bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Long-term stability of neural activity in the motor system

Kristopher T. Jensen[1,2], Naama Kadmon Harpaz[1], Ashesh K. Dhawale[1,3], Steffen B. E. Wolff[1,4], and Bence P. Ölveczky[1]

[1]Department of Organismic and Evolutionary Biology and Center for Brain Science, Harvard University
[2]Computational and Biological Learning Lab, Department of Engineering, University of Cambridge
[3]Present address: Centre for Neuroscience, Indian Institute of Science, Bangalore, India
[4]Present address: Department of Pharmacology, University of Maryland School of Medicine, Baltimore MD 21201, USA

Abstract

How an established behavior is retained and stably produced by a nervous system in constant flux remains a mystery. One possible solution is to fix the activity patterns of single neurons in the relevant circuits. Alternatively, activity in single cells could drift over time provided that the population dynamics are constrained to produce stable behavior. To arbitrate between these possibilities, we recorded single unit activity in motor cortex and striatum continuously for several weeks as rats performed stereotyped motor behaviors – both learned and innate. We found long-term stability in behaviorally-locked single neuron activity patterns across both brain regions. A small amount of neural drift observed over weeks of recording could be explained by concomitant changes in task-irrelevant behavioral output. These results suggest that stereotyped behaviors are generated by stable neural activity patterns.

Introduction

Learning and memory in dynamic motor circuits

When we wake up in the morning, we usually brush our teeth. Some of us then cycle to work, where we log on to the computer by typing our password. After work, we might go for a game of tennis, gracefully hitting the serve in one fluid motion. These motor skills, and many others, are acquired through repeated practice and stored in the motor circuits of the brain, where they are stably maintained and can be reliably executed even after months of no intervening practice (Krakauer & Shadmehr, 2006; Melnick, 1971; Park & Sternad, 2015). The neural circuits underlying such motor skills have been the subject of extensive study (Churchland et al., 2012; Haith & Krakauer, 2013; Kawai et al., 2015; Peters et al., 2014; Wolpert & Ghahramani, 2000), yet little is known about how they persist over time. Given the stability of the behaviors themselves (Park et al., 2013), a possible solution is to dedicate a neural circuit to a given skill or behavior, then leave it untouched. However, cortical areas undergo continual synaptic turnover even in adult animals (Fu et al., 2012; Holtmaat & Svoboda, 2009; Xu et al., 2009; Yang et al., 2009) and have been shown to change their activity patterns over time, both in the presence and absence of explicit learning (Clopath et al., 2017; Driscoll et al., 2017; Kargo & Nitz, 2004; Peters et al., 2017; Schoonover et al., 2021). While neural circuits in constant flux may facilitate learning of new behaviors and reflect the continual acquisition of new memories and associations (Rule et al., 2019), it seems antithetical to the stable storage of previously acquired behaviors.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing theories and predictions

Two main theories have been put forth to explain the apparent paradox of stable memories in plastic circuits. In the commonly held view that motor control is governed by low-dimensional dynamics (Gallego et al., 2017; Jensen et al., 2021; Shenoy et al., 2013; Vyas et al., 2020), the paradox can be resolved by having a degenerate subspace in which neural activity can change without affecting behavior (Rokni et al., 2007) or task performance (Qin et al., 2021). While this would do away with the requirement for stable activity at the level of single neurons (Figure 1A) (Clopath et al., 2017), it requires any drift in population activity to occur exclusively in the degenerate subspace. Whether and how biological circuits can ensure this without continual practice remains largely unknown (Rule et al., 2019). Absent complete degeneracy, it has also been suggested that the connections from drifting neural populations to downstream circuits could continually rewire to maintain stable motor output and behavior (Rule et al., 2020). Finally, it is worth noting that the degenerate subspace could also be aligned with a particular subset of neurons, in which case these neurons would retain stable activity patterns with the remaining neurons free to exhibit representational drift.

A different way to maintain stable motor output is by constraining the changes in neural circuits such that they do not affect single neuron activity associated with already established behaviors (Duncker et al., 2020; Kao, Jensen, et al., 2021; Qin et al., 2021). In this case, the activity patterns of individual neurons locked to the behavior would remain constant or highly similar over time (Figure 1A) (Clopath et al., 2017). This solution has been observed in the specialized zebra finch song circuit, where neural activity patterns associated with a stereotyped song remain stable for months (Katlowitz et al., 2018). However, zebra finches have a neural circuit dedicated exclusively to learning and generating their one song, with plasticity largely restricted to a ‘critical period’ of development (Sizemore & Perkel, 2011). In contrast, humans and other mammals use the same ‘general’ motor network for a wide range of behaviors – both learned and innate. How such generalist brains maintain the stability of complex behaviors remains to be understood.

It has previously been hypothesized that stable circuit function could be associated with stable single-neuron activity in some brain regions and constrained population dynamics with drifting single neuron activity in others (Rule et al., 2019). In this view, brain regions several synapses removed from the periphery, and with high degrees of redundancy, would be more likely to exhibit representational drift (Driscoll et al., 2017; Rule et al., 2019; Rule & O’Leary, 2022). In contrast, regions closer to the periphery, that serve as information bottlenecks for sensory input and motor output, would maintain more stable representations (Clopath et al., 2017; Rule & O’Leary, 2022). However, it is unclear whether the mammalian brain shows such differences in single neuron stability, and, if so, at which stage of the motor hierarchy stable single-neuron activity emerge. More generally, it remains an open question whether single-unit neural activity patterns in the mammalian motor system remain stable over time (Chestek et al., 2007; Flint et al., 2016; Rokni et al., 2007; Stevenson et al., 2011).

Experimental challenges

Arbitrating between the hypotheses outlined above has been attempted by recording neural activity over time during the performance of well-specified behaviors, either by means of electrophysiology (Carmena et al., 2005; Chestek et al., 2007; Flint et al., 2016; Fraser & Schwartz, 2012; Ganguly & Carmena, 2009; Rokni et al., 2007) or calcium imaging (Driscoll et al., 2017; Katlowitz et al., 2018; Liberti et al., 2016). These studies have come to discrepant conclusions, with some suggesting stable single unit acitivity (Chestek et al., 2007; Flint et al., 2016; Ganguly & Carmena, 2009; Katlowitz et al., 2018; Stevenson et al., 2011), and others reporting changing activity for fixed behaviors (Carmena et al., 2005; Liberti et al., 2016;

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Rokni et al., 2007). It remains unclear whether these discrepancies reflect technical differences in recordings and analyses, or whether they reflect biological differences between behaviors, animals, or circuits as suggested above. Importantly, putative drift in neural activity could be caused by factors not directly related to the mapping between neural activity and motor output. These include unstable environmental conditions or fluctuations in the animal's internal state, including attention, satiety, and motivation (Miller et al., 2014; Sadeh & Clopath, 2022; Willett et al., 2020). Notably, many of these processes, driven by constrained and/or cyclic fluctuations in hormones and/or neuromodulators (Sheppard et al., 2019; Willett et al., 2020), drift around a mean. They can therefore be distinguished from drift in neural circuits by recording for durations longer than the autocorrelation time of the various uncontrolled, or 'latent', processes.

