Biological Cybernetics (2022) 116:253–266
https://doi.org/10.1007/s00422-021-00916-3

Biological Cybernetics

MINI REVIEW

Drifting neuronal representations: Bug or feature?

Paul Masset[1,2] · Shanshan Qin[1,3] · Jacob A. Zavatone-Veth[1,4]

Received: 25 August 2021 / Accepted: 17 November 2021 / Published online: 7 January 2022

Abstract
The brain displays a remarkable ability to sustain stable memories, allowing animals to execute precise behaviors or recall stimulus associations years after they were first learned. Yet, recent long-term recording experiments have revealed that single-neuron representations continuously change over time, contravening the classical assumption that learned features remain static. How do unstable neural codes support robust perception, memories, and actions? Here, we review recent experimental evidence for such representational drift across brain areas, as well as dissections of its functional characteristics and underlying mechanisms. We emphasize theoretical proposals for how drift need not only be a form of noise for which the brain must compensate. Rather, it can emerge from computationally beneficial mechanisms in hierarchical networks performing robust probabilistic computations.

Keywords Representational drift · Bayesian learning · Neural networks · Representation learning

1 Introduction

A hallmark of natural intelligence is the ability to continuously learn from external sensory stimuli. Despite this flexibility, the brain achieves consistent perception and stores long-term memories, requiring it to maintain stable neuronal representations of sensory stimuli. The implicit assumption—supported by some experimental evidence—of classical theories of neural coding is that these stable representations are supported by stable single-neuron activity (Barnes et al 1997; Thompson and Best 1990; Tonegawa et al 2015)

While many neurons in sensory cortex show stable responses to simple artificial stimuli, long-term neural population responses to more naturalistic and behaviorally relevant stimuli remained largely unexplored until recent years. Technical advances have enabled researchers to perform longitudinal recordings of large neural populations in complex tasks and in freely behaving animals for weeks and months. These recordings, starting with studies in the hippocampus, have revealed that population activity in brain regions that are responsible for certain tasks changes continuously over time even after the animals have fully learned and maintained the tasks, a phenomenon termed "representational drift" (Kentros et al 2004; Mankin et al 2012; Ziv et al 2013; Rule et al 2019; Mau et al 2020; Gonzalez et al 2019; Schoonover et al 2021; Marks and Goard 2021; Deitch et al 2021; Clopath et al 2017).

At the most basic level, representational drift involves systematic changes in how neurons encode a particular stimulus or behaviorally relevant variable (Rule et al 2019). Here, we use the definition of "neural representation" employed in most experimental studies: the response of a neuron in a short window (on the order of a second) around a defined event (e.g., presentation of a specific stimulus or a position in space). Drift is said to occur if these responses change, for

Communicated by Jean-Marc Fellous.

All authors contributed equally and are listed alphabetically.

✉ Paul Masset
paul_masset@fas.harvard.edu

Shanshan Qin
ssqin@seas.harvard.edu

Jacob A. Zavatone-Veth
jzavatoneveth@g.harvard.edu

1 Center for Brain Science, Harvard University, Cambridge, MA, USA

2 Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

3 School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

4 Department of Physics, Harvard University, Cambridge, MA, USA

all other experimental conditions kept constant, at a slower timescale (days to weeks). Importantly, these changes do not simply result from trial-to-trial variability. Depending on how changes in single-neuron tuning are organized across the population, drift may preserve or disrupt aspects of representational geometry, such as the similarity of responses to different stimuli (Rule et al 2019; Kriegeskorte and Wei 2021; Qin et al 2021).

Classic theoretical work on changes in neural representations over time largely focused on synaptic weight changes due to continual learning in memory-storage areas. In a network with finite-precision synapses, synaptic updates induced by the storage of new memories act as noise on previously stored patterns, eventually leading to forgetting (Amit and Fusi 1994; Fusi and Senn 2006; Fusi and Abbott 2007; Fusi 2021). This leads to a fundamental tradeoff between stability and plasticity, in which the flexibility to memorize new patterns must be balanced with the rate of forgetting. In this view, drift of tasks learned long ago can be at once a bug and a feature: forgetting may be beneficial if the importance of a memory increases with its recency, so long as memories are forgotten elegantly rather than catastrophically (Parisi 1986; Kulhavý and Zarrop 1993; French 1999; Kirkpatrick et al 2017; Zenke et al 2017; Fusi 2021). However, these theories would predict that representations of practiced tasks on which an animal maintains a high degree of proficiency should generally remain stable. Therefore, new normative theories are required to account for drift in neural representations of fixed tasks.

Observations of representational drift naturally raise questions regarding its causes, its ubiquity across brain areas, and its computational implications. In this brief review, we first highlight recent technical advances that enable long-term recording of large populations of neurons in behaving animals, and the resulting experimental observations of representational drift in several brain areas. We then identify three theoretical proposals for how representational drift may be consistent with normative computational principles: i) drift as a signature of Bayesian sampling of the space of solutions ii) drift due to redundancy in the neural code and iii) drift due to compensation for changes in connectivity elsewhere in the network. We emphasize how these three proposals are not mutually exclusive, and can be unified in the framework of stochastic optimization. Finally, we suggest potential directions for future theoretical and experimental studies leveraging these normative principles.

2 Experimental techniques

Recent technical advances such as high-density electrodes (Jun et al 2017b; Steinmetz et al 2021; Chung et al 2019; Dimitriadis et al 2018) and two-photon calcium imaging (Svoboda and Yasuda 2006; Ahrens et al 2013; Li et al 2017) have enabled researchers to simultaneously record the activity of hundreds of neurons over days and weeks. This is further fostered by the development of automated clustering methods which allow isolation of single-neuron activity from either electrophysiological traces (Chung et al 2017; Pachitariu et al 2016; Yger et al 2018; Jun et al 2017a) (Fig. 1Bi) or imaging data (Pnevmatikakis et al 2016; Pachitariu et al 2017; Giovannucci et al 2019; Saxena et al 2020) (Fig. 1Ci), and new surgical methods which provides greater stability of implants (Musk and Neuralink 2019; Juavinett et al 2019; Luo et al 2020; Schoonover et al 2021) and imaging at depth (Attardo et al 2015; Ulivi et al 2019; Li et al 2017). Together, these techniques and methods constitute a high-throughput pipeline that enables researchers to directly probe how the tuning of individual neurons within large ensembles changes across time. These experiments have transformed our understanding of the neural code, from an explanation of representations at the level of single neurons to an explanation at the level of neural populations (Yuste 2015; Saxena and Cunningham 2019; Urai et al 2021; Ebitz and Hayden 2021).

A prerequisite for assessing the stability of tuning of single neurons is the ability to record the same neurons over days or weeks. This requires consistent assignment of neuron identities across days, a process that is complicated by instability in the electrode location or the focal plane. In electrophysiological recordings, the templates of putative single neurons across different electrodes of a probe can be matched across days and shown to be similar within neurons (Fig. 1Bii). In calcium imaging data, there is a wealth of anatomical information that can be used to correctly assign neuron identities across days (Fig. 1Cii).

However, assigning neuron identities based on the similarity of waveforms measured at discrete time points can easily miss a (potentially biased) part of the neural population. As shown in Fig. 1D, the recorded waveform of some neurons can change significantly on the timescale of hours and post hoc corrections are often needed to ensure proper identification (Dhawale et al 2017; Steinmetz et al 2021). Similarly, experiments using one-photon microscopy can suffer from poor resolution of the images and control of the focal plane, leading to a more difficult segmentation problem (Liberti et al 2016) with an increased probability of misattributing neuron identities across days (Fig. 1E). In comparison, two-photon microscopy can more reliably track the same neurons over weeks but is largely confined to head-fixed experimental setups (Fig. 1C).

As highlighted by these examples, the process of identifying neurons across days is still challenging and can be prone to errors, both false positives (different neurons labeled as the same) and false negatives (the same neuron labeled as different). If these errors are biased toward subtypes of neu-

