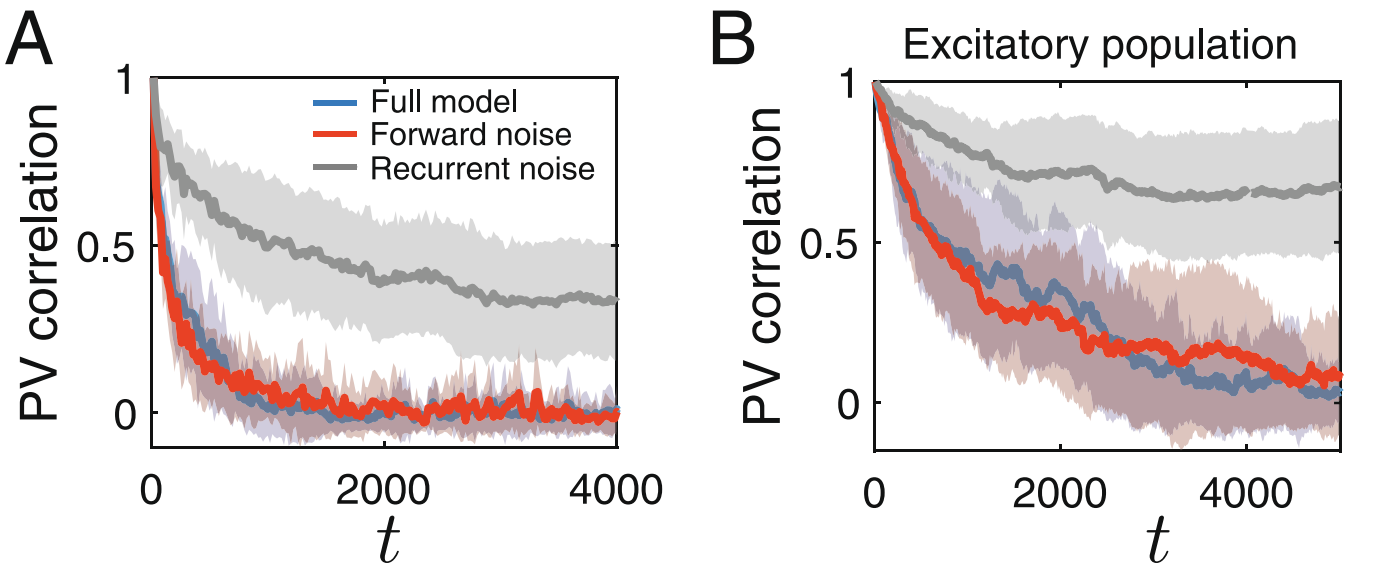
Extended Data Fig. 1 | Performance of the linear Hebbian/anti-Hebbian network in the PSP task. (**a**,**b**) The PSP error as quantified by $\left\|\mathbf{F}_t^\top \mathbf{F}_t - \mathbf{U}\mathbf{U}^\top\right\|_F / \left\|\mathbf{U}\mathbf{U}^\top\right\|_F$, where $\mathbf{U}$ is a $n \times k$ matrix whose columns are the top k left singular vectors of $\mathbf{X} \equiv [\mathbf{x}_1, \cdots, \mathbf{x}_T]$ and $\mathbf{F}_t \equiv \mathbf{M}^{-1}{}_t \mathbf{W}_t$, drops very quickly during training (**a**) and maintains the low error in the presence of synaptic noise (**b**). (**c**) The relative change of the similarity matrix at time t compared to time point 0, corresponding to the point where the network initially learned the task, defined as $\left\|\mathbf{Y}_t^\top \mathbf{Y}_t - \mathbf{Y}_0^\top \mathbf{Y}_0\right\|_F / \left\|\mathbf{Y}_0^\top \mathbf{Y}_0\right\|_F$. (**d**) Estimating rotational diffusion constant D from mean squared angular displacement (MSAD). Gray lines are MSAD estimated based on individual representation trajectory $\mathbf{y}(t)$. The dashed line is a linear fit between $\left\langle (\Delta\varphi)^2 \right\rangle \equiv \left\langle (\varphi(t+\Delta t) - \varphi(t))^2 \right\rangle$ and Δt to estimate the rotational diffusion constant. Inset: illustration of $\Delta\varphi$. Parameters are the same as Fig. 2 in the main text.

https://doi.org/10.1038/s41593-022-01225-z



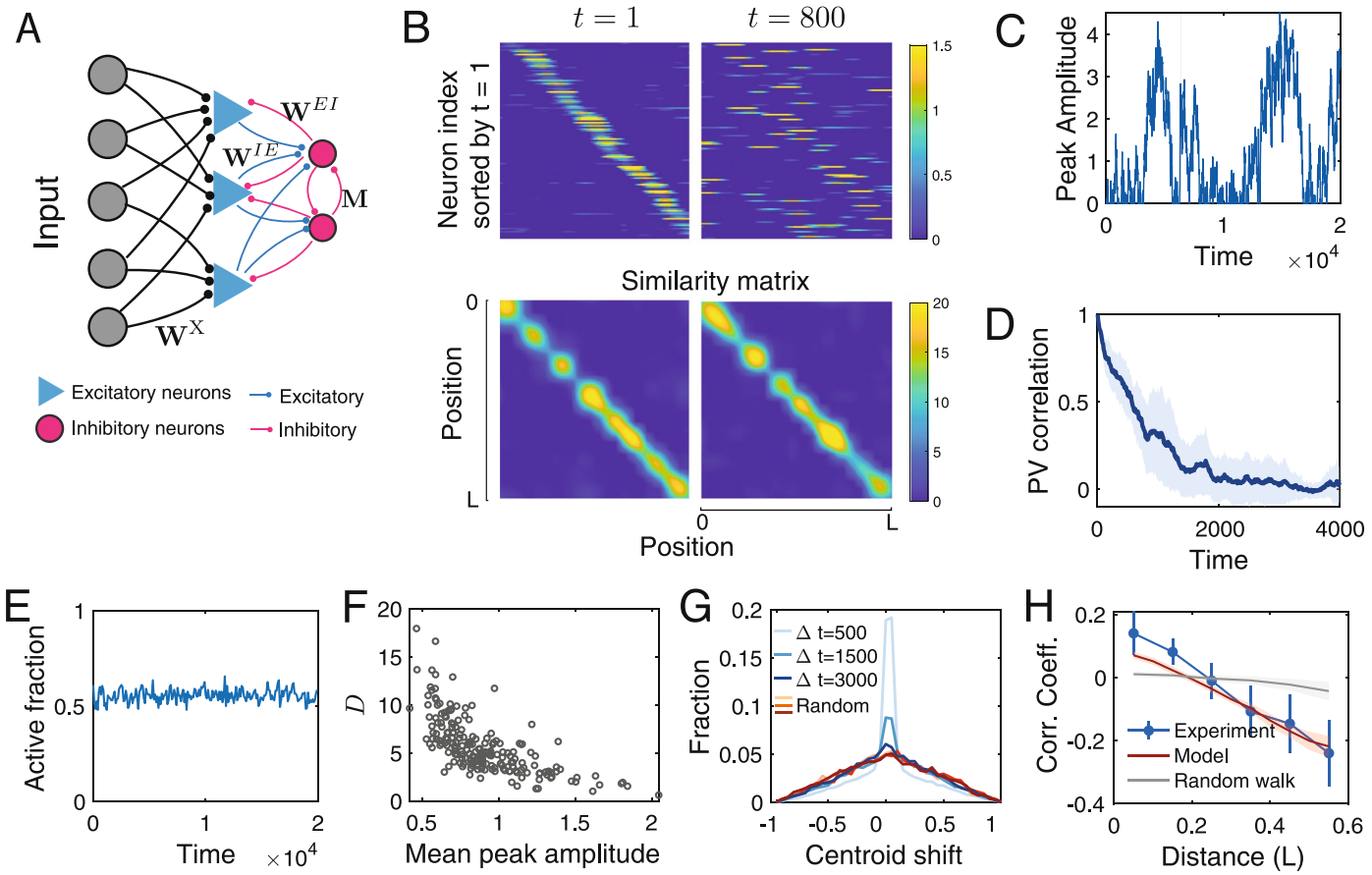
Extended Data Fig. 2 | A single output neuron's RF drift when stimuli lives on a ring. (**a**) With stimuli living on a ring, a single RF has the shape of a truncated cosine curve, whose centroid drifts on the ring like a random walk. (**b**,**c**) The effective diffusion constant D of the centroid position increases with learning rate η both without explicit synaptic noise (σ = 0) (**b**), and with explicit noise (**c**). Error bars: mean ± SD, n = 40 simulations. Magenta lines correspond to theory Eq. (27) in the main text. (**d**) The single RF with larger amplitude has smaller diffusion constant. The amplitude of RF is varied by changing the value of α. Shading: mean ± SD, n = 40 simulations.