However, high-quality long-term recordings of the same neurons can be technically challenging. In lieu of this, a recent approach has considered the stability of low-dimensional latent neural dynamics over extended time periods. This was done for motor cortex by applying linear dimensionality reduction to recordings from each experimental session followed by alignment of the resultant low-dimensional dynamics (Gallego et al., 2020). While this work suggests that latent motor cortical dynamics underlying stable motor behaviors are stable over time, it does not address the source of this stability. In particular, it remains unclear whether such long-term stable latent dynamics result from drifting single-unit activity within a degenerate subspace that produces the same latent trajectories, or whether it is a consequence of neural activity patterns that are stable at the level of single units.

In this work, we first use a recurrent neural network to demonstrate how long-term single-unit recordings during a stably executed behavior can distinguish between the two main models of how stable behaviors are maintained. We then go on to perform such recordings in rats producing stable behaviors, considering two central nodes of the motor system: motor cortex (MC) and dorsolateral striatum (DLS) (Hunnicutt et al., 2016). Importantly, both MC and DLS are high-dimensional circuits with orders of magnitude more neurons than the lower-level control bottlenecks they project to (Bar-Gad et al., 2003; Oorschot, 1996; Zheng & Wilson, 2002), and they therefore exhibit substantial degeneracy with respect to motor output (Bar-Gad et al., 2003; Kao, Sadabadi, et al., 2021). To probe the degree to which our findings generalize across different classes of behaviors relying on different control circuits, we examine both learned (Fig 1B) and innate behaviors. To minimize sources of neural variability not directly related to behavioral control, we performed our experiments in a highly stable and controlled environment. Additionally, we recorded the animals' behavior at high spatiotemporal resolution to account for any changes in task-irrelevant movements (Figure 1C) (Chestek et al., 2007; Musall et al., 2019). Our combined neural and behavioral recordings revealed that neural circuit dynamics are highly stable at the level of single neurons. The small amount of drift in task-related neural activity could be accounted for by a concomitant slow drift in the behavior. These results suggest that stable behaviors are stored and generated by stable single-unit activity in the motor circuits that drive the learned behavior, and that the neural correlates of behavior are also stable in an innate behavior, which does not directly depend on the motor circuits we record from.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

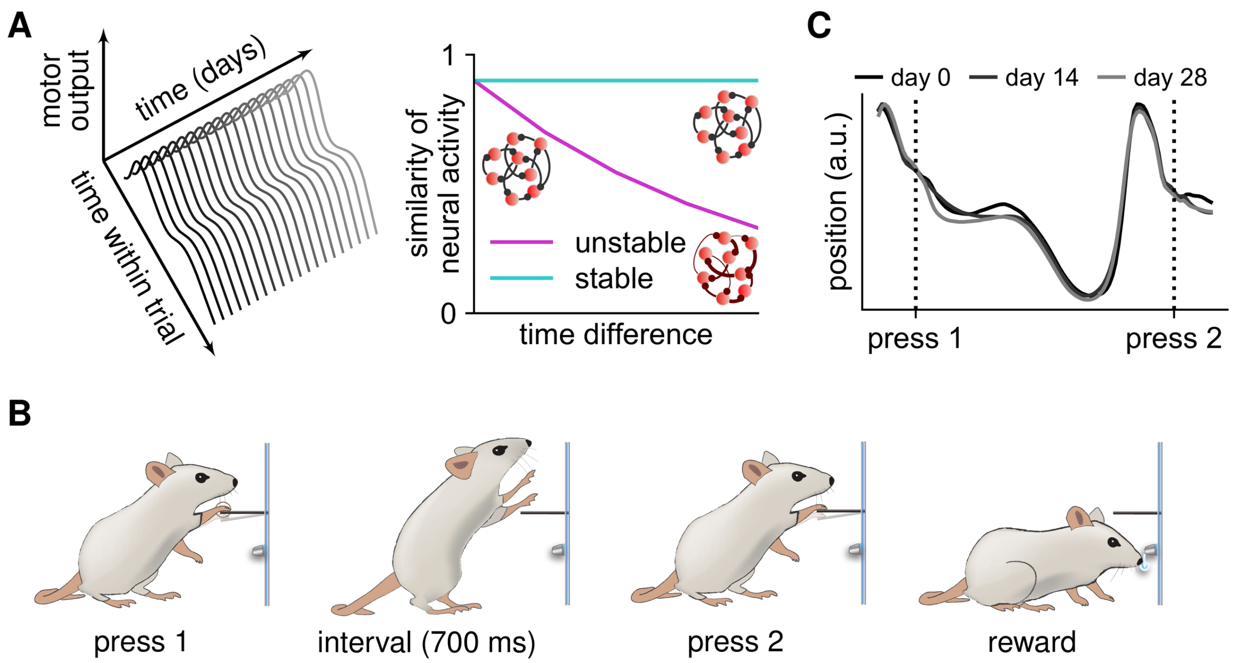


Figure 1: A paradigm for interrogating long-term neural and behavioral stability. (A) For a constant motor output over time (left), the underlying task-related neural activity can either remain stable or change along a behavioral 'null direction' (right) (Kao, Sadabadi, et al., 2021). If single neuron activity is stable over time, similarity of the firing patterns associated with two trials of a stable behavior should not depend on the time separating the trials (cyan). This can be achieved through stable connectivity (RNN insets). Conversely, if the single neuron activity patterns driving the behavior change over time, the similarity of task-associated neural activity should decrease with increasing time difference (magenta) (Clopath et al., 2017). **(B)** Schematic illustration of the task used to train complex stereotyped and stable movement patterns in rats (Kawai et al., 2015). To receive a reward, rats must press a lever twice separated by an interval of 700 ms. **(C)** The mean task-related forelimb trajectory for an example rat on three days, each two weeks apart. Y-axis indicates horizontal forelimb position (parallel to the ground).