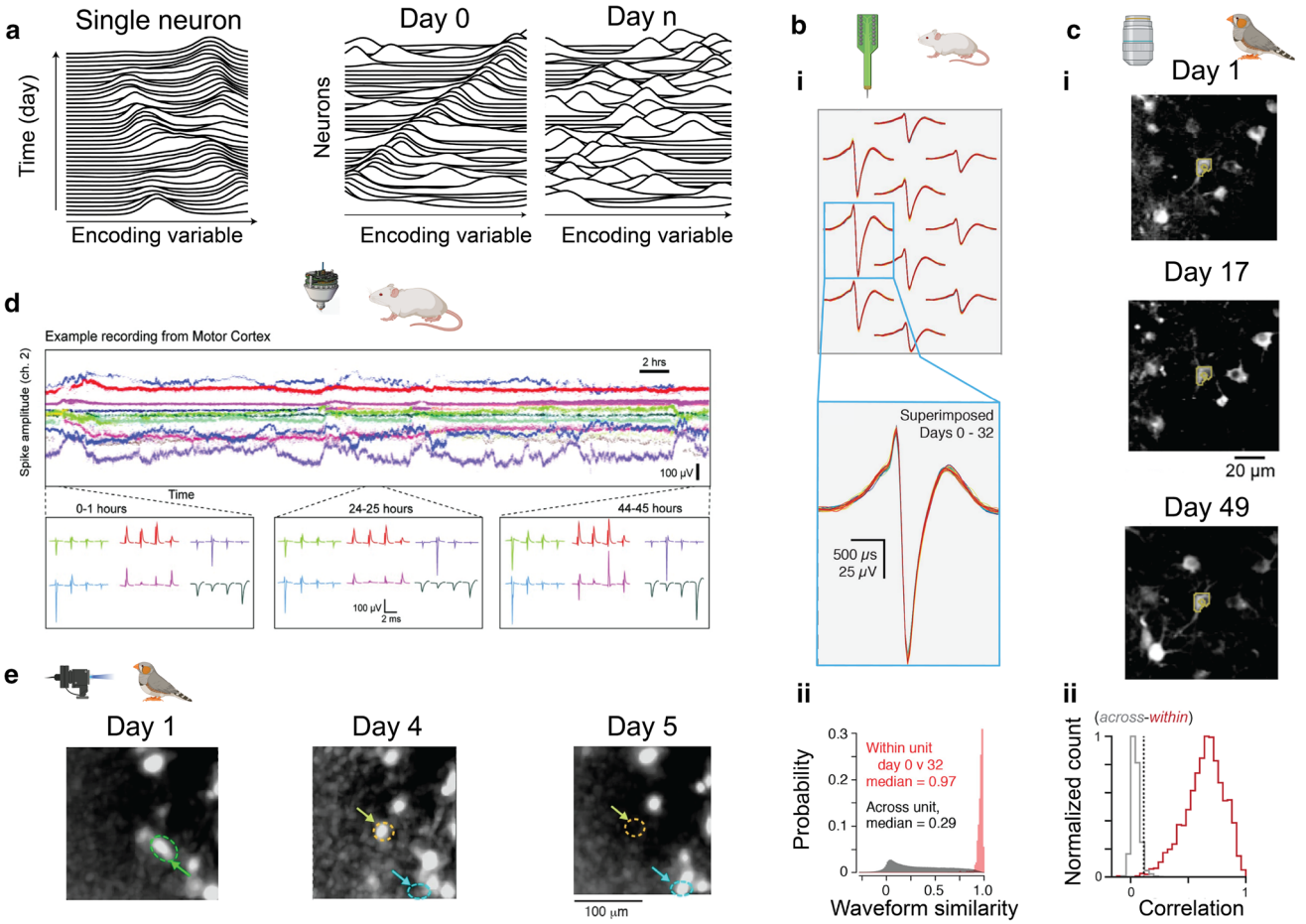

Fig. 1 Drifting neural representations and experimental tools to measure neural activity across days. **a**: Drift in the neural representation is defined as a change over time in how the population encodes a given stimulus. Left: change in the tuning curve of a single neuron. Right: change in the population code. These visualizations were generated using the drift model from Rule and O'Leary (2021). **b**: i. The spike waveform of a single neuron across electrodes can be used to identify it across days. ii. Waveform similarity across days within neurons is much higher than across neurons. Adapted from Schoonover et al (2021). **c**: i. Two-photon microscopy can identify the same neurons across long periods of time. Note that there is a wealth of anatomical information (axon initial segment, dendrites, etc.) that can help in the identification across days. ii. The activity of within-cell pixels is strongly correlated across days in comparison with across-cell pixels. Adapted from Katlowitz et al (2018). **d**: Continuous electrophysiological recordings show that substantial drift in the waveform can occur on a timescale of hours. Experiments recording cells solely during an experimental session could miss a significant (and potentially biased) number of cells. Adapted from Dhawale et al (2017). **e**: One-photon microscopy can also be used for longitudinal recordings and allows recording in a wider set of experimental conditions than two-photon microscopy, including freely moving behavior. However, usually less anatomical information is available and great care has to be taken to confirm neuron identity across days. Arrows indicate neurons coming into and out of the field of view across days. Adapted from Liberti et al (2016)

rons (e.g., those with low firing rates or from a specific cell type) the experiments could lead to misleading interpretations. Yet, with careful registration of neuron identity and knowledge of the limitations of the experimental methods, both electrophysiology and two-photon calcium imaging can achieve stable long-term recordings to investigate the stability of the neural code.

3 Experimental observations

Stable long-term recordings enabled by these technical advances have revealed evidence for drifting neuronal representations across both neocortical and allocortical areas. Below, we review evidence for representational drift in hippocampus and in parietal, primary sensory, and motor areas.

3.1 Hippocampus and parietal cortex

Hippocampal CA1 place cells play a crucial role in spatial navigation and episodic memories. When an animal repeatedly explores a given environment, CA1 pyramidal cells form place fields that tile the physical space. Early studies using low-throughput electrophysiology have suggested that once established, place fields of the same environment remain stable over months in some single neurons (Thompson and Best 1990). However, recent long-term recordings showed that the ensemble of place cells representing a familiar environment change day-by-day (Ziv et al (2013); see also Mankin et al (2012); Kentros et al (2004)) (Fig. 2A). Place cells that are identified in a given day have place fields that tile the linear track, but may drop out of the task-relevant ensemble in the subsequent days. Despite this highly dynamic representation of space, spatial information can be decoded from the small shared subset of neurons that remain active across days (Ziv et al 2013). Similar drift of place cell activity during long-term recordings has been reported by Gonzalez et al (2019) and Sheintuch et al (2020), and further confirmed using two-photon imaging by Lee et al (2020).

Neurons in posterior parietal cortex (PPC) are believed to represent the association between sensation and action, which plays a crucial role in many sensorimotor tasks (Harvey et al 2012; Driscoll et al 2017). Using two-photon microscopy, Driscoll et al (2017) recorded hundreds of PPC neurons while mice were proficiently performing visual cue guided virtual reality "T-maze" task over weeks (Fig. 2B). They found that a subset of neurons fired transiently during task trials, with each neuron firing strongly when the animal was at different locations of the T-maze. Thus, the neuron population exhibits sequential patterns of activity that tile the T-maze. Strikingly, PPC neuronal activity continuously changes over weeks even though the task performance remains stable. As shown in Fig. 2B, many neurons changed their activity pattern (tuning properties), either by exiting or entering the neuron ensemble that represents the task or by changing their tuning curves, i.e., firing strongly at different locations. Despite the single-neuron drift, task-relevant behavioral information can be linearly decoded from some subpopulation of PPC neurons on any given day using a decoder trained on activity from that day.

3.2 Sensory cortex

Representational drift has recently been observed even in primary sensory cortices. In the conventional view, neurons in the primary sensory cortices are tuned directly to features of physical stimuli, and have stable and well-defined tuning properties (Hubel 1995). This intuitive picture is challenged by recent measurements of population responses in mouse primary olfactory (piriform) and visual (V1) cortices.

Piriform cortex is commonly thought to encode odor identity (Stettler and Axel 2009; Roland et al 2017). Yet, Schoonover et al (2021) recently found that odor-evoked population activity in piriform is unstable. Though some population-level statistics—such as the fraction of active neurons, the response sparsity, and the within-day response correlation—remain stable, the representations of odor responses change gradually over time (Fig. 2D). Daily exposure to the same odor can slow down the drift, but once the exposure is halted, the drift continues. Odor-associated fear-conditioning does not reduce the drift. In all, these observations suggest that odor representation in the piriform cortex strongly depends on recent history, and is not anchored by odor valence.

Marks and Goard (2021) compared the stability of the responses of mouse V1 neurons to artificial periodic oriented drifting gratings (PDG) and to naturalistic movies (MOV) over several weeks. They found that neurons responded highly stably to PDG, while the responses to MOV showed strong drift (Fig. 2C). Such drift happened in all cortical layers, in both excitatory and inhibitory neurons. Neurons that showed strong responses tended to be more stable. Interestingly, for neurons that responded to both PDG and MOV, only responses to MOV drifted across time. Similar drift in six mouse visual areas has also been reported by Deitch et al (2021) using the Allen Brain Observatory Dataset.

3.3 Motor areas

Though evidence for representational drift in sensory cortex, parietal cortex, and hippocampus appears compelling, the picture in motor areas is murkier. Intuitively, one might expect greater stability in these regions because their representations are only a few synapses away from terminal motor neurons (Gallego et al 2017, 2020). Thus far, studies of representational stability in motor systems have focused primarily on the circuits that underlie two precise learned behaviors: song in zebra finch and reaching movements in monkeys.