https://doi.org/10.1038/s41593-022-01225-z



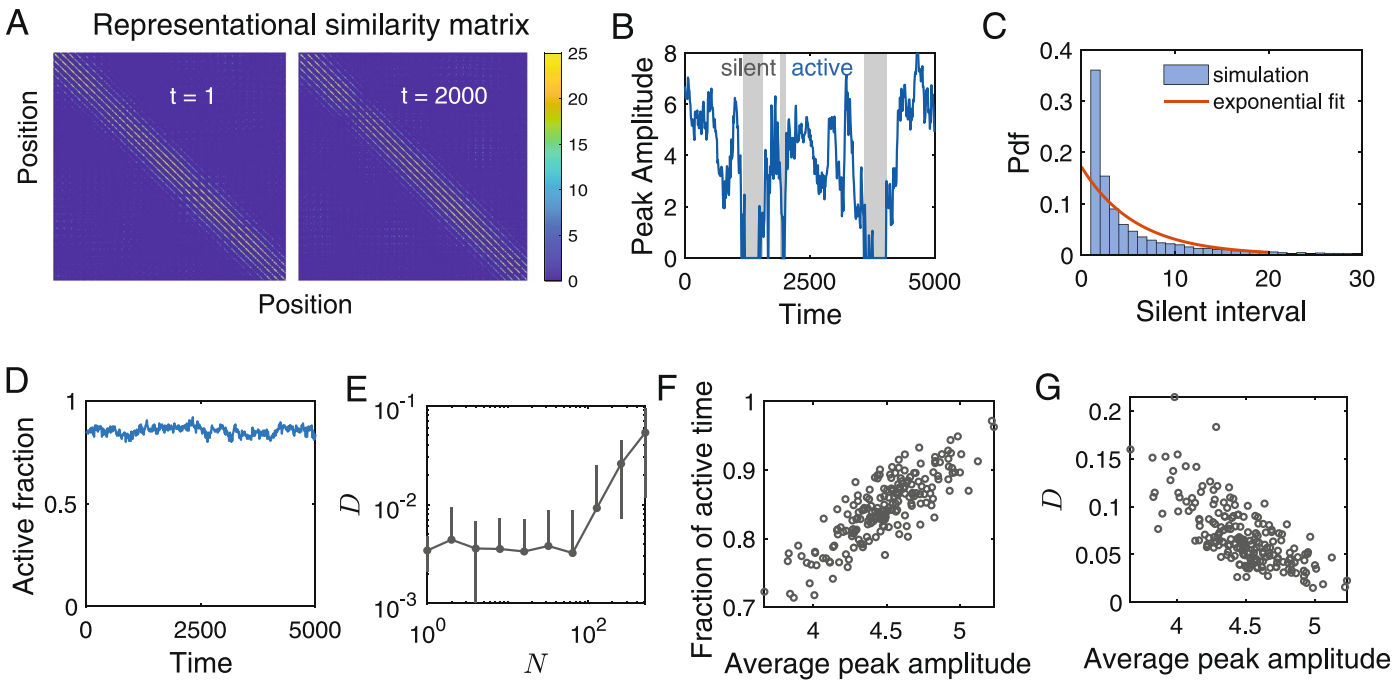
Extended Data Fig. 3 | Distinct contribution of noise from forward synapses and recurrent synapses to representational drift in the 1D place cell model. To further verify the role of noise in feedforward synapses, we simulated models of representational drift in 1D place cells, and compared the correlation coefficient of population vectors of the principal output neurons in three different noise scenarios: full model with all synaptic noises (blue); noise only in the forward synapses $\mathbf{W}$ ($\sigma_M = 0$, red); and noise only in recurrent synapses $\mathbf{M}$ ($\sigma_W = 0$, gray). These models are further explored in main text Fig. 5 and Extended Data Fig. 4. In both the simplified 1D place cell model (**a**) and the more detailed network model with inhibitory neurons (**b**), noise in the forward matrix has much larger influence on the representational drift. For the network with inhibitory neurons, forward noise corresponds to all noises in matrices $\mathbf{M}$, $\mathbf{W}^{EI}$, $\mathbf{W}^{IE}$ are set to 0. Shading: mean ± SD, n = 200 output neurons. Parameters used are in Supplementary Table 1 of SI.

https://doi.org/10.1038/s41593-022-01225-z



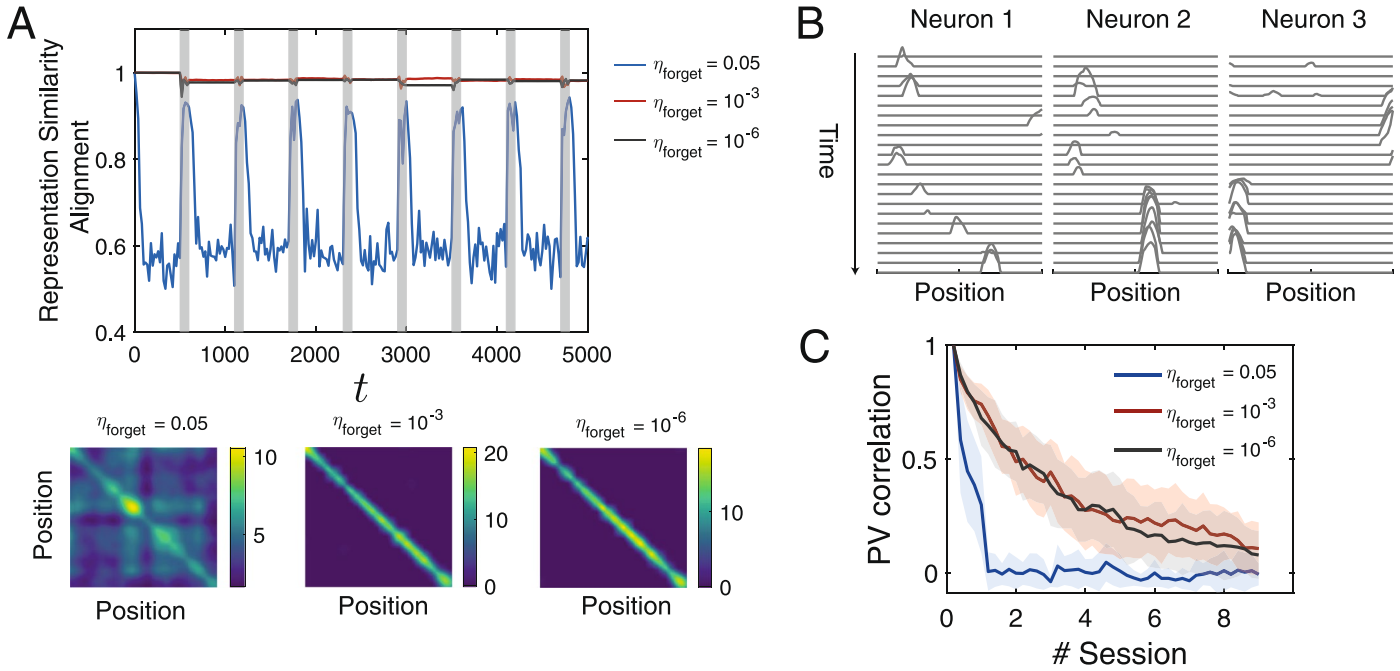
Extended Data Fig. 4 | A Hebbian/anti-Hebbian network model of CA1 with both excitatory and inhibitory neurons exhibits similar representational drift as the network in Fig. 5 of the main text. (**a**) A Hebbian/anti-Hebbian network with inhibitory neurons derived from a similarity matching objective. The derivation is given SI Section 3. (**b**) Upper: learned place fields tile a 1D linear track when sorted by their centroid positions (left), but continuously change over time (right). Lower: Representational similarity matrix $\mathbf{Y}^\top \mathbf{Y}$ of position is stable over time. (**c**) Peak amplitude of an example place field during a simulation. (**d**) Due to the drift, the average autocorrelation coefficient of population vectors decays over time. Shading: mean ± SD, n = 200 places, population vectors consist of only excitatory neurons. (**e**) Despite the continuous reconfiguration of place cell ensembles, the fraction of cells with active place fields is stable over time. (**f**) Neurons whose RFs have larger average amplitude is more stable, as characterized by smaller D. (**g**) Probability distribution of centroid drifts of place cells at three different time intervals. (**h**) Same as Fig. 5k in main text. Drifts of RFs show distance-dependent correlations, quantified by the average Pearson correlation coefficient. Shading: mean ± SD, n = 20 repeats. Error bars: mean ± SD, n = 13 animals. Parameters used are in Supplementary Table 1 of SI.