Results

Network models of stable and unstable motor circuits

When analyzing the stability of task-associated neural activity, it is important to consider stability not only at the population level (e.g. in the form of stable latent dynamics), but also at the level of single task-associated neurons. We start by simulating a degenerate artificial control network to highlight this distinction and motivate the use of longitudinal single-unit recordings to address the neural mechanisms of long-term behavioral stability. Our simulations show how 'stable' and 'drifting' single-neuron activity can both drive stable latent dynamics and behavior, providing validation and motivation for the experimental strategy and analyses considered below (Lindsay, 2022). Our simulated neural circuit was a recurrent neural network (RNN) producing a stereotyped output, akin to those previously used to model pattern generator functions (Hennequin et al., 2014, 2018; Laje & Buonomano, 2013; Sussillo & Abbott, 2009) (Methods). We trained the network using gradient descent (Kingma & Ba, 2014) to generate five smooth target output trajectories (Figure 2A; Methods). After training, we simulated the noisy dynamics for 100 trials (Methods) and generated spikes from a Poisson observation model to constitute a simulated experimental 'session' (Figure 2B, C).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Importantly, due to the degeneracy of the circuit (250 neurons with 60,000 parameters controlling a 5-dimensional time varying output), multiple distinct networks with different single-unit activity patterns can achieve the same target output. This allowed us to compare network dynamics of RNNs producing the same output with either identical or differing connectivity. When comparing the peri-event time histograms (PETHs) of individual units from identical networks across separate simulated 'sessions', the activity of most units was highly correlated as expected (Figure 2D; left). We then compared the activity of neurons across two different networks trained independently to generate the same output and found a near-uniform distribution of PETH correlations with a mean close to zero (Figure 2D; left). Thus, while individual units from the same network had similar activity profiles in different sessions, units from different networks were, on average, only weakly correlated due to the heterogeneity of the underlying activity patterns.

We compared this measure of single-unit similarity with the similarity of aligned latent dynamics across sessions (Gallego et al., 2020). To do this, we simulated the experimental scenario in which only a subset of the total population is tracked by randomly sampling 50 neurons. We further reduced the dimensionality of each recording from 50 to 10 using PCA and aligned the resulting latent trajectories by applying canonical correlation analysis (CCA) to each pair of simulated sessions, using the canonical correlations as a measure of similarity (Methods) (Gallego et al., 2020). In contrast to the single-unit correlations, the latent dynamics were highly similar between pairs of identical and pairs of different networks (Figure 2D). This reflects the fact that even though distinct networks differ in the activity patterns of individual neurons, they have similar population level statistics due to the conserved nature of the task.

To intuit how activity patterns change over time in an unstable network, we performed a linear interpolation between the parameters of the two independently trained RNNs. Re-optimizing the recurrent weight matrix while fixing the readout weights for each network in the interpolation series led to 7 RNNs with progressively more dissimilar connectivity, yet which all produced the same output – a phenomenological model of neural drift, where the position of a network within the interpolation series is a proxy for time (Figure 2F; Extended Data Fig. 1; Methods). This resulted in a recurrent network with parameters that turned over in a simulated 'week' of recording, comparable to the timescale of drift reported in previous studies (Driscoll et al., 2017; Liberti et al., 2016).

We proceeded to investigate the degree to which single-unit activity changed as a function of this measure of time. When inspecting the activity of individual units in the RNNs, we found that their firing patterns tended to change from session to session, with sessions close in time generally exhibiting more similar firing patterns than distant sessions (Figure 2E). To quantify this at the population level, we computed the correlation between single-unit PETHs for all pairs of sessions. This measure of similarity exhibited a systematic decrease as a function of time difference between sessions (Figure 2F). For comparison with a stable network, we performed an interpolation as above, but now between two instances of the same RNN, such that the changes in connectivity corresponded to fluctuations around a single local minimum (Methods). As expected, the network output and the single-unit neural activity in this stable network were highly stable over time (Figure 2F).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