The adult zebra finch song is exceptionally stereotyped, with less than 2% timing variability across renditions (Glaze and Troyer 2006). This precise timing depends on the premotor nucleus HVC, which generates bursts of activity with submillisecond precision across hours of recording (Hahnloser et al 2002; Long et al 2010). While chronic one-photon imaging experiments indicated instability in the activity of nearly 40% of recorded HVC projection neurons over a 5-day period (Liberti et al 2016) (Fig. 1E), subsequent two-photon imaging recordings over 60 days showed that HVC activity remained highly stable at both single-neuron and network levels (Katlowitz et al 2018) (Fig. 1C). Given the limitations of one-photon imaging, further study will be required to conclusively determine whether these discrepant conclusions reflect differences between recording methods and/or exper-

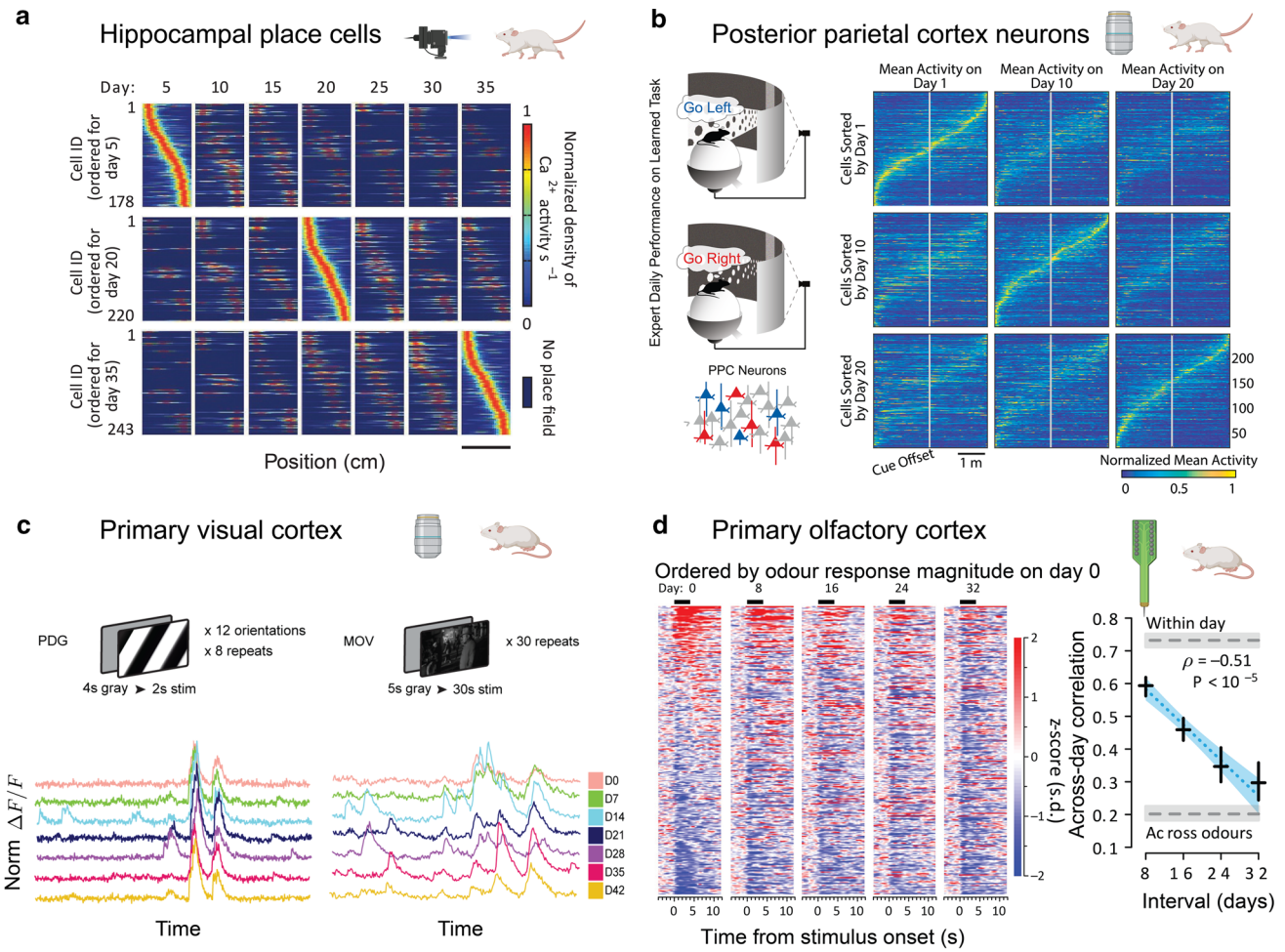

Fig. 2 Observation of representational drift in various brain areas. **a**: Place field maps for CA1 pyramidal cells found across different days after the mice were familiar with the linear track environment. Neurons are ordered by the response centroid position at day 5 (first row), day 20 (second row), and day 35 (last row). Adapted from Ziv et al (2013). **b**: Mice were trained to turn left or right based on the visual patterns when navigating through a virtual reality "T-maze." PPC neurons developed tuning to either left-turn or right-turn task. Right: mean activity of neurons identified at three different days. Neurons were sorted for the same day in each row. Adapted from Driscoll et al (2017). **c**: Responses of mouse V1 neurons to natural movies (MOV) drift, while responses to periodic drifting gratings (PDG) remain relatively stable. Lower: Fluorescence trace for an example neuron over weeks. Adapted from Marks and Goard (2021). **d**: Left: Odor-evoked response of piriform cortex neurons over 32 days. Black bars mark the duration of odor stimuli. Neurons are ordered based on their response at day 0. Right: Correlation coefficient of population vectors decrease gradually over days. Adapted from Schoonover et al (2021)

imental conditions, and thus whether or not single-neuron activity in HVC is stable. Nevertheless, the two-photon imaging results suggest that HVC representations likely remain substantially stable.

Studies of mammalian motor cortex have primarily focused on the stability of primary motor cortex (M1) population representations in monkeys trained to perform precise reaching tasks (Rokni et al 2007; Chestek et al 2007; Dickey et al 2009; Stevenson et al 2011; Kaufman et al 2014; Stavisky et al 2017; Gallego et al 2017, 2020). Some early work claimed that single-neuron activity in M1 is inherently variable (Rokni et al 2007), while other studies asserted that it is stable within measurement noise (Chestek et al 2007; Stevenson et al 2011). However, the multielectrode arrays used in most of these chronic recordings cannot sample a stable set of neurons over long periods (Dickey et al 2009; Stevenson et al 2011; Gallego et al 2020). As a result, experiments were roughly limited to careful characterization of the population code, and could not interrogate the possibility of single-neuron drift (Kappel et al 2015; Stavisky et al 2017; Gallego et al 2017, 2020). Within these constraints, Gallego et al (2020) proposed that M1 activity during reaching lies on a low-dimensional "neural manifold" whose geometry remains highly stable throughout recordings lasting up to two years (Gallego et al 2017, 2020). We note that a very recent report suggests that representations of innate behaviors in

rodent motor cortex may indeed be stable at a single-neuron level (Jensen et al 2021), consistent with representations of learned behavior in the finch (Katlowitz et al 2018).

To conclude, there is a growing body of work suggesting that single-neuron representations in diverse brain areas drift on a timescale of days to weeks. Further work will help elucidate the differences in the structure of drift observed across tasks and brain areas, as well as the factors governing the rate of drift across the cortical hierarchy (Pérez-Ortega et al 2021; Katlowitz et al 2018; Jensen et al 2021).

4 Theoretical ideas

Observations of representational drift seem to contradict the idea that stable neural activities underlie stereotyped behaviors, contravening simple theories of neural coding (Thompson and Best 1990; Barnes et al 1997; Tonegawa et al 2015). Here, we discuss how Bayesian inference can provide a unifying framework for understanding the computational role of drift.

4.1 Noisy plasticity and sampling

For clarity of exposition, we introduce these theoretical ideas within a feedforward toy model of a population of neurons responding to a single input. We assume that the population activity $\mathbf{r}$ is given in terms of the input $\mathbf{x}$ by a simple linear–nonlinear model:

$$\mathbf{r}(\mathbf{x}) = g(\mathbf{V}\mathbf{x}), \tag{1}$$

where g is a fixed activation function and $\mathbf{V}$ is a matrix of tunable synaptic weights. We suppose that the "goal" of this network is to produce some desired output $\mathbf{y}$, and that it has access to a function $\mathcal{E}$ that measures the error between the target and actual activities:

$$\mathcal{E} = \mathcal{E}(\mathbf{r}(\mathbf{x}), \mathbf{y}). \tag{2}$$

A typical example of an error metric is the squared error $\mathcal{E} = \|\mathbf{r}(\mathbf{x}) - \mathbf{y}\|^2$. This picture could of course be generalized to the average error over multiple inputs, but we will focus on the single-input case for clarity.

In neuroscience and machine learning, learning tasks of this form are most often framed as searches for a single set of synaptic weights $\mathbf{V}_*$ that minimizes the error $\mathcal{E}$. A simple plasticity rule that aims to accomplish this goal is gradient descent, which updates the weights as a function of time via

$$\mathbf{V}_t = \mathbf{V}_{t-1} - \eta \nabla \mathcal{E}, \tag{3}$$

where $(\nabla \mathcal{E})_{ij} = \partial \mathcal{E} / \partial V_{ij}$ is the gradient of the error with respect to the synaptic weights and η is a learning rate, which can be understood as the inverse of the learning timescale. From this perspective, the parameters should remain static after the minimization procedure has converged, yielding stable network activity.