https://doi.org/10.1038/s41593-022-01225-z



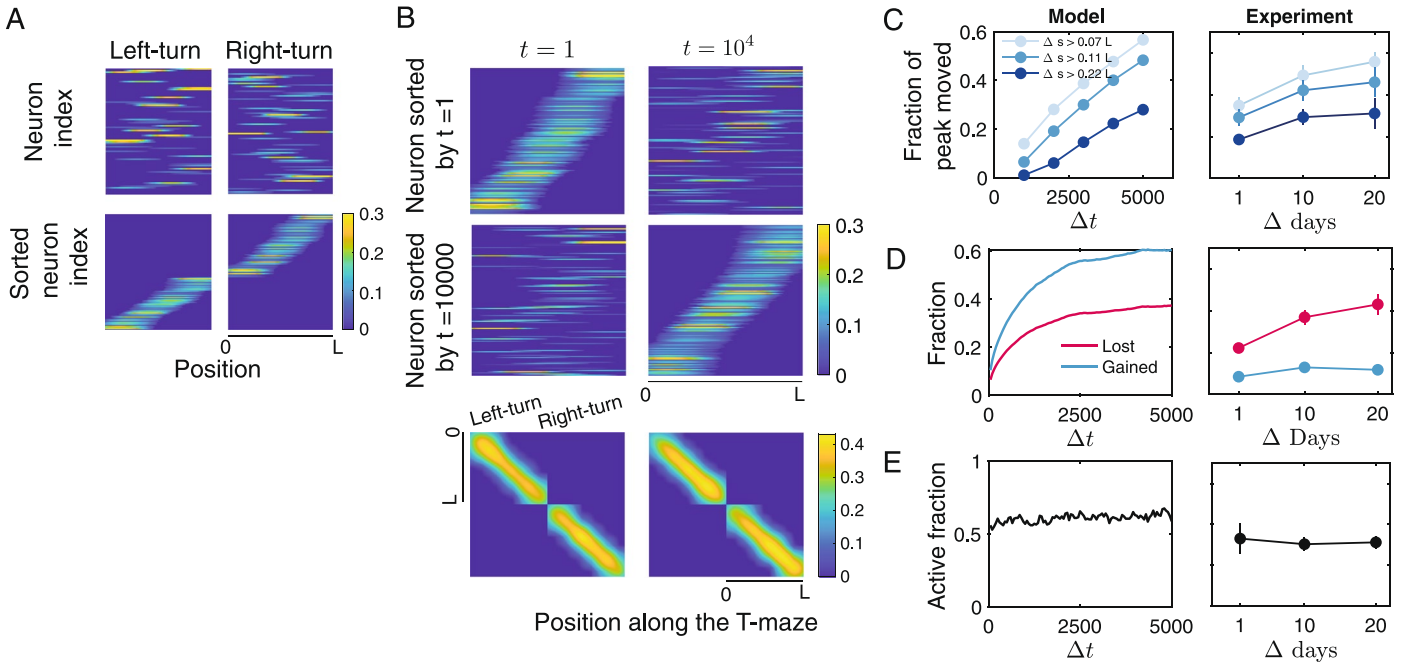
Extended Data Fig. 5 | Drift of 2D place cells in the model. (**a**) Representational similarity is preserved despite the continuous drift of place cell RFs. Positions on the plane (discretized as 32 × 32 lattice) are represented by an index from 1 to 1024. (**b**) The dynamics of RFs are intermittent. The peak amplitude of an example place field has active and silent bouts. (**c**) The intervals of silent bouts follow approximately an exponential distribution. (**d**) At the population level, there is a constant fraction of active RFs over time. (**e**) Dependence of the effective diffusion constant on the total number output neurons. Error bars: mean ± SD, n = 40 simulations. (**f**,**g**) Place cells that have stronger place fields tend to be active more often (**f**) and also more stable as indicated by smaller diffusion constant (**g**). Parameters used are in Supplementary Table 1 of SI.

https://doi.org/10.1038/s41593-022-01225-z



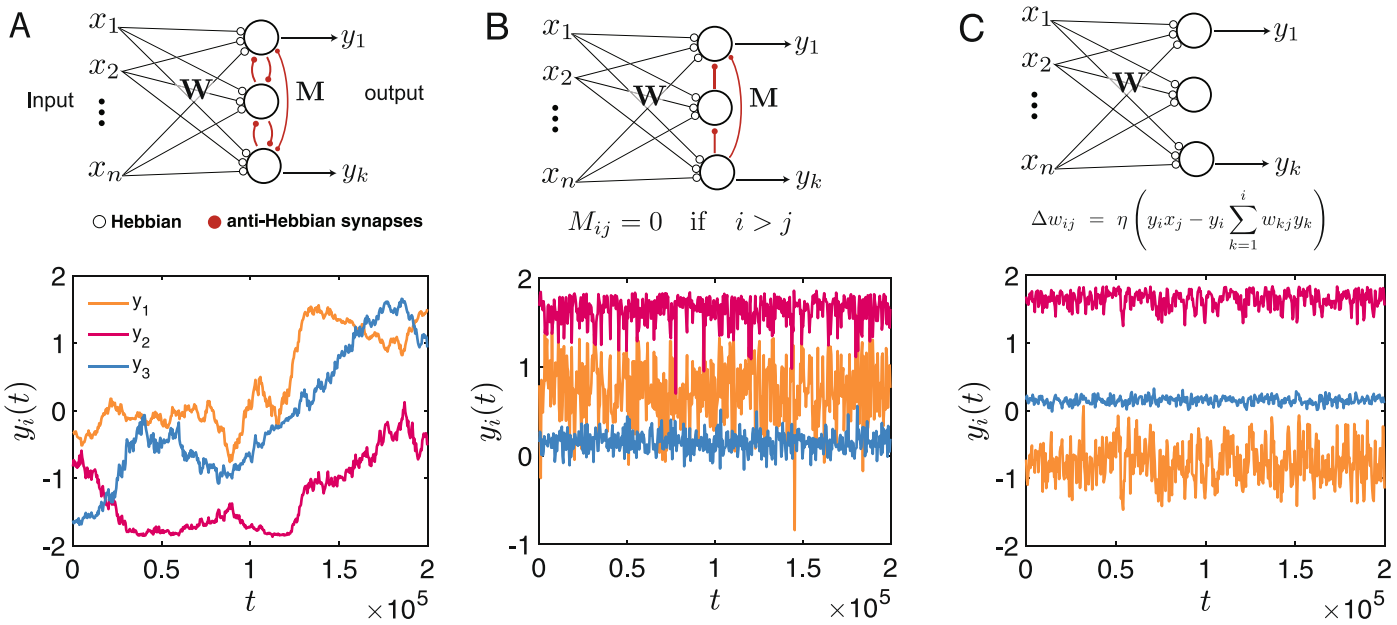
Extended Data Fig. 6 | Representational drift in a modified 1D place cell model with alternating learning and forgetting periods. We introduced a forgetting time scale ($1/\eta_{forget}$) to our learning rules. The model is described in detail in SI Section 4. (**a**) 100 synaptic updates (shaded region) are sequentially followed by a forgetting period with 500 synaptic updates. Including a slower forgetting time scale significantly enhances the stability of learned representation as quantified by the similarity matrix alignment (RSA), defined in equation (41) of SI (upper). The representational similarity matrices $\mathbf{Y}^{\top}\mathbf{Y}$ after the last forgetting period for three different forgetting time scales (lower). (**b**) Place fields of 3 exemplar output neurons in the presence of input and synaptic noise. Time starts from when the system has fully learned the representation. (**c**) Even with slow forgetting time scale, the representation still drifts during 'experiment' sessions as shown by the decay of coefficients of population vectors across learning sessions (shaded regions in (**a**)). Parameters are listed in Supplementary Table 1 of SI. Shading: mean ± SD, n = 200 output neurons. In (**a**) and (**b**), $\eta_{forget} = 10^{-3}$.