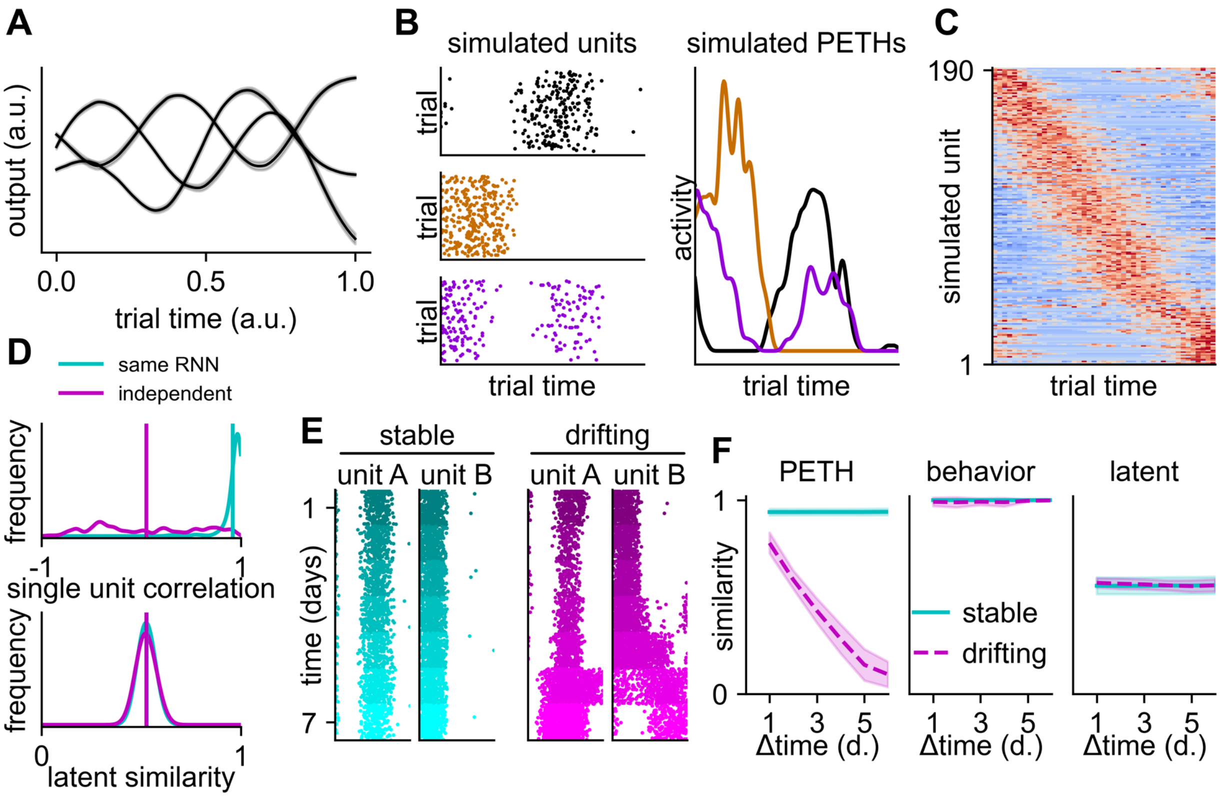


Figure 2: Analyzing neural stability in a recurrent network model. (A) Example RNN outputs after training (mean across 100 simulated trials). **(B)** Activity of three example recurrent neurons after training (Methods). (Left) Raster plot of spike times across 100 trials. (Right) Peri-event time histograms (PETHs) computed across the corresponding trials. **(C)** PETHs for all units firing at least 100 spikes, sorted according to the PETH peak from a set of held-out trials and plotted as a heatmap with color indicating spike count from low (blue) to high (red). **(D; top)** Distribution across neurons of the correlation between two PETHs constructed from separate sets of trials from the same network (cyan). A second RNN was independently trained to produce the same output, and PETH correlations were approximately uniformly distributed between these distinct networks (magenta). Vertical lines indicate average similarity. **(D; bottom)** Distribution of latent similarities between the same (cyan) or different (magenta) networks after alignment using CCA, considering 250 random samples of 50 neurons. **(E)** Example raster plots as in (B) across 7 different sessions (y-axis) for a network exhibiting either stability (left) or drifting neural activity (right). **(F)** Quantification of the similarity in the space of PETHs (left), network output (middle), and aligned latent trajectories (right) as a function of time difference (change in y value from (E)) for the stable RNN (cyan) and the drifting RNN (magenta). Lines and shadings indicate mean and standard deviation across 10 networks.

Finally, we again considered how such single-unit analyses differ from approaches that consider the stability of low-dimensional latent dynamics (Gallego et al., 2020). Similar to the single-unit analyses, we computed the neural similarity as a function of time difference, but now with similarity measured as the correlation between aligned latent trajectories as described above. As expected for a network with constant output, the latent dynamics of the RNN with drifting single unit activity did not become more dissimilar over time. These observations reflect the fact that the network retained similar population statistics despite the changing activity profiles of individual units (Figure 2F). They are also consistent with work in visual cortex showing that the low-dimensional structure of neural activity can be preserved despite drifting single-unit activity patterns (Deitch et al., 2021). Our artificial model system thus highlights the importance of long-term recordings of single units to complement studies of latent space stability when investigating how stable behaviors are generated. In particular, we illustrate how stability at the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

level of single units implies stability at the level of latent trajectories, while the stable latent dynamics reported in previous work (Gallego et al., 2020) can be driven by either stable or drifting single-unit activity patterns.

Long-term recordings of neural activity and kinematics during a learned motor task

To investigate the stability of biological motor circuits experimentally, we trained rats (n=6) to perform a timed lever-pressing task in which they received a water reward for pressing a lever twice with an inter-press interval of 700 ms. Rats learned to solve the task by developing complex stereotyped movement patterns (Figure 3A, Extended Data Figure 2A) (Dhawale et al., 2021; Kawai et al., 2015). Since the task is kinematically unconstrained (meaning it has many 'motor solutions') and acquired through trial-and-error, each animal converged on its own idiosyncratic solution (Figure 3B). However, once acquired, the individually distinct behaviors persisted over long periods of time (Figure 3A).

To reduce day-to-day fluctuations in environmental conditions that could confound our assessment of neural stability over time, animals were trained in a fully automated home-cage training system with a highly regimented training protocol in a very stable and well-controlled environment (Poddar et al., 2013). After reaching expert performance, animals were implanted with tetrode drives for neural recordings (Dhawale et al., 2017) targeting motor cortex (MC) and dorsolateral striatum (DLS) (Hunnicutt et al., 2016) (Methods). While the stability of single units in cortical regions has previously been addressed with inconsistent findings (Chestek et al., 2007; Clopath et al., 2017; Dhawale et al., 2017; Rokni et al., 2007; Stevenson et al., 2011), studies of neural stability in sub-cortical regions, and specifically the striatum, are scarce (Kubota et al., 2009; Sheng et al., 2019). DLS is, in this case, particularly relevant as it is essential for the acquisition and control of the motor skills we train (Dhawale et al., 2021).

Three animals were implanted in Layer 5 of MC, and three animals in DLS. Following implantation and recovery, animals were returned to their home-cage training boxes and resumed the task. Neural activity was then recorded continuously over the course of the experiment (Dhawale et al., 2017). Importantly, our semi-automated and previously benchmarked spike-sorting routine (Dhawale et al., 2017) allowed us to track the activity of the same neurons over days to weeks in both DLS and MC (Figure 3C, 3D). The task-relevant movements of all animals were tracked using high-resolution behavioral recordings (Insafutdinov et al., 2016; Mathis et al., 2018), and both behavior (kinematic features) and neural activity were aligned to the two lever-presses to account for minor variations in the inter-press interval (Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