Yet, considering the stochasticity of synaptic dynamics, it is implausible that synaptic weight updates in the brain would be noise-free (Attardo et al 2015; Kappel et al 2015; Mongillo et al 2017). The simplest possible noisy generalization of the plasticity rule (3) would be to add a Gaussian noise term to the update (Rokni et al 2007; Kappel et al 2015; Gardiner 1985; Øksendal 2003):

$$\mathbf{V}_t = \mathbf{V}_{t-1} - \eta \nabla \mathcal{E} + \sqrt{2\beta^{-1}\eta}\, \mathbf{\Xi}_t. \tag{4}$$

Here, the noise matrices $\mathbf{\Xi}_t$ are independent and identically distributed in time, with elements that are independent and identically distributed standard Gaussian random variables, and the parameter $\beta > 0$ sets the variance of the noise. In the limit $\eta \to 0$ of long learning timescales, these noisy updates tend to the Langevin dynamics (Gardiner 1985; Øksendal 2003)

$$d\mathbf{V}_t = -\nabla \mathcal{E}\, dt + \sqrt{2\beta^{-1}}\, d\mathbf{W}_t, \tag{5}$$

where the components of the matrix $\mathbf{W}_t$ are independent standard Brownian motions.

With the addition of this noise term, the synaptic weights no longer converge to a steady state, and the network function is never static. At long times, instead of converging to a fixed value, the Langevin dynamics will explore the weight space. Concretely, under mild assumptions (Gardiner 1985; Øksendal 2003), these dynamics sample an equilibrium distribution with density

$$p_\infty(\mathbf{V}) \propto \exp\left[-\beta \mathcal{E}\right], \tag{6}$$

as illustrated in Fig. 3b.

Though the noise in the dynamics (5) renders the representation unstable, it serves a concrete computational purpose. In particular, a principled probabilistic choice for $\mathcal{E}$ is the energy associated to the Bayesian posterior probability of the weights $\mathbf{V}$ given the data $(\mathbf{x}, \mathbf{y})$:

$$\beta \mathcal{E} = -\log p(\mathbf{V} \mid \mathbf{x}, \mathbf{y}). \tag{7}$$

Then, the sampling procedure allows the network to perform Bayesian inference. Concretely, by averaging outputs over time once the weights have equilibrated, the network can average predictions over the possible parameters consistent with the given data, as measured by the posterior probability (Neal 1993; Welling and Teh 2011). In contrast

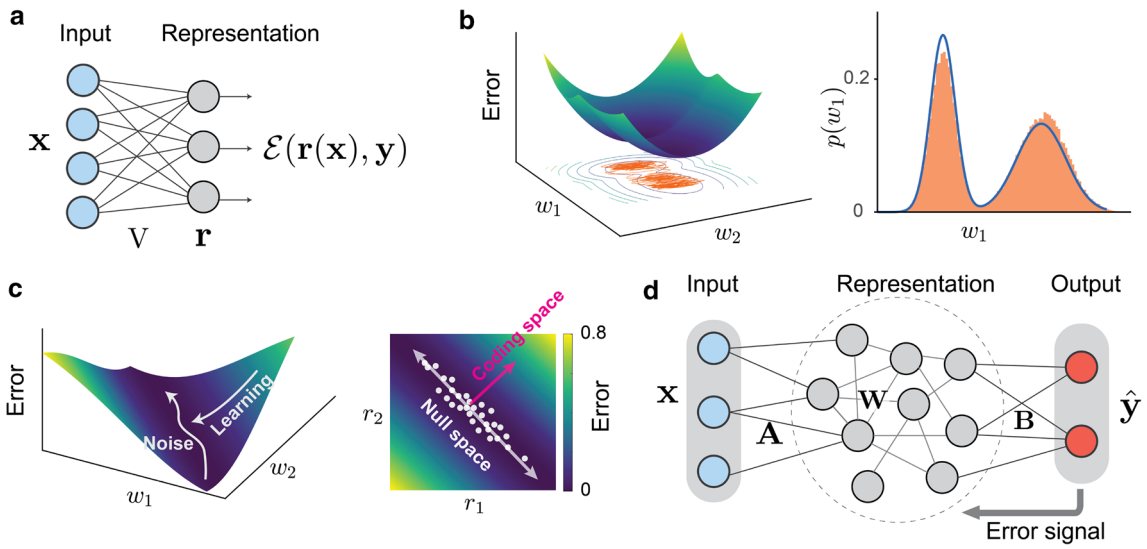

Fig. 3 Theoretical explanations for drifting neural representations. **a**: A single-layer neural network takes an input **x** and maps it to an output **r**. The error function $\mathcal{E}$ quantifies task performance. **b**: Noisy learning as sampling in weight space. *Left:* Shown is an example of task whose error function has two local minimal in the $w_1 - w_2$ space. The orange trace corresponds to the sampled trajectory in weight space (equation 5). *Right:* The equilibrium probability distribution of w_1 is bimodal, corresponding to the two minima of $\mathcal{E}$. The projection of the trajectory onto w_1 in the left panel samples this distribution. **c** Networks with redundancy can exhibit representational drift while maintaining stable performance. *Left:* Once the learning process reaches the flat valley in the error landscape $\mathcal{E}$, any noise will cause drift along this valley, corresponding to different representations that yield equally good task performance. *Right:* In the neural activity space, a task performance may be invariant to the change of neural activities in certain directions (Null space). For example, if task performance only depends on r_1 and r_2 through their sum $r_1 + r_2$, then both r_1 and r_2 can change without compromising performance so long as $r_1 + r_2$ remains constant. **d**: Various sources can contribute to representational drift in hierarchical networks. To maintain stable performance, the representation in the layer of interest might have to change to compensate for drift in upstream and downstream layers, including changes in the input **A** and output **B** synaptic weights.

to simply choosing a single error-minimizing weight matrix $\mathbf{V}_*$, this allows the network to account for the uncertainty inherent in inference from limited data. When only a few task examples are available, the posterior will be broad, but as more data is received and the evidence becomes stronger, the posterior predictive will become increasingly concentrated. This natural accounting for uncertainty may allow better generalization compared to maximum-likelihood estimation (Kappel et al 2015; Neal 1993; Welling and Teh 2011; Wilson and Izmailov 2020; Izmailov et al 2021).

In neuroscience, this idea has principally been explored from the point of view of short-timescale activity inference. Instead of sampling the weights as described above, these models consider sampling the network activity **r** rather than fixing the input–output map (Buesing et al 2011; Orbán et al 2016; Aitchison and Lengyel 2017; Fiser et al 2010; Savin and Deneve 2014; Denève et al 2017). This proposal has successfully explained a range of experimental measurements of short-timescale variability (Orbán et al 2016; Echeveste et al 2020).

Recent work has extended these ideas to the learning of synaptic weights in the manner described above (Kappel et al 2015; Aitchison et al 2021; Llera-Montero et al 2019). Instead of sampling the space of neural representations consistent with a given stimulus, the network explores the space of synaptic weights consistent with the set of stimuli experienced over a longer time period. In these models, the stochasticity in dendritic spine dynamics induces stochasticity in synaptic weights (Attardo et al 2015). When coupled with error signals, this process can implement Bayesian sampling at a slower timescale than activity sampling (Kappel et al 2015; Hiratani and Fukai 2018; Aitchison et al 2021). The changing synaptic weights lead to drifting single-neuron representations, but these stochastic dynamics are not arbitrary. Rather, they allow the network to explore the space of parameters consistent with its history.

4.2 Sampling in the presence of redundancy

In the sampling picture, the structure of the error landscape strongly affects the statistics of drift (Kwon et al 2005). In biological contexts, one particularly salient form of structure in error landscapes is redundancy. Redundancy in task performance is thought to be necessary to achieve robust function, as it allows for accommodation of the parameter imprecision inherent in biological systems (Marder et al 2015; Goaillard and Marder 2021; Li et al 2016). In our toy model, redundancy would correspond to the existence of multiple sets

of synaptic weights $\mathbf{V}$ that achieve minimal error. A simple form of this redundancy would be a flat valley in the error landscape (Fig. 3C). Once the learning process has found the valley, even small amounts of noise will drive substantial drift in synaptic weights along the flat direction of the valley. As a result, neural representations will drift. Importantly, the structure of the error landscape can also suppress drift; variance will be reduced along high-curvature directions.

Rokni *et al.* first applied the idea of redundancy in activity to explain their measurements of drifting single-neuron tuning curves in monkey motor cortex (Rokni et al 2007). They built a redundant neural network and trained it to perform simple reaching tasks. They found that the random component in synaptic modifications during learning drives synaptic weights wandering in a subspace that give the same behavior but different neural representations. More generally, high-dimensional representations of low-dimensional tasks have been observed in various brain regions (Gao et al 2017; Gallego et al 2018). If only the low-dimensional task-encoding manifold affects task performance—as quantified by the error $\mathcal{E}$—many distinct high-dimensional configurations could be functionally equivalent (Rule et al 2019; Gallego et al 2020). A closely related idea is the concept of a "coding null space," a subspace of neural activity that is orthogonal to the task-encoding subspace (Fig. 3C). This idea has been used to explain preparatory and inter-area communication activity in primary motor cortex by Kaufman et al (2014).