https://doi.org/10.1038/s41593-022-01225-z



Extended Data Fig. 7 | A Hebbian/anti-Hebbian network model of the PPC with both excitatory and inhibitory neurons exhibits similar representational drift as the network in Fig. 6 of main text. (**a**) Population activity of excitatory neurons for the left-turn and right-turn task before (upper) and after (lower) sorting based on the centroids of their RFs. Only neurons that have active RFs at the given time point are shown. (**b**) Population activity drifts but representational similarity is stable over time. Activity of excitatory neurons that are active (either tuned to left turn or right turn) in the sorted time (upper and middle). Representational similarity matrix is stable for both left-turn and right-turn task (lower panels). (**c**,**d**) Comparison of drift statistics between model and experiment, corresponding to panels **d**–**f** of Fig. 6 in the main text. Error bars: mean ± SD, n = 5, 5, 4 mice for Δ = 1, 10, 20 days.

https://doi.org/10.1038/s41593-022-01225-z



Extended Data Fig. 8 | Degeneracy of the learning objective function and representational drift. We compare the long-term behavior of learned representations in three different networks. (**a**) Upper: the Hebbian/anti-Hebbian network for PSP. Lower: the evolution of the three components of a representation $\boldsymbol{y}_t$. (**b**) Upper: The network differs from the Hebbian/anti-Hebbian network only in the recurrent matrix **M** which breaks the rotational symmetry of the PSP solution. The learning rule is the same. Lower: the learned representation is stabilized and only fluctuates around its equilibrium. (**c**) A single feedforward network that perform online principal component analysis with Sanger's rule[80]. This network has only feedforward input matrix W and the learning rule is nonlocal. Lower: learned representation is relatively stable in the presence of noise. Parameters are the same as in the Fig. 2 of main text except that η = 0.01.

nature portfolio

Corresponding author(s): Cengiz Pehlevan

Last updated by author(s): Sep 26, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Codes for numerical experiments were written in MATLAB (R2020b).
Data analysis	Analysis and figures were made using MATLAB (R2020b) except Fig. 4A-B, which is made by R (version 4.2.0). All codes are available in the Github repository https://github.com/Pehlevan-Group/representation-drift.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

No new data was generated for this study.
Experimental data presented in Fig. 5 are originally described in [10]. We used the processed data and MATLAB code, which are available at the Caltech Research Data Repository (https://doi.org/10.22002/d1.1229) to produce these plots.
Experimental data presented in Fig. 6 is extracted from Figure 2C and 2D of [9]. The data is freely available in [79].

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In reporting our simulations, we either average over the number of output neurons, which is specified in supplementary table 1, or independent simulations, which are specified in figure captions.
Data exclusions	No data were excluded.
Replication	Number of repeated independent simulations are specified in figure captions.
Randomization	All the numerical simulations were intrinsically randomized, each simulation started from different random number generator seed.
Blinding	N/A- This is a computational/theoretical study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

nature neuroscience

Article

https://doi.org/10.1038/s41593-022-01225-z

Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning

In the format provided by the authors and unedited

Supplementary Information for "Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning"

Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M. Sengupta, Dmitri B. Chklovskii, and Cengiz Pehlevan

I. DERIVATION OF THE ROTATIONAL DIFFUSION CONSTANT IN THE LINEAR HEBBIAN/ANTI-HEBBIAN NETWORK

In this section, we derive an analytical expression for the rotational diffusion constant defined by [1, 2]

$$D_\varphi \equiv \lim_{t\to\infty} \frac{1}{2(k-1)t}\langle|\vec{\varphi}(t)-\vec{\varphi}(0)|^2\rangle, \tag{1}$$

where k is the dimension of the space in which rotation occurs, and brackets mean averaging over different realizations of the noise. $\vec{\varphi}(t)$ is a measure of angular displacement defined as follows. We assume that to arrive at the data cloud at time step i, the data points at time $i-1$ are rotated by the rotation matrix $\mathbf{R}=\exp(-\Delta\vec{\varphi}_i(t)\cdot\vec{\mathbf{L}})$, where $\vec{\mathbf{L}}$ are the $k\times k$ infinitesimal rotation generators [3]. These generators are given by 1) when $k>3$, $(\mathbf{L}_{(mn_m)})_{ij}=\delta_{mi}\delta_{n_mj}-\delta_{mj}\delta_{n_mi}$. Here, $\{i,j\}\in\{1,\dots k\}$ are the matrix element indices; $m\in\{1,\dots,k\}$ and $n_m\in\{1,\dots,(m-1)\}$ label the $k(k-1)/2$ generators, 2) when $k=3$, $(\mathbf{L}_i)_{jk}=\epsilon_{ijk}$, where $\{i,j,k\}\in\{1,2,3\}$ and ϵ_{ijk} is the completely antisymmetric tensor, and 3) when $k=2$, $\mathbf{L}=\begin{pmatrix}0&1\\-1&0\end{pmatrix}$. We define

$$\Delta\vec{\varphi}_i=\vec{\varphi}(i)-\vec{\varphi}(i-1),\quad \vec{\varphi}(t)\equiv\vec{\varphi}(0)+\sum_{i=1}^{t}\Delta\vec{\varphi}(i). \tag{2}$$

Obtaining an exact expression for D_φ is difficult, but we were able to derive an approximation that matches numerical experiments well, as shown in Fig. 2 E and F of main text. Our approach relies on two simplifications. Note that

$$\langle|\vec{\varphi}(t)-\vec{\varphi}(0)|^2\rangle=\sum_{i=1}^{t}\langle|\Delta\vec{\varphi}_i|^2\rangle+\sum_{i=1}^{t}\sum_{j=1,i\neq j}^{t}\langle\Delta\vec{\varphi}_i\cdot\Delta\vec{\varphi}_j\rangle. \tag{3}$$

We assume that the correlation between angular displacements at different times is negligible. Therefore, we approximate

$$D_\varphi\approx\lim_{t\to\infty}\frac{1}{2(k-1)t}\sum_{k=1}^{t}\langle|\Delta\vec{\varphi}_i|^2\rangle. \tag{4}$$

Second, we assume that the network weights start at a configuration that is already an optimal solution to the similarity matching objective, projecting the input to its principal subspace, and the drift keeps the weights in the optimal solution space. This is a reasonable approximation because of a linear stability analysis presented in [4, 5]. We now review that argument.

We first define the feature map $\mathbf{F}=\mathbf{M}^{-1}\mathbf{W}$, which relates the output to input at the fixed point of the network dynamics defined by equation (2) in the main text. In other words, when the dynamics converges, $\mathbf{y}_t=\mathbf{F}\mathbf{x}_t$. We refer to rows of $\mathbf{F}$ as neural filters. We refer to an optimal solution of the similarity matching problem in the offline setting without noise as a fixed point, and denote it with $\hat{}$. We note that $\hat{\mathbf{M}}$ is symmetric, which we will use throughout. It was shown in [5] that $\hat{\mathbf{F}}\hat{\mathbf{F}}^\top=\mathbf{I}$ and $\hat{\mathbf{F}}$ projects to the principal subspace.