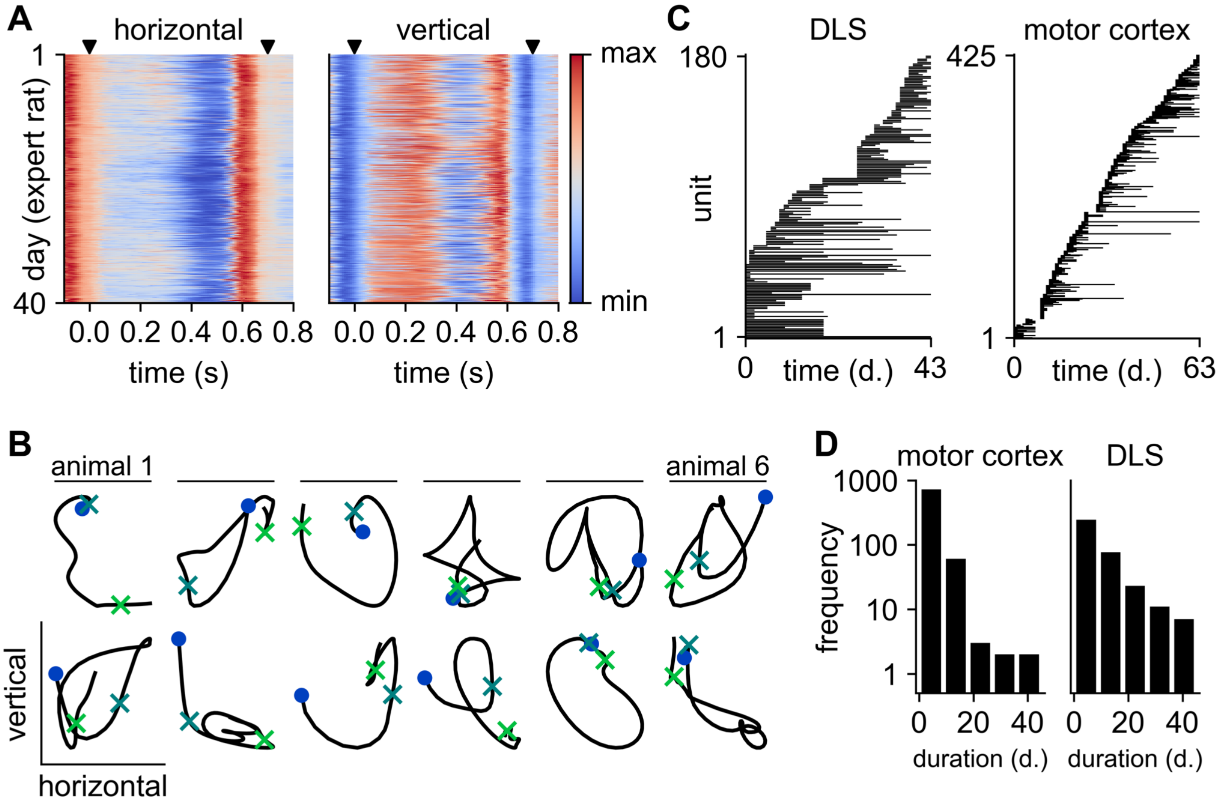


Figure 3: **Experimental recordings of behavior and neural activity. (A)** Right forelimb trajectories in the horizontal and vertical directions (c.f. Figure 1C) for an example expert rat (see Extended Data Figure 2 for data from the other 5 animals). Color indicates forelimb position. Kinematics were linearly time-warped to align the two lever-presses for all analyses (Methods; warping coefficient = 1.00 ± 0.07), and black triangles indicate the times of the lever presses. The rat uses the same motor sequence to solve the task over many days with only minor variations. **(B)** Mean trajectories across all trials of the left (top row; left side view) and right (bottom row; right side view) forelimbs for each rat (columns), illustrating the idiosyncratic movement patterns learned by different animals to solve the task. Circles indicate movement initiation; dark and light green crosses indicate the times of the 1st and 2nd lever press respectively. **(C)** Time of recording for each unit for two example rats recording from DLS (left) and MC (right). Units are sorted according to the time of first recording. **(D)** Distribution of recording times pooled across units from all animals recording from DLS (left) or MC (right). Note that the data used in this study has previously been analyzed by Dhawale et al. (Dhawale et al., 2017, 2021).

Behaviorally locked activity of single neurons in MC and DLS is stable

The combination of controlled and regimented experimental conditions, stable behavior, and continuous neural recordings provides a unique setting for quantifying the stability of an adaptable circuit driving a complex learned motor behavior (Clopath et al., 2017). Importantly, this experimental setup mirrors the scenario considered in our RNN model (Figure 2) and thus facilitates analyses of neural stability at the level of single neurons. We first considered the PETHs of all units combined across all trials and found that units in both MC and DLS fired preferentially during particular phases of the learned behavior (Figure 4A) (Dhawale et al., 2017). Importantly, we found that the behaviorally locked activity profiles of individual units were highly stable over long periods of time (Figure 4B), reminiscent of the 'stable' RNN model (Figure 2C).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To quantitatively compare neural activity profiles across days, we constructed PETHs for each neuron by summing the spike counts across all trials on each day and convolving them with a 15 ms Gaussian filter (Figure 4B; Methods). We then computed the Pearson correlation ρ between pairs of PETHs across different recording days as a function of the time difference between days (Extended Data Figure 3A), similar to our RNN analyses (Figure 2) and to previous studies in visual and motor circuits (Deitch et al., 2021; Dhawale et al., 2017). When considering neurons recorded for at least two weeks, the mean PETH similarity remained high in both DLS and MC (Figure 4C; see Extended Data Figure 4 for other recording thresholds). This is consistent with results from the stable RNN model (Figure 2D), and it suggests that learned motor behaviors are driven by single neuron activity patterns that do not change over the duration of our recordings, despite the life-long structural and functional plasticity in these circuits (Holtmaat & Svoboda, 2009; Peters et al., 2017; Wolff et al., 2019; Xu et al., 2009).

To see how this compares to a hypothetical circuit where population statistics are retained but individual neurons change their firing patterns, we also computed pairwise correlations between non-identical neurons recorded on different days. These correlations were near zero in both MC and DLS, confirming that the high correlation over time for individual units is not due to a particular population structure of neural activity imposed by the task (Figure 4C). These results suggest that neural activity associated with the learned motor skill is qualitatively stable over periods of several days and weeks (Figure 4B, 4C). Our findings also suggest that the stable latent dynamics identified in previous work (Gallego et al., 2020) could be a result of such stable single-unit dynamics (c.f. Figure 2F), which is further supported by the fact that alignment of the neural dynamics using CCA did not increase stability further (Extended Data Figure 5A).

In contrast to our RNN model, the experimental data contained neurons that were recorded for different durations (Figure 3D). This introduces additional variability and makes it difficult to assess stability across neurons without either discarding neurons recorded for short durations or losing information about neurons recorded for long durations. To combine information across more neurons, we instead considered the PETH similarity as a function of the time difference between PETHs for each neuron individually. An exponential model of the form $\rho = \beta e^{\alpha \delta t}$ was fitted to the Pearson correlation (ρ) between PETHs as a function of time difference δt for each neuron (Methods; see Extended Data Figure 6 for example fits). We denote $\alpha = -\tau^{-1}$ as the 'stability index' since it corresponds to the negative inverse time constant τ in an exponential decay model, and this stability index provides a single parameter summarizing the rate of drift for each neuron.