However, these previous works have not proposed biologically plausible mechanistic models for how drift can arise in redundant circuits. Recent work by Qin et al (2021) studied a mechanistic model for the dynamics of drift during noisy representation learning. This model considers a neural population that learns to represent stimuli in a way that optimizes a representational similarity objective. This objective has many degenerate minima (Qin et al 2021; Pehlevan et al 2017; Sengupta et al 2018), hence even relatively small amounts of noise in synaptic updates can drive the network to explore the space of near-optimal representations. Importantly, unlike in the Rokni et al (2007) model, this redundancy does not result from having a large number of neurons in the representation layer, but directly from the structure of the task. The drift of single-neuron receptive fields in this model can be described by a coordinated random walk, a prediction which is consistent with the statistics of measured drift in the responses of hippocampal CA1 neurons (Qin et al 2021; Gonzalez et al 2019). Despite this instability in single-neuron representations, the geometry of the population code remains approximately stable over time, as enforced by the representational similarity error. In all, this model recapitulates many features of the experimental observations in hippocampus and PPC through sampling in an error landscape with a subspace of degenerate minima.

4.3 Sampling in hierarchical networks

In the preceding sections, we introduced sampling in a single-layer feedforward network. However, biological neural networks are both hierarchical and recurrent. In this section, we highlight some of the new phenomena that can emerge during weight sampling in such networks.

For simplicity, we will illustrate these ideas using a rate network. We consider a network with input $\mathbf{x}$ weighted by a feedforward matrix $\mathbf{A}$, neuron firing rates $\mathbf{r}$, recurrent weights $\mathbf{W}$, output weights $\mathbf{B}$ and output $\hat{\mathbf{y}}$ (Fig. 3D). The goal of the network is to minimize the error $\mathcal{E}(\hat{\mathbf{y}}, \mathbf{y})$ between the network output and the target y. The dynamics of the population activity $\mathbf{r}$ is described by a simple firing rate model:

$$\tau \dot{\mathbf{r}} = -\mathbf{r} + \phi(\mathbf{W}\mathbf{r} + \mathbf{A}\mathbf{x}), \tag{8}$$

where τ is the neurons' membrane time constant and $\phi(\cdot)$ is an element-wise nonlinear activation function. We refer to the steady-state neural activity of single neurons in (8) as their "representation" of the input. The network output is given by a simple feedforward readout of the rate:

$$\hat{\mathbf{y}} = f(\mathbf{B}\mathbf{r}), \tag{9}$$

where $\mathbf{B}$ is the decoding weight matrix and $f(\cdot)$ is some element-wise nonlinear function. We assume that the timescale of synaptic plasticity is much slower than the timescale of neural dynamics.

In a hierarchical network (Fig. 3D), learning (and drift) can occur in populations downstream and upstream of the recorded population. In our framework, this would correspond to changes in the matrices $\mathbf{A}$ and $\mathbf{B}$. To maintain task proficiency, the recurrent connectivity $\mathbf{W}$ would need to change to compensate for this drift. These compensatory changes will in turn result in drift in the experimentally measured representations $\mathbf{r}$. Similarly, if drift in $\mathbf{r}$ due to changes in $\mathbf{A}$ or $\mathbf{W}$ is confined to some subspace, a static readout can achieve robust decoding. However, if the drift is less structured, then the decoder weights $\mathbf{B}$ must compensate for changes in activity in order to produce the desired output $\mathbf{y}$. Such compensation appears to be necessary in order to linearly read out task and behavioral information from PPC activity during T-maze tasks, as the performance of a fixed decoder degrades significantly after several days due to measured drift (Rule et al 2020). Rule *et al.* have explored different possible biologically plausible adaptation rules that achieve robust readout under various constraints, including the absence or presence of explicit external error signals (Rule et al 2020; Rule and O'Leary 2021). However, these studies have yet to investigate compensatory drift in hierarchical networks performing probabilistic computations.

5 Outlook

In this review, we presented the technical advances and experimental observations that have recently led researchers to question the stability of single-neuron representations. We then presented possible mechanisms through which drift could occur within a network performing normative computation. Importantly, these mechanisms are not mutually exclusive. They affect different parts of neural computations, and—considering that they have distinct roles—it is likely that they coexist, with each accounting for some fraction of the drift observed in neural population codes. Here, we outline some promising directions for future experimental and theoretical research.

The stability of the geometry of population representations remains an important open question. Though the tuning of individual neurons determines the manifold of neural activity, the same coding geometry in a redundant circuit could arise from different single-neuron tuning (see Sect. 4.2 and Kriegeskorte and Wei (2021)). Analyses of neural responses in the monkey motor cortex and rodent visual cortex suggest that the geometry of the population code is stable (Gallego et al 2020; Xia et al 2021), while analysis of drifting representations in olfactory cortex suggests a changing geometry (Schoonover et al 2021). In higher areas, the representational similarity of spatial representations in CA1 appears to remain stable over weeks (Qin et al 2021; Gonzalez et al 2019); it will be interesting to test whether this is also true in PPC.

In the simple model described in Sect. 4.3, drift in the representation $\mathbf{r}$ is driven by two sets of synaptic weights: the feedforward weights $\mathbf{A}$ and the recurrent weights $\mathbf{W}$. In a biologically plausible model, it is likely that these two sets of synaptic weights would have different learning dynamics. Different classes of stimuli will recruit different degrees of recurrent interaction, resulting in representations that may be driven primarily by forward input $\mathbf{Ax}$ or recurrent input $\mathbf{Wr}$. Differences in the rate at which these two weights matrices drift would thus lead to differences in measured drift rates of representations across stimulus classes. In visual cortex, representations of stimuli known to elicit maximum activation (drifting gratings) appear relatively stable in comparison with representations of more complex visual scenes (Deitch et al 2021; Marks and Goard 2021). This could be interpreted as stability of the feedforward weights establishing the classical (center) receptive field and drift of the recurrent and feedback weights contributing to the extra-classical (surround) receptive field (Rao and Ballard 1999; Angelucci et al 2002; Chalk et al 2017; Carandini and Heeger 2012). Extending these concepts to modalities in which the notion of classical and extra-classical receptive fields is less well understood remains a challenge (Carandini and Heeger 2012; Wanner and Friedrich 2020).

Thus far, most studies have defined representations by the steady-state firing or average activity over a relatively large window ($\sim$ 1 s). Yet, neurons also have rich transient dynamics; the stability of these dynamics has not been comprehensively characterized. Comparing drift in transient and steady-state responses could potentially afford some insight into the relative contributions of feedforward and feedback weights (Wanner and Friedrich 2020).

This issue highlights one of the key difficulties faced by neuroscientists. Learning in artificial neural networks is defined through synaptic weight dynamics, but experimental neuroscientists typically have access to only time-resolved large-scale measurements of neural activity, and not of synaptic weights. Inferring weights and learning rules from neural activity is extremely challenging. Some theoretical work has started to attempt inference of network weights from activity, but much work remains before these methods can be applied to neural recordings (Pereira and Brunel 2018; Nayebi et al 2020; Goldt et al 2021; Chalk et al 2021). This work will be important to directly link drift in representations with changes in synaptic weights beyond comparison between transient and steady-state responses.

Attempts to probe redundancies in neural circuits will also face similar challenges. Within the illustrative model of Sect. 4.3, a simple way to model redundancy due to the presence of a coding null space would be for the error to depend on the representation $\mathbf{r}$ only through some linear readout $\mathbf{Br}$. If the matrix $\mathbf{B}$ has a nontrivial null space, any drift in the population representation $\mathbf{r}$ within this null space would keep the error constant. However, in most settings the error landscape will be more complex, and experimentally identifying the redundant directions in neural representations will be challenging. Even for artificial networks, for which one has access to complete information about every neuron, precise characterization of high-dimensional error landscapes is a highly nontrivial task (Baldi and Hornik 1989; Kawaguchi 2016; Geiger et al 2019). Therefore, precise characterization of redundancies in complex neural systems is likely to be a challenging task.

Previous attempts to experimentally probe circuit redundancies have focused either on characterizing the task-relevant activity subspace (Kaufman et al 2014; Gallego et al 2017, 2020) or on characterizing perturbations that severely impair long-term neural dynamics and behavioral performance (Li et al 2016; Inagaki et al 2019). To obtain a fine-grained understanding of circuit redundancies, it will be necessary to perform high-precision perturbation experiments (Adesnik and Abdeladim 2021; Banerjee et al 2021; Jazayeri and Afraz 2017). Such experiments would elucidate causal links between neural activity and behavioral function, which cannot conclusively be established with correlative measurement of activity. If one can identify redundant directions in activity space, one would then expect to observe

greater drift along those directions than along task-relevant directions. However, for the abovementioned reasons, comprehensive mapping of the high-dimensional landscape is likely to be quite challenging.