We note that a general perturbation of feature map $\delta\mathbf{F}$ around a fixed point $\hat{\mathbf{F}}=\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}}$ can be decomposed as

$$\delta\mathbf{F}=\delta\mathbf{A}\hat{\mathbf{F}}+\delta\mathbf{S}\hat{\mathbf{F}}+\delta\mathbf{B}\hat{\mathbf{G}}, \tag{5}$$

where $\delta\mathbf{A}$ is a $k\times k$ antisymmetric matrix, $\delta\mathbf{S}$ is a $k\times k$ symmetric matrix, $\delta\mathbf{B}$ is a $k\times(n-k)$ matrix, and $\hat{\mathbf{G}}$ is a $(n-k)\times n$ matrix with orthonormal rows. These rows are chosen to be orthogonal to the rows of $\mathbf{F}$ [5]. So we have $\delta\mathbf{A}+\delta\mathbf{S}=\delta\mathbf{F}\hat{\mathbf{F}}^\top$. The first term in (5) corresponds to a rotation of the neural filter basis in the principal subspace, the second term captures deviations from orthogonality of the basis vectors within the subspace, and the third term captures perturbations of the weight vectors that lead to projecting outside the principal subspace. As shown in [5],

the fixed point is stable to the perturbations due to the second and third terms, meaning they decay exponentially to zero, making a principal subspace projection linearly stable. Therefore, we consider drift due to the first term, which rotates neural filters and, in turn, the data cloud. Then $\delta\mathbf{A} = -\Delta\vec{\varphi}(t)\cdot\vec{\mathbf{L}}$, which follows from $\mathbf{y} = \mathbf{F}\mathbf{x}$. In this setup, $\langle|\Delta\vec{\varphi}_i|^2\rangle$ is independent of time step i (see below). Therefore, our final approximation is

$$D_\varphi \approx \frac{1}{2(k-1)}\langle|\Delta\vec{\varphi}|^2\rangle, \tag{6}$$

where $\Delta\vec{\varphi}$ arises from a noisy synaptic update to the network with an optimal set of synapses. This quantity is called mean squared angular displacement (MSAD). This approximation turns out to match simulations very well as shown in Fig. 2 E and F in the main text.

Next, we calculate D_φ. In the linear Hebbian/anti-Hebbian network for the principal subspace projection task, the learning rule with synaptic noise is

$$\Delta\mathbf{W} = \eta(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}) + \boldsymbol{\xi}_t^W, \quad \Delta\mathbf{M} = \eta(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}) + \boldsymbol{\xi}_t^M, \tag{7}$$

where $\langle\xi_{ij,t}^W\rangle = \langle\xi_{ij,t}^M\rangle = 0$ and $\langle(\xi_{ij,t}^W\xi_{kl,t'}^W\rangle = \eta\sigma_1^2\delta_{ik}\delta_{jl}\delta_{tt'}$, $\langle\xi_{ij,t}^M\xi_{kl,t'}^M\rangle = \eta\sigma_2^2\delta_{ik}\delta_{jl}\delta_{tt'}$.

By estimating the variance of the rotation of the learned representation during a single-step update under rule (7), we can define an effective rotational diffusion constant that is related to this variance. More specifically, in the small update and noise regime, $\delta\mathbf{A}$ is related to an infinitesimal rotation, $\mathbf{R}$, of the output vectors, by $\mathbf{R} = \exp(\delta\mathbf{A}) = \exp(-\Delta\vec{\varphi}\cdot\vec{\mathbf{L}})$, where $\vec{\mathbf{L}}$ are the infinitesimal rotation generators [3].

We start by writing $\delta\mathbf{F}$ in terms of the perturbations of $\hat{\mathbf{W}}, \hat{\mathbf{M}}$:

$$\delta\mathbf{F} = \hat{\mathbf{M}}^{-1}\delta\mathbf{W} - \hat{\mathbf{M}}^{-1}\delta\mathbf{M}\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}} = \hat{\mathbf{M}}^{-1}(\delta\mathbf{W} - \delta\mathbf{M}\hat{\mathbf{F}}). \tag{8}$$

Right-multiplying (8) by $\hat{\mathbf{F}}^\top$ and using (7), we have

$$\delta\mathbf{F}\hat{\mathbf{F}}^\top = \hat{\mathbf{M}}^{-1}(\delta\mathbf{W}\hat{\mathbf{F}}^\top - \delta\mathbf{M}) = \hat{\mathbf{M}}^{-1}\left(\eta(\mathbf{y}\mathbf{x}_t^\top - \hat{\mathbf{W}})\hat{\mathbf{F}}^\top - \eta(\mathbf{y}\mathbf{y}^\top - \hat{\mathbf{M}}) + \boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \boldsymbol{\xi}^M\right) = \hat{\mathbf{M}}^{-1}(\boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \boldsymbol{\xi}^M), \tag{9}$$

where we have used the property $\hat{\mathbf{F}}\hat{\mathbf{F}}^\top = \mathbf{I}$ and

$$\hat{\mathbf{M}}^{-1}\left((\mathbf{y}_t\mathbf{x}_t^\top - \hat{\mathbf{W}})\hat{\mathbf{F}}^\top - (\mathbf{y}_t\mathbf{y}_t^\top - \hat{\mathbf{M}})\right) = \hat{\mathbf{M}}^{-1}\left(\mathbf{y}_t\mathbf{y}_t^\top - \hat{\mathbf{W}}\hat{\mathbf{F}}^\top - \mathbf{y}_t\mathbf{y}_t^\top + \hat{\mathbf{M}}\right) = -\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}}\hat{\mathbf{F}}^\top + \mathbf{I} = \mathbf{0}. \tag{10}$$

Now, $\delta\mathbf{A} = \frac{1}{2}(\delta\mathbf{F}\hat{\mathbf{F}}^\top - \hat{\mathbf{F}}\delta\mathbf{F}^\top)$ can be written down explicitly:

$$\delta\mathbf{A} = \frac{1}{2}\left[\left(\hat{\mathbf{M}}^{-1}\boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \hat{\mathbf{F}}\boldsymbol{\xi}^{W^\top}\hat{\mathbf{M}}^{-1}\right) + \left(\boldsymbol{\xi}^{M^\top}\hat{\mathbf{M}}^{-1} - \hat{\mathbf{M}}^{-1}\boldsymbol{\xi}^M\right)\right]. \tag{11}$$

The mean squared angular displacement (MSAD) is related to $\delta\mathbf{A}$. To see this more clearly, note that $\delta\mathbf{A} = -\Delta\varphi\hat{\mathbf{n}}\cdot\vec{\mathbf{L}}$, where $\Delta\varphi$ is the magnitude of the rotation and $\hat{\mathbf{n}}$ is a unit vector pointing along $\Delta\vec{\varphi}$. From here, using the definiton of generators, it follows that

$$\mathrm{Tr}(\delta\mathbf{A}\delta\mathbf{A}^\top) = 2(\Delta\varphi)^2. \tag{12}$$

Also, the variance of δA_{ij} is

$$\langle\delta A_{ij}^2\rangle = \frac{\eta}{4}(\sigma_1^2 + \sigma_2^2)\sum_k\left(\tilde{M}_{kj}^2 + \tilde{M}_{ki}^2 - 2\delta_{ij}\tilde{M}_{ki}\tilde{M}_{kj}\right), \tag{13}$$

where $\tilde{\mathbf{M}} \equiv \hat{\mathbf{M}}^{-1}$, the average $\langle\rangle$ is over the noise distribution, and we used $\hat{\mathbf{F}}\hat{\mathbf{F}}^\top = \mathbf{I}$ to simplify. Using the fact that $\mathrm{eig}(\hat{\mathbf{M}}) = [\lambda_1, \cdots, \lambda_\mathrm{k}]$, which was shown in [5], and $\mathrm{Tr}\tilde{\mathbf{M}}^2 = \sum_{i=1}^k 1/\lambda_i^2$, we have

$$\langle\mathrm{Tr}\delta\mathbf{A}\delta\mathbf{A}^\top\rangle = \sum_{ij}\langle\delta A_{ij}^2\rangle = \frac{1}{2}\eta(k-1)(\sigma_1^2 + \sigma_2^2)\sum_{i=1}^k\frac{1}{\lambda_i^2}. \tag{14}$$

Given our approximation of considering only a single-step, it follows from Eq. (1) that $\langle|(\Delta\varphi)^2|\rangle = 2(k-1)D_\varphi$. With Eq.(12) and Eq.(14) and $\sigma_1 = \sigma_2 = \sigma$, we arrive at Eq. 4 in the main text. Note that $k = 3$ in Figure 2.

II. DERIVATION OF THE EFFECTIVE DIFFUSION CONSTANT IN THE RING MODEL

Here, we calculate the diffusion constant in the ring model for a single output neuron. We again consider the approximation that the diffusion constant can be approximated by the mean squared displacement around a fixed point by a noisy synaptic update.