We then considered the distribution of stability indices across neurons recorded for at least 4 days. In a null-model where single-neuron activity remains constant, the PETH similarity should be independent of the time difference for all units (c.f. Figure 2F). The stability indices should thus be centered around zero with some spread due to trial-to-trial variability, corresponding to an infinitely slow exponential decay. The population-level distributions over α were indeed centered near zero (Figure 4E). However, a permutation test across time differences revealed that all DLS recordings and two of the animals with recordings from MC did in fact exhibit slow but significant neural drift ($p < 0.05$). We saw this also when combining data for all neurons across animals within each experimental group (DLS: $\alpha_{median} = -0.014$, $\tau_{median} = 71$ days, $p < 0.001$; MC: $\alpha_{median} = -0.027$, $\tau_{median} = 37$ days, $p < 0.001$; permutation test).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

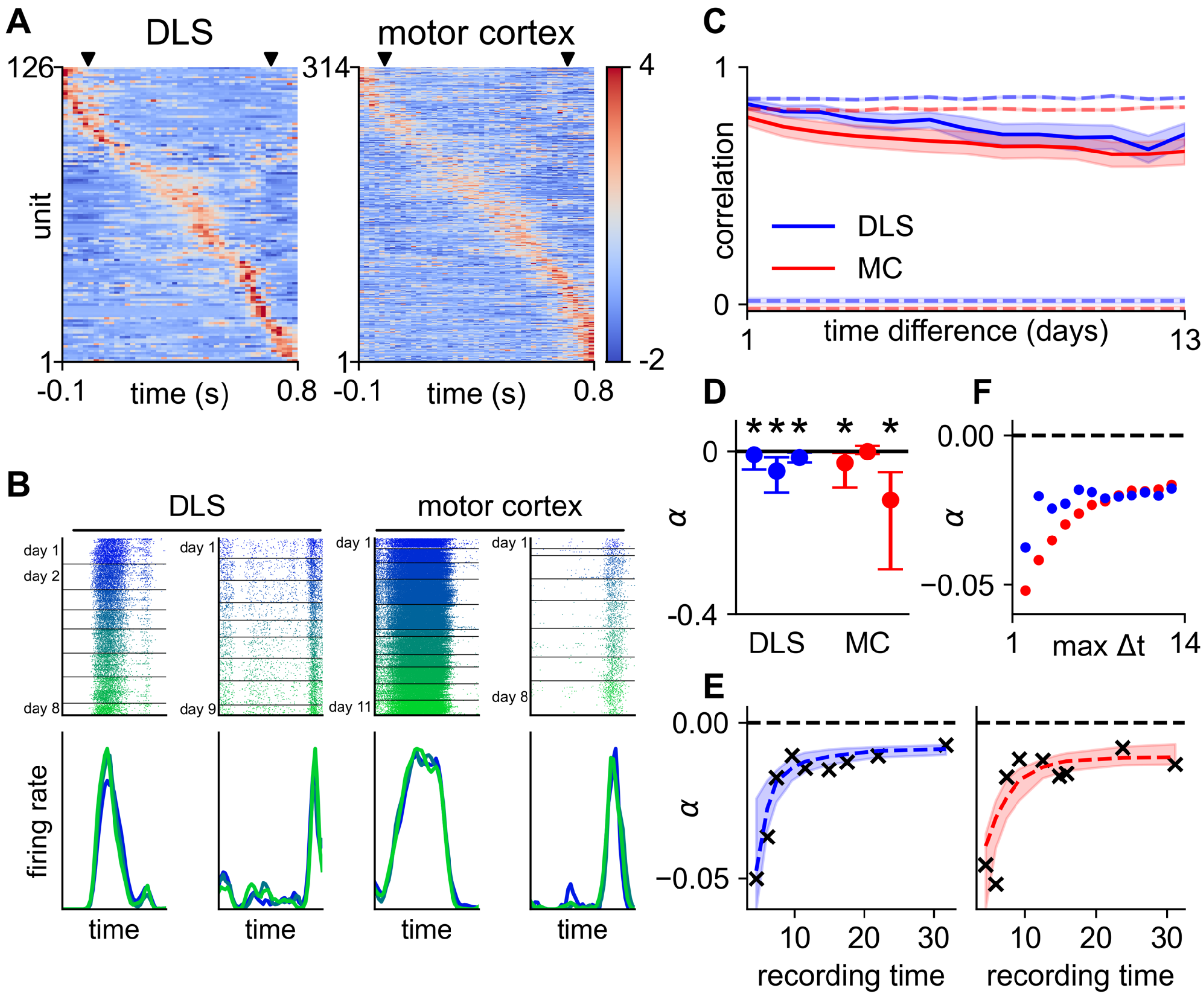


Figure 4: Single-unit activity is stable over time in DLS and MC. (A) z-scored PETHs across trials and sessions for all units firing at least 100 spikes during the lever-pressing task for two example rats. Units were sorted according to the activity peak from a set of held-out trials. X-axis indicates time-within-trial relative to the first lever press, spike times were linearly time-warped to align the two lever presses (Methods), and black triangles indicate the times of the presses. **(B; top)** Raster plots for two example units in DLS (left) and MC (right) illustrating firing patterns that are time-locked to the behavior over days to weeks. Horizontal lines indicate the beginning of a new day, and color indicates the progression of time from day 1 (blue) to the last day of recording (green). **(B; bottom)** Normalized PETHs for the four example units computed on three different days (early, middle, late) with corresponding colors in the raster. Our quantification of neural similarity is based on the correlations between such PETHs. **(C)** Mean value of the correlation between PETHs calculated on separate days, averaged over all units recorded for at least 14 days from MC (red) or DLS (red) and plotted as a function of time between days (n = 38 neurons for DLS; n = 27 neurons for MC; see Extended Data Figure 4A for different recording thresholds). Shaded regions indicate standard error across units. Colored dashed lines indicate the similarity between non-identical neurons (lower) and in a resampled dataset with neural activity drawn from a stationary distribution (upper; Methods). **(D)** Quartiles of the distribution of stability indices for each animal. Horizontal dashed line indicates $\alpha = 0$. Asterisks indicate $p < 0.05$ for the median stability index being smaller than zero (permutation test; Methods; n = [88, 22, 13] neurons for DLS; [122, 6, 7] for MC). **(E)** Rolling median of the stability index (crosses) for units recorded for different recording durations. Bins are overlapping with each neuron occurring in two bins (see Extended Data Figure 7 for the non-binned data). Dashed lines indicate exponential model fits to the non-binned data, and shadings indicate interquartile intervals from bootstrapping the units included in the model fits (Methods). **(F)** Stability indices of models fitted to increasing subsets of the data from (C), illustrating how longer recording durations lead to longer time constants. The maximum

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

time difference considered for the model fit is indicated on the x-axis (see Extended Data Figure 8 for the full model fits).

Short recording durations underestimate the stability of neural activity

Our analyses at the level of single neurons included units recorded for short durations of time. However, as noted in the introduction, recording over such short time spans could underestimate stability in the presence of latent processes that affect circuit dynamics, even if the circuit itself is stable. Such processes may vary over timescales of hours or days (Kanwal et al., 2021) but be constrained over longer timescales by homeostatic mechanisms, biological rhythms, or task constraints (Miller et al., 2014; Sheppard et al., 2019; Willett et al., 2020). Even though such bounded physiological fluctuations will manifest as short-term drift in neural firing patterns, their contributions to estimates of neural stability will diminish as neural recording durations exceed the characteristic timescales of the underlying processes.