On the theoretical side, studies of weight sampling in neural networks have thus far mostly focused on the properties of the prior and Bayes posterior, and do not consider nonequilibrium properties of the sampling process. Even at this level, the link between the choice of prior distribution over weights and the prior distribution over network activity remains poorly understood (Wilson and Izmailov 2020; Izmailov et al 2021; Aitchison 2020; Zavatone-Veth and Pehlevan 2021; Yang 2019). One can predict the approximate equilibrium statistics of representations for simple network architectures and error metrics in certain regimes (Aitchison 2020; Zavatone-Veth et al 2021), but these calculations remain challenging even for feedforward nonlinear networks. As a result, significant theoretical work will be required in order to make experimentally testable predictions for the statistics of drifting representations during sampling.

In our discussion of theories of drift, we focused on the learning of a single task. However, it is important to note that representational drift due to Bayesian sampling and classic proposals for representational changes due to continual learning are not mutually exclusive. Indeed, these ideas can be unified under the umbrella of Bayesian continual learning (Opper 1999; Kirkpatrick et al 2017; Kulhavý and Zarrop 1993). Depending on the timescales of intrinsic synaptic noise and of the arrival of new tasks, equilibrium sampling may or may not be a good approximation for probabilistic computations in the brain. If different tasks are well separated in time, drift after learning a single task can result from equilibrium sampling, particularly in the presence of redundancy. The introduction of a new task would alter the error landscape, introducing additional sources of representational modification beyond sampling variability (Fusi 2021; Schoonover et al 2021). Further changes could result from the need to compensate for neuron death (Barrett et al 2016; Calaim et al 2020). In addition to the ideas around Bayesian sampling proposed here, ongoing synaptic plasticity has been posited to be crucial for maintaining network dynamics near criticality, which has been proposed to be computationally beneficial (Zeraati et al 2021; Beggs and Timme 2012; Yu et al 2017; Das and Levina 2019; de Andrade Costa et al 2015). These processes could all contribute to measurable drift—as defined experimentally—but their timescales and statistical structure are likely to differ.

Disentangling the relative contributions of different sources of drift will require both new theoretical work and new experiments spanning larger stimulus sets and wider spatial and temporal scales. Mechanistic modeling of circuits performing Bayesian continual learning will help elucidate how the temporal structure of tasks contributes to the structure of drift. Moreover, it will be important to theoretically elucidate how stimulus statistics and circuit architecture affect differences in the statistical structure of representational changes due to continual learning and drift due to sampling (Qin et al 2021; Zavatone-Veth et al 2021). Experimentally, it will be important to probe how stimulus structure and task complexity affect the structure of drift (Schoonover et al 2021; Pashkovski et al 2020). Moreover, multi-area recordings (Luo et al 2020; Sofroniew et al 2016; Steinmetz et al 2021; Chung et al 2019) may allow experimentalists to quantify co-variation in representational drift across the cortical hierarchy. This may allow separation of compensatory drift from other mechanisms.

In summary, we propose that drift could be a feature of robust probabilistic computation in hierarchical networks. Nonetheless, significant challenges remain to properly understand and characterize drift in neuronal representations.

Acknowledgements We thank Cengiz Pehlevan and Venkatesh Murthy for support and mentorship. We also thank Matthew Farrell, Michael Goard, Siddharth Jayakumar, and Torben Ott for helpful comments on our manuscript.

Author Contributions All authors contributed equally to conceptualization, literature review, and writing. They are listed alphabetically.

Funding PM was supported by a grant from the Harvard Mind Brain Behavior Interfaculty Initiative. SQ and JAZ-V were supported by the NIH (1UF1NS111697-01), the Intel Corporation (through the Intel Neuromorphic Research Community), and a Google Faculty Research Award. JAZ-V was also partially supported by the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard, the Harvard Quantitative Biology Initiative, and the Harvard FAS Dean's Competitive Fund for Promising Scholarship.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no other competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Adesnik H, Abdeladim L (2021) Probing neural codes with two-photon holographic optogenetics. Nat Neurosci pp 1–11. https://doi.org/10.1038/s41593-021-00902-9

Ahrens MB, Orger MB, Robson DN et al (2013) Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat Methods 10(5):413–420. https://doi.org/10.1038/nmeth.2434

Aitchison L (2020) Why bigger is not always better: on finite and infinite neural networks. In: III HD, Singh A (eds) Proceedings of the 37th

International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 119. PMLR, pp 156–164
Aitchison L, Lengyel M (2017) With or without you: predictive coding and Bayesian inference in the brain. Curr Opin Neurobiol 46:219–227. https://doi.org/10.1016/j.conb.2017.08.010
Aitchison L, Jegminat J, Menendez JA et al (2021) Synaptic plasticity as Bayesian inference. Nat Neurosci 24(4):565–571. https://doi.org/10.1038/s41593-021-00809-5
Amit DJ, Fusi S (1994) Learning in neural networks with material synapses. Neural Comput 6(5):957–982. https://doi.org/10.1162/neco.1994.6.5.957
de Andrade Costa A, Copelli M, Kinouchi O (2015) Can dynamical synapses produce true self-organized criticality? Journal of Statistical Mechanics: Theory and Experiment 2015(6):P06004. https://doi.org/10.1088/1742-5468/2015/06/P06004
Angelucci A, Levitt JB, Walton EJ et al (2002) Circuits for local and global signal integration in primary visual cortex. J Neurosci 22(19):8633–8646. https://doi.org/10.1371/journal.pcbi.1005582
Attardo A, Fitzgerald JE, Schnitzer MJ (2015) Impermanence of dendritic spines in live adult CA1 hippocampus. Nature 523(7562):592–596. https://doi.org/10.1038/nature14467
Baldi P, Hornik K (1989) Neural networks and principal component analysis: Learning from examples without local minima. Neural Netw 2(1):53–58. https://doi.org/10.1016/0893-6080(89)90014-2
Banerjee A, Egger R, Long MA (2021) Using focal cooling to link neural dynamics and behavior. Neuron. https://doi.org/10.1016/j.neuron.2021.05.029
Barnes CA, Suster MS, Shen J et al (1997) Multistability of cognitive maps in the hippocampus of old rats. Nature 388(6639):272–275. https://doi.org/10.1038/40859
Barrett DG, Deneve S, Machens CK (2016) Optimal compensation for neuron loss. eLife 5(e12):454. https://doi.org/10.7554/eLife.12454
Beggs JM, Timme N (2012) Being critical of criticality in the brain. Front Physiol 3:163. https://doi.org/10.3389/fphys.2012.00163
Buesing L, Bill J, Nessler B et al (2011) Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1002211
Calaim N, Dehmelt FA, Gonçalves PJ, et al. (2020) Robust coding with spiking networks: a geometric perspective. bioRxiv https://doi.org/10.1101/2020.06.15.148338
Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. Nat Rev Neurosci 13(1):51–62. https://doi.org/10.1038/nrn3136
Chalk M, Masset P, Deneve S et al (2017) Sensory noise predicts divisive reshaping of receptive fields. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1005582
Chalk M, Tkacik G, Marre O (2021) Inferring the function performed by a recurrent neural network. Plos ONE. https://doi.org/10.1371/journal.pone.0248940
Chestek CA, Batista AP, Santhanam G et al (2007) Single-neuron stability during repeated reaching in macaque premotor cortex. J Neurosci 27(40):10742–10750. https://doi.org/10.1523/JNEUROSCI.0959-07.2007
Chung JE, Magland JF, Barnett AH et al (2017) A fully automated approach to spike sorting. Neuron 95(6):1381-1394 https://doi.org/10.1016/j.neuron.2017.08.030
Chung JE, Joo HR, Fan JL et al (2019) High-density, long-lasting, and multi-region electrophysiological recordings using polymer electrode arrays. Neuron 101(1):21–31. https://doi.org/10.1016/j.neuron.2018.11.002
Clopath C, Bonhoeffer T, Hübener M et al (2017) Variance and invariance of neuronal long-term representations. Philos Trans R Soc B: Biol Sci 372(1715):20160161. https://doi.org/10.1098/rstb.2016.0161
Das A, Levina A (2019) Critical neuronal models with relaxed timescale separation. Phys Rev X. https://doi.org/10.1103/PhysRevX.9.021062
Deitch D, Rubin A, Ziv Y (2021) Representational drift in the mouse visual cortex. Curr Biol. https://doi.org/10.1016/j.cub.2021.07.062
Denève S, Alemi A, Bourdoukan R (2017) The brain as an efficient and robust adaptive learner. Neuron 94(5):969–977. https://doi.org/10.1016/j.neuron.2017.05.016
Dhawale AK, Poddar R, Wolff SB, et al. (2017) Automated long-term recording and analysis of neural activity in behaving animals. eLife 6:e27702 https://doi.org/10.7554/eLife.27702
Dickey AS, Suminski A, Amit Y et al (2009) Single-unit stability using chronically implanted multielectrode arrays. J Neurophysiol 102(2):1331–1339. https://doi.org/10.1152/jn.90920.2008
Dimitriadis G, Neto JP, Aarts A, et al. (2018) Why not record from every channel with a CMOS scanning probe? bioRxiv https://doi.org/10.1101/275818
Driscoll LN, Pettit NL, Minderer M et al (2017) Dynamic reorganization of neuronal activity patterns in parietal cortex. Cell 170(5):986–999. https://doi.org/10.1016/j.cell.2017.07.021
Ebitz RB, Hayden BY (2021) The population doctrine in cognitive neuroscience. Neuron. https://doi.org/10.1016/j.neuron.2021.07.011
Echeveste R, Aitchison L, Hennequin G et al (2020) Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. Nat Neurosci 23(9):1138–1149. https://doi.org/10.1038/s41593-020-0671-1
Fiser J, Berkes P, Orbán G et al (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 14(3):119–130. https://doi.org/10.1016/j.tics.2010.01.003
French RM (1999) Catastrophic forgetting in connectionist networks. Trends Cogn Sci 3(4):128–135. https://doi.org/10.1016/s1364-6613(99)01294-2
Fusi S (2021) Memory capacity of neural network models. arXiv preprint arXiv:2108.07839
Fusi S, Abbott L (2007) Limits on the memory storage capacity of bounded synapses. Nat Neurosci 10(4):485–493. https://doi.org/10.1038/nn1859
Fusi S, Senn W (2006) Eluding oblivion with smart stochastic selection of synaptic updates. Chaos: An Interdisciplinary Journal of Nonlinear Science 16(2):026112. https://doi.org/10.1063/1.2213587
Gallego JA, Perich MG, Miller LE et al (2017) Neural manifolds for the control of movement. Neuron 94(5):978–984. https://doi.org/10.1016/j.neuron.2017.05.025
Gallego JA, Perich MG, Naufel SN et al (2018) Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. Nat Commun 9(1):4233. https://doi.org/10.1038/s41467-018-06560-z
Gallego JA, Perich MG, Chowdhury RH et al (2020) Long-term stability of cortical population dynamics underlying consistent behavior. Nat Neurosci 23(2):260–270. https://doi.org/10.1038/s41593-019-0555-4
Gao P, Trautmann E, Yu B, et al (2017) A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv p 214262. https://doi.org/10.1101/214262
Gardiner CW (1985) Handbook of stochastic methods, vol 3. Springer, Berlin
Geiger M, Spigler S, d'Ascoli S et al (2019) Jamming transition as a paradigm to understand the loss landscape of deep neural networks. Phys Rev E 100(1):012115. https://doi.org/10.1103/PhysRevE.100.012115