The position on the ring is represented by the input vectors $\mathbf{x} = [\cos\theta, \sin\theta]^\top$. The response of the neuron is given by

$$y(\theta) = \frac{1}{m+\beta}[w_1 \cos\theta + w_2 \sin\theta - \alpha b]_+. \tag{15}$$

Here and after, we use $[x]_+$ to denote the rectified linear function. We define a steady state where the average update to the weights is zero. Denoting the stationary state weights as $\{w_1^*, w_2^*, m^*, b^*\}$, this leads to the conditions:

$$w_1^* = \langle y(\theta)\cos\theta\rangle_\theta, \quad w_2^* = \langle y(\theta)\sin\theta\rangle_\theta, \quad m^* = \langle y^2(\theta)\rangle_\theta, \quad b^* = \alpha\langle y(\theta)\rangle_\theta, \tag{16}$$

where $\langle\cdot\rangle_\theta$ means averaging over the input distribution $\theta \in [-\pi, \pi)$ which we assume to be uniform.

To solve these equations, we use an ansatz of the form

$$y_\phi(\theta) = \mu[\cos(\theta-\phi) - \cos(\psi)]_+, \tag{17}$$

where $\psi \geq 0$ determines the width of the RF, $\mu(1-\cos\psi)$ is the peak amplitude, and ϕ is the centroid of the receptive field. We solve for the parameters of the ansatz self-consistently. Due to the symmetry of the problem, any ϕ gives a plausible solution. Plugging (17) into (15) and (16), we find that

$$w_1^* = \frac{\mu}{4\pi}(2\psi - \sin 2\psi)\cos\phi, \tag{18}$$

$$w_2^* = \frac{\mu}{4\pi}(2\psi - \sin 2\psi)\sin\phi, \tag{19}$$

$$m^* = \frac{\mu^2}{4\pi}(4\psi + 2\psi\cos 2\psi - 3\sin 2\psi), \tag{20}$$

$$b^* = \frac{\alpha\mu}{\pi}(\sin\psi - \psi\cos\psi). \tag{21}$$

Equation (15) can be rewritten as

$$y(\theta) = \frac{\sqrt{w_1^2+w_2^2}}{m+\beta}\left[\frac{w_1}{\sqrt{w_1^2+w_2^2}}\cos\theta + \frac{w_2}{\sqrt{w_1^2+w_2^2}}\sin\theta - \frac{\alpha b}{\sqrt{w_1^2+w_2^2}}\right]_+. \tag{22}$$

Comparing to (17), we have

$$\mu = \frac{\sqrt{w_1^{*2}+w_2^{*2}}}{m^*+\beta}, \quad \alpha b^* = \sqrt{w_1^{*2}+w_2^{*2}}\cos\psi. \tag{23}$$

Combining (18)-(21) and (23), we get the dependence of μ and ψ on α and β, given parametrically by

$$\mu^2 = \frac{2\psi - \sin 2\psi - 4\beta\pi}{4\psi + 2\psi\cos 2\psi - 3\sin 2\psi}, \quad \alpha^2 = \frac{\cos\psi(2\psi - \sin 2\psi)}{4(\sin\psi - \psi\cos\psi)}. \tag{24}$$

Given α and β, one can solve these equations for μ and ψ and plug them back into (18)-(21) to recover all parameters. One can also check for the self-consistency of this solution by starting from (15) and plugging in (18)-(21) and (24).

Next, we proceed to estimate the drift due to noisy synaptic updates. First, to simplify the following calculations, using (22) we define

$$\hat{\mu} \equiv w_1^*\cos\phi + w_2^*\sin\phi = \sqrt{w_1^{*2}+w_2^{*2}} = \frac{\mu}{4\pi}(2\psi - \sin 2\psi). \tag{25}$$

Next, we want to estimate how the centroid changes. We define the centroid ϕ to be the angle at which the response is maximum and positive. Then, $dy(\theta)/d\theta = 0$ at $\theta = \phi$. This implies $\tan(\phi) = w_2/w_1$. Note that this is true for

any set of weights, not only $\{w_1^*, w_2^*\}$. We can then approximate the change in the centroid $\Delta\phi$ due to a small weight change from the configuration $\{w_1^*, w_2^*\}$ by:

$$\Delta\phi = \frac{\cos^2\phi}{w_1^{*2}}(\Delta w_2 w_1^* - \Delta w_1 w_2^*) = \frac{1}{\hat{\mu}}\left(\Delta w_2 \cos\phi - \Delta w_1 \sin\phi\right), \tag{26}$$

where we have used the fact $w_1^* = \hat{\mu}\cos\phi$ and $w_2^* = \hat{\mu}\sin\phi$ from Eq.(25). In particular, we are interested in how the centroid of the RF changes when a perturbation is added to the stationary weight vector

$$\Delta w_1 = \eta(y(\theta)\cos\theta - w_1^*) + \xi_1, \tag{27}$$
$$\Delta w_2 = \eta(y(\theta)\sin\theta - w_2^*) + \xi_2. \tag{28}$$

By their definition, the statistics of the Gaussian white noise terms are: $\langle\xi_1\rangle = \langle\xi_2\rangle = 0$, $\langle\xi_1^2\rangle = \langle\xi_2^2\rangle = \eta\sigma^2$. From (18), (19) and (25), we also have $w_1^* = \hat{\mu}\cos\phi$ and $w_2^* = \hat{\mu}\sin\phi$. Then $\Delta\phi$ can be written as

$$\begin{aligned}\Delta\phi &= \frac{1}{\hat{\mu}}\left(\eta\mu[\cos(\theta-\phi)-\cos\psi]_+\sin\theta\cos\phi - \eta\mu[\cos(\theta-\phi)-\cos\psi]_+\cos\theta\sin\phi + (\xi_2\cos\phi - \xi_1\sin\phi)\right)\\ &= \frac{1}{\hat{\mu}}\{\eta\mu[\cos(\theta-\phi)-\cos\psi]_+\sin(\theta-\phi) + (\xi_2\cos\phi - \xi_1\sin\phi)\}.\end{aligned} \tag{29}$$

Since in online learning θ is sampled randomly, we can average over θ to get mean squared displacement

$$\langle(\Delta\phi)^2\rangle = \frac{\mu^2}{\hat{\mu}^2}\eta^2\langle([\cos(\theta-\phi)-\cos\psi]_+^2\sin^2(\theta-\phi))\rangle_\theta + \frac{1}{\hat{\mu}^2}(\cos^2\phi\langle\xi_2^2\rangle + \sin^2\phi\langle\xi_1^2\rangle) = \gamma\eta^2 + \frac{\eta\sigma^2}{\hat{\mu}^2}, \tag{30}$$

where

$$\gamma \equiv \frac{\mu^2}{\hat{\mu}^2}\frac{1}{2\pi}\int_{-\pi}^{\pi}[\cos(\theta-\phi)-\cos\psi]_+^2\sin^2(\theta-\phi)d\theta = \frac{\pi}{6}\frac{36\psi + 24\psi\cos(2\psi) - 28\sin(2\psi) - \sin(4\psi)}{(2\psi-\sin(2\psi))^2}. \tag{31}$$

Now, to calculate the diffusion constant, we again resort to approximation. We assume that the diffusion constant can be obtained from a single-step update, $\langle(\Delta\phi)^2\rangle \approx 2D$. Then, we have

$$D \approx \frac{1}{2}\left(\gamma\eta^2 + \frac{\eta\sigma^2}{\hat{\mu}^2}\right). \tag{32}$$

Since $\gamma, \hat{\mu}$ depend on the tuning width ψ of the RF, the dependence of D on the peak amplitude $\mu(1-\cos(\psi))$ is complicated. Numerical simulations shows that neurons with larger RFs typically have smaller D (Extended Data Fig. 2D).

As a final note, we want to point that tracking peak activity is numerically prone to noise in simulations. Therefore, we track center of mass of RFs. However, due to the symmetry of RFs, these two metrics largely coincide.