To better estimate drift over longer timescales, we binned the stability indices of all neurons by their recording duration. This revealed that the apparent stability ranged from $\alpha \approx -0.05$ for short recording durations to $\alpha \approx -0.01$ for long recording durations (Figure 4F). To extrapolate to longer recording durations, we fitted an exponential model to the data of the form $\alpha = -a - b\exp(-c\ t)$ (Figure 4F). The parameter $\tau_\infty = a^{-1}$ provides an estimate of the asymptotic stability of the population and took values of $\tau_\infty = 139$ days for DLS and $\tau_\infty = 92$ days for motor cortex (interquartile ranges of 115-211 for DLS and 75-161 for MC; bootstrapped model fits; Methods).

To confirm that this apparent increase in stability with recording duration was not due to a bias in our data collection, we returned to the average similarity across neurons recorded for at least 14 days (Figure 4C). We subsampled the data from these neurons to different maximum time differences, thus varying the effective recording duration for a fixed set of neurons. We then computed stability indices by fitting our exponential model to the average correlation across neurons as a function of time difference (Methods; Extended Data Figure 8). The stability indices increased with subsampled recording duration in both DLS and MC (Figure 4D), consistent with the results across the whole population of neurons (Figure 4F). These findings suggest the presence of constrained fluctuations in the physiology or behavior of the animals, which affect estimates of neural stability on shorter timescales. Our results therefore motivate long-duration single neuron recordings for estimating long-term neural stability.

Finally, if the observed increase in stability with recording duration is due to latent processes with autocorrelations on the order of days, we would expect the neural similarity to decrease to some saturating baseline value. We therefore proceeded to fit a model to the average similarity across neurons as a function of time difference, which assumes a decay to a constant baseline ($\rho = \beta e^{\delta t/\tau} + \gamma$; Extended Data Figure 8). This yielded an asymptotic correlation of $\gamma = 0.61$ for DLS and $\gamma = 0.62$ for MC, suggesting a high degree of neural similarity even at very long timescales. In summary, we find that estimates of stability can be biased by short recording durations, possibly due to the presence of physiological or behavioral processes with characteristic timescales on the order of hours and days. Extrapolating to longer recording durations, we found the average drift to be on a timescale of 100 days in a single-timescale exponential model (Figure 4F). When considering a three-parameter model with a baseline, we found evidence of a decay to a constant above-chance similarity of $\rho \approx 0.6$. These findings suggest that motor memories are retained by maintaining stable task-associated activity patterns over several months.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Neural drift is correlated with behavioral changes

In the previous section, we quantified the stability of motor circuits during a learned behavior and showed how such estimates can be affected by internal or external processes that affect neural activity. However, any residual drift in neural activity is still likely to be an underestimate of the true stability in the mapping from circuit activity to motor output. Indeed, even a perfectly stable neural system should exhibit drifting neural activity patterns locked to the behavior if the behavior itself is changing (Chaisanguanthum et al., 2014; Chestek et al., 2007). This tends to be the case even after performance saturates and stabilizes in a motor task. In particular, humans and animals alike exhibit small behavioral changes both in terms of trial-to-trial variability (Churchland, 2015) and systematic drifts in the mean behavioral output (Chaisanguanthum et al., 2014). If such systematic behavioral drift is present in the motor task we analyze, it could explain some of the short-timescale constrained drift as well as any residual drift in neural activity over longer timescales (Figure 4). This, in turn, would suggest a more stable circuit linking neural activity to behavior than revealed by analyses of neural data alone. Thus, to quantify the degree to which the neural drift we see can be accounted for by accompanying changes in task-related motor output, we proceeded to analyze the kinematics of the timed lever-pressing behavior and how they changed over time.

We first examined whether minor behavioral changes could be observed in the motor output by visualizing the z-scored forelimb velocities; that is, we subtracted the mean velocity across all trials for each time point and normalized by the standard deviation. This discarded the dominant mean component of the motor output and revealed a slow drift in the behavior over periods of days to weeks (Figure 5A, Extended Data Figure 2B). To quantify this drift, we computed the correlation between mean forelimb velocities across trials on a given day as a function of the time separating each pair of days. This confirmed the presence of a small but consistent decrease in the similarity of motor output as a function of time difference (Figure 5B; Extended Data Fig. 9). Importantly, such behavioral drift occurred despite stable task performance (Extended Data Fig. 10). This is consistent with previous work considering behavioral variability in expert performers as an underlying ‘random walk’ in behavioral space with added motor noise (Chaisanguanthum et al., 2014), although our work considers drift over the course of several weeks rather than hours. If the physical environment remains unchanged, any long-term behavioral drift must ultimately arise from changes in neural activity. Additionally, DLS is known to be involved in driving the behavioral output during this lever pressing task (Dhawale et al., 2021). These considerations suggest that the observed neural drift could be in a motor potent subspace and reflect the changing behavioral output. To investigate this, we followed previous work (Chestek et al., 2007; Driscoll et al., 2017) and showed that the performance of a decoder predicting behavior from neural activity did not deteriorate over time (Extended Data Figure 5B). However, this analysis only considered a small subset of our data, where a large population of neurons was recorded simultaneously (c.f. Figure 3C), and it has previously been shown that decoding analyses looking at ‘effective connectivity’ can suffer from omitted variable bias (Mehler & Kording, 2020; Stevenson, 2018). We therefore proceeded to investigate the relationship between neural and behavioral drift at the level of single neurons.

We controlled for the confounding factor of time (Marinescu et al., 2018) by computing both the similarity of neural PETHs and the similarity of forelimb velocity profiles for each pair of consecutive days. We then exploited the fact that behavioral drift varies across days (Figure 5A) and computed the correlation between neural and behavioral drift rates across all consecutive days for each neuron. This correlation should be positive if the drift in neural activity is related to the drift in motor output (Figure 5C). The analysis was repeated for all units to generate a distribution over correlations between the drift in neural activity and the drift in forelimb kinematics (Figure 5D). The mean of the distribution was $\bar{\rho} = 0.33$ for DLS and $\bar{\rho} = 0.26$ for MC, both of which were significantly larger than a null distribution generated by

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

permuting the behavioral data to break any correlations with the neural drift (Figure 5E; p < 0.001; permutation test). This finding confirms that the shorter timescale drift in neural activity is directly related to changes in behavior and suggests that neural drift could be even slower for behaviors with stronger kinematic constraints.