Giovannucci A, Friedrich J, Gunn P, et al (2019) Caiman an open source tool for scalable calcium imaging data analysis. eLife 8:e38173. https://doi.org/10.7554/eLife.38173

Glaze CM, Troyer TW (2006) Temporal structure in zebra finch song: implications for motor coding. J Neurosci 26(3):991–1005. https://doi.org/10.1523/JNEUROSCI.3387-05.2006

Goaillard JM, Marder E (2021) Ion channel degeneracy, variability, and covariation in neuron and circuit resilience. Annual Review of Neuroscience 44. https://doi.org/10.1146/annurev-neuro-092920-121538

Goldt S, Krzakala F, Zdeborová L, et al (2021) Bayesian reconstruction of memories stored in neural networks from their connectivity. arXiv preprint arXiv:2105.07416

Gonzalez WG, Zhang H, Harutyunyan A et al (2019) Persistence of neuronal representations through time and damage in the hippocampus. Science 365(6455):821–825. https://doi.org/10.1126/science.aav9199

Hahnloser RH, Kozhevnikov AA, Fee MS (2002) An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature 419(6902):65–70. https://doi.org/10.1038/nature00974

Harvey CD, Coen P, Tank DW (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484(7392):62–68. https://doi.org/10.1038/nature10918

Hiratani N, Fukai T (2018) Redundancy in synaptic connections enables neurons to learn optimally. Proc Natl Acad Sci 115(29):E6871–E6879. https://doi.org/10.1073/pnas.1803274115

Hubel D (1995) Eye, Brain, and Vision. Scientific American Library series, Henry Holt and Company

Inagaki HK, Fontolan L, Romani S et al (2019) Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature 566(7743):212–217. https://doi.org/10.1038/s41586-019-0919-7

Izmailov P, Vikram S, Hoffman MD, et al (2021) What are Bayesian neural network posteriors really like? In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 4629–4640

Jazayeri M, Afraz A (2017) Navigating the neural space in search of the neural code. Neuron 93(5):1003–1014. https://doi.org/10.1016/j.neuron.2017.02.019

Jensen KT, Harpaz NK, Dhawale AK, et al (2021) Long-term stability of neural activity in the motor system. bioRxiv https://doi.org/10.1101/2021.10.27.465945

Juavinett AL, Bekheet G, Churchland AK (2019) Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice. eLife 8. https://doi.org/10.7554/eLife.47188

Jun JJ, Mitelut C, Lai C, et al (2017) Real-time spike sorting platform for high-density extracellular probes with ground-truth validation and drift correction. bioRxiv https://doi.org/10.1101/101030

Jun JJ, Steinmetz NA, Siegle JH et al (2017) Fully integrated silicon probes for high-density recording of neural activity. Nature 551(7679):232–236. https://doi.org/10.1038/nature24636

Kappel D, Habenschuss S, Legenstein R et al (2015) Network plasticity as Bayesian inference. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1004485

Katlowitz KA, Picardo MA, Long MA (2018) Stable sequential activity underlying the maintenance of a precisely executed skilled behavior. Neuron 98(6):1133–1140. https://doi.org/10.1016/j.neuron.2018.05.017

Kaufman MT, Churchland MM, Ryu SI et al (2014) Cortical activity in the null space: permitting preparation without movement. Nat Neurosci 17(3):440–448. https://doi.org/10.1038/nn.3643

Kawaguchi K (2016) Deep learning without poor local minima. In: Lee D, Sugiyama M, Luxburg U et al (eds) Advances in Neural Information Processing Systems, vol 29. Curran Associates Inc

Kentros CG, Agnihotri NT, Streater S et al (2004) Increased attention to spatial context increases both place field stability and spatial memory. Neuron 42(2):283–295. https://doi.org/10.1016/s0896-6273(04)00192-8

Kirkpatrick J, Pascanu R, Rabinowitz N et al (2017) Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci 114(13):3521–3526. https://doi.org/10.1073/pnas.1611835114

Kriegeskorte N, Wei XX (2021) Neural tuning and representational geometry. Nat Rev Neurosci. https://doi.org/10.1038/s41583-021-00502-3

Kulhavý R, Zarrop MB (1993) On a general concept of forgetting. Int J Control 58(4):905–924. https://doi.org/10.1080/00207179308923034

Kwon C, Ao P, Thouless DJ (2005) Structure of stochastic dynamics near fixed points. Proceed Natl Acad Sci 102(37):13029–13033. https://doi.org/10.1073/pnas.0506347102

Lee JS, Briguglio JJ, Cohen JD et al (2020) The statistical structure of the hippocampal code for space as a function of time, context, and value. Cell 183(3):620–635. https://doi.org/10.1016/j.cell.2020.09.024

Li M, Liu F, Jiang H et al (2017) Long-term two-photon imaging in awake macaque monkey. Neuron 93(5):1049–1057. https://doi.org/10.1016/j.neuron.2017.01.027

Li N, Daie K, Svoboda K et al (2016) Robust neuronal dynamics in premotor cortex during motor planning. Nature 532(7600):459–464. https://doi.org/10.1038/nature17643

Liberti WA, Markowitz JE, Perkins LN et al (2016) Unstable neurons underlie a stable learned behavior. Nat Neurosci 19(12):1665–1671. https://doi.org/10.1038/nn.4405

Llera-Montero M, Sacramento J, Costa RP (2019) Computational roles of plastic probabilistic synapses. Curr Opin Neurobiol 54:90–97. https://doi.org/10.1016/j.conb.2018.09.002

Long MA, Jin DZ, Fee MS (2010) Support for a synaptic chain model of neuronal sequence generation. Nature 468(7322):394–399. https://doi.org/10.1038/nature09514

Luo TZ, Bondy AG, Gupta D, et al (2020) An approach for long-term, multi-probe neuropixels recordings in unrestrained rats. eLife 9. https://doi.org/10.7554/eLife.59716

Mankin EA, Sparks FT, Slayyeh B et al (2012) Neuronal code for extended time in the hippocampus. Proceed Natl Acad Sci 109(47):19462–19467. https://doi.org/10.1073/pnas.1214107109

Marder E, Goeritz ML, Otopalik AG (2015) Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. Curr Opin Neurobiol 31:156–163. https://doi.org/10.1016/j.conb.2014.10.012

Marks TD, Goard MJ (2021) Stimulus-dependent representational drift in primary visual cortex. Nat Commun 12(1):5169. https://doi.org/10.1038/s41467-021-25825-8

Mau W, Hasselmo ME, Cai DJ (2020) The brain in motion: How ensemble fluidity drives memory-updating and flexibility. eLife 9:e63550. https://doi.org/10.7554/eLife.63550