III. DERIVATION OF THE HEBBIAN/ANTI-HEBBIAN NEURAL NETWORK WITH INHIBITORY NEURONS

In the main text, we presented Hebbian/anti-Hebbian networks with only principal cells for simplicity. The mutual inhibition among these principal neurons violates Dale's law because in the brain principal neurons are mostly excitatory. Here, we derive a Hebbian/anti-Hebbian networks with inhibitory neurons from a minimax similarity matching objective

$$\min_{\forall t\in\{1,\dots,T\}:\mathbf{y}_t\geq 0}\ \max_{\forall t\in\{1,\dots,T\}:\mathbf{z}_t\geq 0}\ \frac{1}{2T^2}\sum_{t,t'}\left[\left(\mathbf{x}_t^\top\mathbf{x}_{t'} - \mathbf{y}_t^\top\mathbf{y}_{t'} - \alpha^2\right)^2 - \left(\mathbf{y}_t^\top\mathbf{y}_{t'} - \mathbf{z}_t^\top\mathbf{z}_{t'}\right)^2\right] + \frac{1}{T}\sum_t^T\left(2\beta_1\|\mathbf{y}_t\|_1 + \beta_2\|\mathbf{y}_t\|_2^2\right). \tag{33}$$

Here $\mathbf{y}_t$ are N_E-dimensional (nonnegative) vectors and $\mathbf{z}_t$ are N_I-dimensional (nonnegative) vectors. These vectors will map to activations of excitatory and inhibitory neurons respectively. Notice that if the rank of $\mathbf{Z} \equiv [\mathbf{z}_1, \cdots, \mathbf{z}_T]$ is larger than the rank of $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_T]$ at the optimum, which would always be true if $N_I \geq N_E$, optimal solutions satisfy $\mathbf{y}_t^\top\mathbf{y}_{t'} - \mathbf{z}_t^\top\mathbf{z}_{t'} = 0$, and the above objective function is equivalent to the Nonnegative Similarity Matching one in the main text (Eq. 15). If not, Eq. (33) can be seen as an approximation to (Eq. 15) in the main text. In Extended

Data Fig. 4, we show a simulation where $N_I < N_E$, and yet the principal neurons still learn localized receptive fields and show drift.

In the online setting, we can follow a similar procedure to the one sketched in the Methods of the main text and consider a dual problem to the above objective function by 1) introducing auxiliary variables $\mathbf{W}^{\mathrm{X}}$, $\mathbf{M}$, and $\mathbf{W}^{\mathrm{IE}}$ through the same procedure as in Eqs. 6 and 7 of the main text, and 2) changing orders of optimization:

$$\min_{\mathbf{W}^{\mathrm{X}}}\min_{\mathbf{M}}\max_{\mathbf{W}^{\mathrm{IE}}}\max_{\mathbf{b}}\frac{1}{T}\sum_{t=0}^{T}\left[\mathrm{Tr}\,\mathbf{W}^{\mathrm{X}^\top}\mathbf{W}^{\mathrm{X}}-\mathrm{Tr}\,\mathbf{W}^{\mathrm{IE}^\top}\mathbf{W}^{\mathrm{IE}}+\frac{1}{2}\mathrm{Tr}\mathbf{M}^\top\mathbf{M}-\|\mathbf{b}\|_2^2+\min_{\mathbf{y}_t\geq 0}\max_{\mathbf{z}_t\geq 0}\mathrm{l}_t(\mathbf{y}_t,\mathbf{z}_t,\mathbf{W}^{\mathrm{X}},\mathbf{W}^{\mathrm{IE}},\mathbf{M},\mathbf{b})\right] \tag{34}$$

where

$$l_t\equiv -2\mathbf{y}_t^\top\mathbf{W}^X\mathbf{x}_t+2\alpha\mathbf{y}_t^\top\mathbf{b}+2\mathbf{z}_t^\top\mathbf{W}^{\mathrm{IE}}\mathbf{y}_t-\mathbf{z}_t^\top\mathbf{M}\mathbf{z}_t+2\beta_1\|\mathbf{y}_t\|_1+\beta_2\|\mathbf{y}_t\|_2^2. \tag{35}$$

We solve this dual optimization problem again in an online manner. For each input, first, we optimize l_t by a neural minimax dynamics [6, 7]. To see how we do this, consider the following. If we treat $\mathbf{y}_t$ as constant, then an inhibitory neuron dynamics which maximizes l_t is given by [8, 9]:

$$\tau_v\dot{v}_i=-v_i+\sum_j W_{ij}^{IE}y_j-\sum_{j\neq i}\bar{M}_{ij}z_j,\quad z_i=\left[\frac{v_i}{M_{ii}}\right]_+. \tag{36}$$

Similarly, for fixed $\mathbf{z}_t$, the following excitatory neuron dynamics minimizes l_t [8, 9]:

$$\tau_u\dot{u}_i=-u_i+\sum_j W_{ij}^X x_j-\sum_k W_{ik}^{EI}z_k-\alpha b_i,\quad y_i=\left[\frac{u_i-\beta_1}{\beta_2}\right]_+. \tag{37}$$

We define a heuristic optimization algorithm by running these equations simultaneously. In simulations, we observe that fast inhibitory neuron dynamics helps with convergence, consistent with the max operator appearing inside the min in the $\mathbf{y},\mathbf{z}$ optimization. In particular, in our simulations of Extended Data Figures 4 and 7, we used instantaneous inhibitory neurons, but observed that stable dynamics could be achieved for $\tau_v/\tau_u \lessapprox 1/100$. We note that such instantaneous neurons were used before in the literature, for example in [10, 11].

Second, we do gradient updates to the synaptic matrices and biases

$$\begin{aligned}\Delta\mathbf{W}^X&=\eta(\mathbf{y}\mathbf{x}^\top-\mathbf{W}^X),\\ \Delta\mathbf{W}^{EI}&=\eta(\mathbf{y}\mathbf{z}^\top-\mathbf{W}^{EI}),\\ \Delta\mathbf{W}^{IE}&=\eta(\mathbf{z}\mathbf{y}^\top-\mathbf{W}^{IE}),\\ \Delta\mathbf{M}&=\eta(\mathbf{z}\mathbf{z}^\top-\mathbf{M}),\\ \Delta\mathbf{b}&=\eta(\alpha\mathbf{y}-\mathbf{b}).\end{aligned} \tag{38}$$

We modeled the hippocampal place cell formation in a 1D linear track environment using the above neural dynamics and learning rules with independent random Gaussian noise. All the drifting dynamics and statistics are similar to the simplified model without inhibitory neurons (Extended Data Fig. 4).

IV. MODIFIED MODEL WITH A SLOW FORGETTING TIMESCALE

In the main text, we considered the learning process only in the presence of sensory inputs. Animals can retain their learned memories during long periods when such sensory inputs are absent. If our model's synapses were updated during these latter periods in a way independent of the task-relevant sensory variables, then the task-relevant receptive fields would be rapidly forgotten. To resolve this, we note that memory processes in the brain operate at a spectrum of timescales [12, 13]. By introducing a slow synaptic timescale that corresponds to a slower forgetting process as in [14, 15], we show that learned representations in Hebbian/anti-Hebbian networks can be maintained for a long time even in the absence of sensory inputs.

Our modified learning rules are the following:

$$\begin{aligned}\Delta\mathbf{W}&=\underbrace{-\eta_{\mathrm{forget}}\mathbf{W}+\sqrt{\eta_{\mathrm{forget}}\sigma^2}\boldsymbol{\zeta}_t^W}_{\text{forgetting}}+\underbrace{\eta(\mathbf{y}_t\mathbf{x}_t^\top-\mathbf{W})+\sqrt{\eta\sigma^2}\boldsymbol{\xi}_t^W}_{\text{sensory input}},\\ \Delta\mathbf{M}&=-\eta_{\mathrm{forget}}\mathbf{M}+\sqrt{\eta_{\mathrm{forget}}\sigma^2}\boldsymbol{\zeta}_t^M+\eta(\mathbf{y}_t\mathbf{y}_t^\top-\mathbf{M})+\sqrt{\eta\sigma^2}\boldsymbol{\xi}_t^M,\\ \Delta\mathbf{b}&=-\eta_{\mathrm{forget}}\mathbf{b}+\eta(\alpha\mathbf{y}_t-\mathbf{b}).\end{aligned} \tag{39}$$

The first two terms in the right hand side of the first equation above represent the forgetting process, while the last two terms represent learning in the presence of sensory input. The other rules are modified similarly. $\zeta^A_{ij,t}, \xi^A_{ij,t}, A \in \{W, M\}$ are unit Gaussian random variables with mean 0 and standard deviation 1. Both types of processes come with their own noise terms reflecting that these processes are different. In the absence of sensory inputs, the synaptic weights undergo a purely forgetting process:

$$\begin{aligned}\Delta\mathbf{W} &= -\eta_{\text{forget}}\mathbf{W} + \sqrt{\eta_{\text{forget}}\sigma^2}\boldsymbol{\zeta}^W_t \\ \Delta\mathbf{M} &= -\eta_{\text{forget}}\mathbf{M} + \sqrt{\eta_{\text{forget}}\sigma^2}\boldsymbol{\zeta}^M_t \\ \Delta\mathbf{b} &= -\eta_{\text{forget}}\mathbf{b}.\end{aligned} \tag{40}$$

In these equations, $1/\eta_{forget}$ sets the time scale for number of synaptic updates that lead to decay of memories.

In the numerical simulation shown in Extended Data Fig. 6, we start by learning representations of sensory inputs using the above learning rules. We find that the system quickly develops place cell like representations and these representations drift over time in the presence of synaptic noise (Extended Data Fig. 6A). We next simulate the network with alternating learning (shaded region) and forgetting sessions. In each learning session, 100 randomly selected inputs are presented to the network while the network weight matrices $\mathbf{W}, \mathbf{M}$ and $\mathbf{b}$ are updated with rule (39). Then, the network goes through a forgetting session with 500 iterations without sensory input in which the synaptic weights undergo a biased random walk based on (40). We quantify how the representation similarity matrix changes during this process with the "Representational Similarity Alignment" (RSA) metric [16]:

$$\text{RSA} = \frac{\|\mathbf{Y}_t\mathbf{Y}_0^\top\|_F^2}{\|\mathbf{Y}_0^\top\mathbf{Y}_0\|_F\|\mathbf{Y}_t^\top\mathbf{Y}_t\|_F}, \tag{41}$$

where $\mathbf{Y}_0$ and $\mathbf{Y}_t$ are the output matrices at time 0 and t respectively. An RSA with 1 means a perfect alignment and 0 means totally orthogonal representations. We computed RSA with different forgetting timescales and found that the network with slower forgetting timescale (smaller η_{forget}) has a much longer memory of representations, as indicated by higher value of RSA and slower decay over time (Extended Data Fig. 6A). Longer sequences of forgetting trials can be tolerated by the network by appropriate choice of η_{forget}. When focusing on the data from only the learning sessions (similar to experimental data acquisition), we observed a decay of correlation coefficients of population vectors consistent with experimental observations (Extended Data Fig. 6C). Together, these results support the idea that slow forgetting timescale of synaptic plasticity can retain the learned representations/memories in the absence of sensory input while still produce observed representational drift.

V. SIMULATION PARAMETERS

We collect in Table I all the simulation parameters. The MATLAB code for our simulations is available in the Github repository: https://github.com/Pehlevan-Group/representation-drift.

In the simulation of PSP task Fig. 2E of the main text, the first 3 eigenvalues of the input covariance matrix $\mathbf{C}$ are $3.1, 3.1, 3.1$ and the rest are set to be 0.01. In Fig. 2F of the main text, the first 3 eigenvalues are randomly sampled from a log-normal distribution and normalized such that the summation of all the eigenvalues is 10.

TABLE I. Simulation parameters used in the figures

	N_{out} (or n)	N_{in} (or k)	η	α^2	σ	β_1	β_2	λ	N_l	N_θ	N_x	N_y	l_0
Fig. 2A-D	3	10	0.05	-	0.01	-	-	-	-	-	-	-	-
Fig. 2E	3	10	0.1	-	0.01	-	-	-	-	-	-	-	-
Fig. 2F	3	10	0.05	-	0.01	-	-	-	-	-	-	-	-
Fig. 3B,C	5	2	0.05	0	0.02	0	0.02	-	-	-	-	-	-
Fig. 3D-F Fig. 4A,B	200	2	0.02	0	0.001	0	0.01	-	-	-	-	-	-
Fig. 4C,D	100	2	$[10^{-3}, 10^{-1}]$	0	0.01	0	0.05	-	-	-	-	-	-
Fig. 4E	-	2	0.01	0	0.01	0	0.05	-	-	-	-	-	-
Fig. 4F	200	2	0.05	0	0.002	0	0.05	-	-	-	-	-	-
Fig. 5B,C	200	750	0.005	95	0.02	0.02	0.05	1.42	5	6	5	5	$0.2L$
Fig. 5D-I	200	270	0.05	60	0.005	0.02	0.05	1.42	5	6	3	3	$0.25L$
Fig. 6	300	3	0.01	1.5	4×10^{-5}	2×10^{-5}	10^{-3}	-	-	-	-	-	-
ED Fig. 2B	1	2	0.05	0	0	0	0	0	-	-	-	-	-
ED Fig. 2C	1	2	0.01	0	-	0	0	0	-	-	-	-	-
ED Fig. 3A	200	480	0.02	65	0.05	0.04	0.01	1.42	5	6	4	4	$0.25L$
ED Fig. D4,3B	$N_E = 200, N_I = 20$	480	0.02	60	0.05	0.01	0.05	1.42	5	6	4	4	$0.25L$
ED Fig. 5E	200	750	0.005	95	0.008	0.01	0.05	1.42	5	6	5	5	$0.2L$
ED Fig, 6	200	270	0.05	35	0.005	0.0	0.05	1.42	5	6	3	3	$0.25L$
ED Fig. 7	$N_E = 200, N_I = 20$	3	0.03	1.5	3×10^{-4}	5×10^{-6}	10^{-3}	-	-	-	-	-	-
ED Fig. 8	3	10	0.01	-	0.01	-	-	-	-	-	-	-	-

[1] S. Kämmerer, W. Kob, and R. Schilling, Dynamics of the rotational degrees of freedom in a supercooled liquid of diatomic molecules, Physical Review E **56**, 5450 (1997).

[2] M. G. Mazza, N. Giovambattista, F. W. Starr, and H. E. Stanley, Relation between rotational and translational dynamic heterogeneities in water, Physical Review Letters **96**, 057803 (2006).

[3] A. Zee, *Group theory in a nutshell for physicists*, Vol. 17 (Princeton University Press, 2016).

[4] C. Pehlevan and D. Chklovskii, A normative theory of adaptive dimensionality reduction in neural networks, in *Advances in neural information processing systems* (2015) pp. 2269–2277.

[5] C. Pehlevan, A. M. Sengupta, and D. B. Chklovskii, Why do similarity matching objectives lead to hebbian/anti-hebbian networks?, Neural computation **30**, 84 (2018).

[6] H. S. Seung, T. Richardson, J. Lagarias, and J. J. Hopfield, Minimax and hamiltonian dynamics of excitatory-inhibitory networks, Advances in neural information processing systems **10** (1997).

[7] Q. Li and C. Pehlevan, Minimax dynamics of optimally balanced spiking networks of excitatory and inhibitory neurons, Advances in Neural Information Processing Systems **33**, 4894 (2020).

[8] P. T. P. Tang, Convergence of lca flows to (c) lasso solutions, arXiv preprint arXiv:1603.01644 (2016).

[9] C. Pehlevan, A spiking neural network with local learning rules derived from nonnegative similarity matching, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019) pp. 7958–7962.

[10] S. Druckmann, T. Hu, and D. Chklovskii, A mechanistic model of early sensory processing based on subtracting sparse representations, Advances in Neural Information Processing Systems **25** (2012).

[11] A. A. Koulakov and D. Rinberg, Sparse incomplete representations: a potential role of olfactory granule cells, Neuron **72**, 124 (2011).

[12] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory., Psychological review **102**, 419 (1995).

[13] A. Roxin and S. Fusi, Efficient partitioning of memory systems and its importance for memory consolidation, PLoS computational biology **9**, e1003146 (2013).

[14] U. Rokni, A. G. Richardson, E. Bizzi, and H. S. Seung, Motor learning with unstable neural representations, Neuron **54**, 653 (2007).

[15] S. Fusi, P. J. Drew, and L. F. Abbott, Cascade models of synaptically stored memories, Neuron **45**, 599 (2005).

[16] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, On kernel-target alignment, Advances in neural information processing systems **14** (2001).