Mongillo G, Rumpel S, Loewenstein Y (2017) Intrinsic volatility of synaptic connections–a challenge to the synaptic trace theory of memory. Curr Opin Neurobiol 46:7–13. https://doi.org/10.1016/j.conb.2017.06.006

Musk E, Neuralink, (2019) An integrated brain-machine interface platform with thousands of channels. J Med Internet Res. https://doi.org/10.2196/16194

Nayebi A, Srivastava S, Ganguli S, et al (2020) Identifying learning rules from neural network observables. arXiv preprint arXiv:2010.11765

Neal RM (1993) Bayesian learning via stochastic dynamics. In: Advances in Neural Information Processing Systems, pp 475–482

Øksendal B (2003) Stochastic differential equations. Springer, Berlin

Opper M (1999) A Bayesian approach to on-line learning. In: Saad D (ed) On-Line Learning in Neural Networks. Cambridge University Press, Publications of the Newton Institute, p 363-378, https://doi.org/10.1017/CBO9780511569920.017

Orbán G, Berkes P, Fiser J et al (2016) Neural variability and sampling-based probabilistic representations in the visual cortex. Neuron 92(2):530–543. https://doi.org/10.1016/j.neuron.2016.09.038

Pachitariu M, Steinmetz NA, Kadir S, et al (2016) Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. bioRxiv https://doi.org/10.1101/061481

Pachitariu M, Stringer C, Dipoppa M, et al (2017) Suite2p: beyond 10,000 neurons with standard two-photon microscopy. bioRxiv https://doi.org/10.1101/061507

Parisi G (1986) A memory which forgets. J Phys A: Math Gen 19(10):L617. https://doi.org/10.1088/0305-4470/19/10/011

Pashkovski SL, Iurilli G, Brann D et al (2020) Structure and flexibility in cortical representations of odour space. Nature 583(7815):253–258. https://doi.org/10.1038/s41586-020-2451-1

Pehlevan C, Sengupta AM, Chklovskii DB (2017) Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? Neural Comput 30(1):84–124. https://doi.org/10.1162/neco_a_01018

Pereira U, Brunel N (2018) Attractor dynamics in networks with learning rules inferred from in vivo data. Neuron 99(1):227–238. https://doi.org/10.1016/j.neuron.2018.05.038

Pérez-Ortega J, Alejandre-García T, Yuste R (2021) Long-term stability of cortical ensembles. eLife 10(e64):449. https://doi.org/10.7554/eLife.64449

Pnevmatikakis EA, Soudry D, Gao Y et al (2016) Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron 89(2):285–299. https://doi.org/10.1016/j.neuron.2015.11.037

Qin S, Farashahi S, Lipshutz D, et al (2021) Coordinated drift of receptive fields during noisy representation learning. bioRxiv https://doi.org/10.1101/2021.08.30.458264

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87. https://doi.org/10.1038/4580

Rokni U, Richardson AG, Bizzi E et al (2007) Motor learning with unstable neural representations. Neuron 54(4):653–666. https://doi.org/10.1016/j.neuron.2007.04.030

Roland B, Deneux T, Franks KM, et al (2017) Odor identity coding by distributed ensembles of neurons in the mouse olfactory cortex. eLife 6:e26337. https://doi.org/10.7554/eLife.26337

Rule ME, O'Leary T (2021) Self-healing neural codes: Hebbian and homeostatic mechanisms can track evolving neural representations. bioRxiv https://doi.org/10.1101/2021.03.08.433413

Rule ME, O'Leary T, Harvey CD (2019) Causes and consequences of representational drift. Current Opinion in Neurobiology 58:141–147. https://doi.org/10.1016/j.conb.2019.08.005

Rule ME, Loback AR, Raman D, et al (2020) Stable task information from an unstable neural population. eLife 9:e51121. https://doi.org/10.7554/eLife.51121

Savin C, Deneve S (2014) Spatio-temporal representations of uncertainty in spiking neural networks. In: Advances in Neural Information Processing Systems, pp 2024–2032

Saxena S, Cunningham JP (2019) Towards the neural population doctrine. Curr Opin Neurobiol 55:103–111. https://doi.org/10.1016/j.conb.2019.02.002

Saxena S, Kinsella I, Musall S, et al (2020) Localized semi-nonnegative matrix factorization (LocaNMF) of widefield calcium imaging data. PLoS Computational Biology 16(4):e1007791. https://doi.org/10.1371/journal.pcbi.1007791

Schoonover CE, Ohashi SN, Axel R et al (2021) Representational drift in primary olfactory cortex. Nature. https://doi.org/10.1038/s41586-021-03628-7

Sengupta AM, Pehlevan C, Tepper M et al (2018) Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. In: Bengio S, Wallach H, Larochelle H et al (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates Inc

Sheintuch L, Geva N, Baumer H et al (2020) Multiple maps of the same spatial context can stably coexist in the mouse hippocampus. Current Biology. https://doi.org/10.1016/j.cub.2020.02.018

Sofroniew NJ, Flickinger D, King J, et al (2016) A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. eLife 5:e14472. https://doi.org/10.7554/eLife.14472

Stavisky SD, Kao JC, Ryu SI et al (2017) Motor cortical visuomotor feedback activity is initially isolated from downstream targets in output-null neural state space dimensions. Neuron 95(1):195–208. https://doi.org/10.1016/j.neuron.2017.05.023

Steinmetz NA, Aydin C, Lebedeva A, et al (2021) Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science 372(6539):eabf4588. https://doi.org/10.1126/science.abf4588

Stettler DD, Axel R (2009) Representations of odor in the piriform cortex. Neuron 63(6):854–864. https://doi.org/10.1016/j.neuron.2009.09.005

Stevenson IH, Cherian A, London BM et al (2011) Statistical assessment of the stability of neural movement representations. J Neurophysiol 106(2):764–774. https://doi.org/10.1152/jn.00626.2010

Svoboda K, Yasuda R (2006) Principles of two-photon excitation microscopy and its applications to neuroscience. Neuron 50(6):823–839. https://doi.org/10.1016/j.neuron.2006.05.019

Thompson L, Best P (1990) Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats. Brain Res 509(2):299–308. https://doi.org/10.1016/0006-8993(90)90555-p

Tonegawa S, Pignatelli M, Roy DS et al (2015) Memory engram storage and retrieval. Curr Opin Neurobiol 35:101–109. https://doi.org/10.1016/j.conb.2015.07.009

Ulivi AF, Castello-Waldow TP, Weston G et al (2019) Longitudinal two-photon imaging of dorsal hippocampal CA1 in live mice. J Vis Exp 148(e59):598. https://doi.org/10.3791/59598

Urai AE, Doiron B, Leifer AM, et al (2021) Large-scale neural recordings call for new insights to link brain and behavior. arXiv preprint arXiv:210314662 https://arxiv.org/abs/arXiv:2103.14662

Wanner AA, Friedrich RW (2020) Whitening of odor representations by the wiring diagram of the olfactory bulb. Nat Neurosci 23(3):433–442. https://doi.org/10.1038/s41593-019-0576-z

Welling M, Teh YW (2011) Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning, pp 681–688

Wilson AG, Izmailov P (2020) Bayesian deep learning and a probabilistic perspective of generalization. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 4697–4708

Xia J, Marks TD, Goard MJ et al (2021) Stable representation of a naturalistic movie emerges from episodic activity with gain variability. Nat Commun. https://doi.org/10.1038/s41467-021-25437-2

Yang G (2019) Scaling limits of wide neural networks with weight sharing. arXiv preprint arXiv:1902.04760

Yger P, Spampinato GL, Esposito E, et al (2018) A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. eLife 7. https://doi.org/10.7554/eLife.34518

Yu S, Ribeiro TL, Meisel C et al (2017) Maintained avalanche dynamics during task-induced changes of neuronal activity in nonhuman primates. Elife 6(e27):119. https://doi.org/10.7554/eLife.27119

Yuste R (2015) From the neuron doctrine to neural networks. Nat Rev Neurosci 16(8):487–497. https://doi.org/10.1038/nrn3962

Zavatone-Veth JA, Pehlevan C (2021) Exact marginal prior distributions of finite Bayesian neural networks. In: Ranzato M, Beygelzimer A, Liang P et al (eds) Advances in Neural Information Processing Systems, vol 34. Curran Associates Inc

Zavatone-Veth JA, Canatar A, Ruben BS et al (2021) Asymptotics of representation learning in finite Bayesian neural networks. In: Ran-

zato M, Beygelzimer A, Liang P et al (eds) Advances in Neural Information Processing Systems, vol 34. Curran Associates Inc

Zenke F, Poole B, Ganguli S (2017) Continual learning through synaptic intelligence. In: International Conference on Machine Learning, PMLR, pp 3987–3995

Zeraati R, Priesemann V, Levina A (2021) Self-organization toward criticality by synaptic plasticity. Front Phys 9:103. https://doi.org/10.3389/fphy.2021.619661

Ziv Y, Burns LD, Cocker ED et al (2013) Long-term dynamics of CA1 hippocampal place codes. Nat Neurosci 16(3):264. https://doi.org/10.1038/nn.3329