We proceeded to investigate how this experimental correlation compared to a hypothetical system where the drift in neural activity was driven entirely by the drift in behavior, i.e., where there was a stable mapping between single unit activities and behavioral output. To do this, we fitted a linear-nonlinear Poisson GLM (Pillow et al., 2008) to predict neural activity from behavior using data from a single day of recording for each unit. This model was used to generate synthetic neural activity on each trial from the recorded behavior, allowing us to compute the correlation between the simulated neural drift and the experimentally observed behavioral drift. Here, we found an average correlation with behavior of $\bar{\rho} = 0.40$ for DLS and $\bar{\rho} = 0.24$ for MC. The correlation values found in the experimental data were substantially more similar to this stable synthetic circuit than to the null distribution with no relation between the drift in neural activity and behavior (Figure 5E; p = 0.06 and p = 0.73 for the synthetic average correlation being smaller than the experimental value).

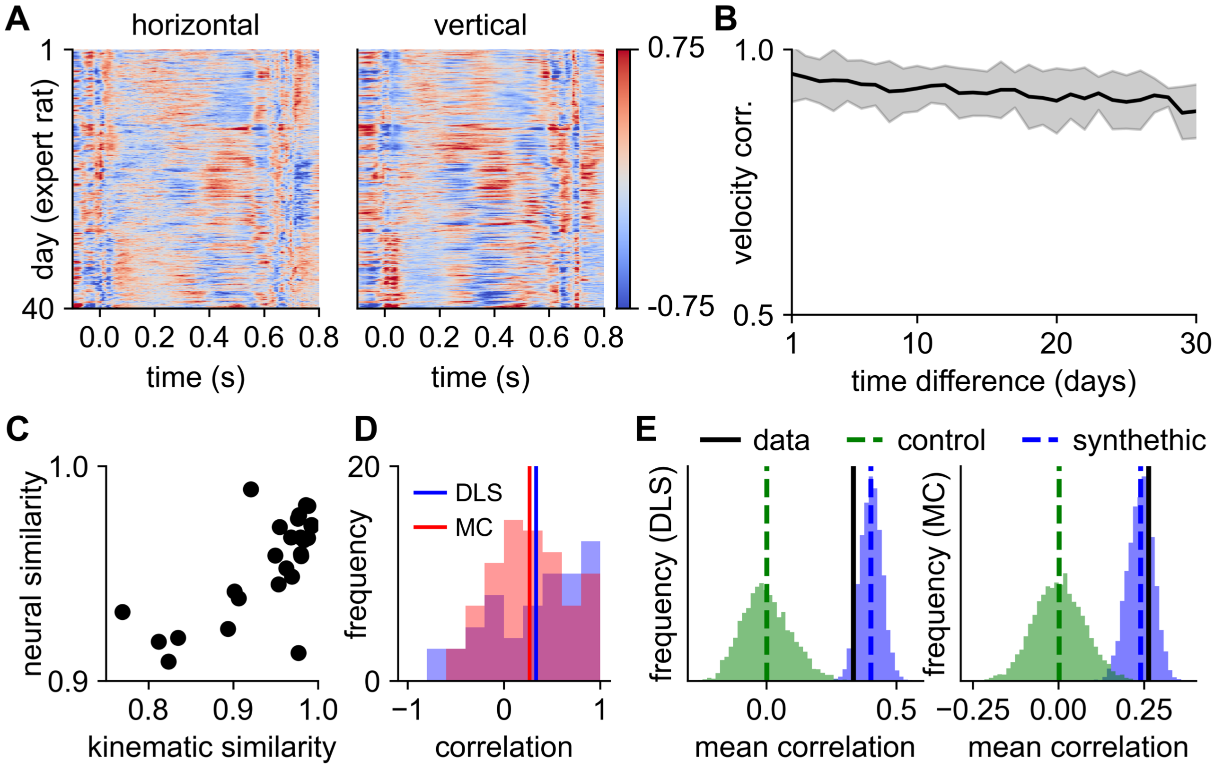


Figure 5: Long-term drift of task specific movement patterns in the lever-pressing task. (A) Forelimb velocities for the example animal from Figure 3A, plotted as z-scores with the column-wise mean subtracted (see Extended Data Figure 2B for the other 5 animals). The stereotyped motor sequence masks a behavioral drift across days and weeks. **(B)** Mean and standard deviation of behavioral similarity as a function of time difference, averaged across all pairs of days for the example animal in (A) (see Extended Data Fig. 9 for data across all animals). **(C)** Similarity between PETHs on consecutive days plotted against the similarity in kinematic output across the corresponding days for an example unit. Each point corresponds to a single pair of consecutive days. **(D)** Distribution of the correlation between neural similarity and behavioral similarity on consecutive days for neurons recorded in DLS (blue) and MC (red). Vertical lines indicate average correlations. **(E)** Mean correlation between neural and behavioral similarity across neurons (black). Green histogram indicates a control distribution constructed by permuting the indices of the days

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in the behavioral data. Blue histogram indicates the distribution of correlations in synthetic datasets where neural activity is determined entirely by behavior via a GLM. Left panel includes all units from DLS, right panel from MC.

Taken together, our behavioral analyses confirm that the drift in neural activity is driven, at least in part, by a concomitant task-irrelevant behavioral drift. Additionally, the correlation between the drift in neural activity and behavior is comparable to a synthetic system where the systematic changes in neural activity are exclusively caused by behavioral changes. This suggests that behavioral changes can account for much of the experimentally observed neural drift.

Neural activity remains stable during an innate behavior

The majority of studies on neural stability have considered behaviors that are either learned or adapted to artificial settings, such as navigating a maze (Driscoll et al., 2017), reaching for points on a screen (Chestek et al., 2007; Flint et al., 2016; Rokni et al., 2007), controlling a BCI (Carmena et al., 2005; Flint et al., 2016; Ganguly & Carmena, 2009), or singing a song (Katlowitz et al., 2018; Liberti et al., 2016). However, many of the behaviors we express are species-typical, or 'innate'. For example, sneezing, crying, and shivering require intricate patterns of sequential muscle activity but are not consciously controlled or learned. While we know less about the neural circuits controlling such innate behaviors, we can probe the stability with which they are encoded and compare them to behaviors that explicitly require plasticity. We therefore considered an innate behavior in the rat known as the 'wet-dog shake' (WDS), which is characterized by whole-body oscillations (Fletcher & Harding, 1981; Kleinrok & Turski, 1980; Marshall et al., 2021; Martin et al., 1963). Importantly, while we know that MC and DLS are necessary for learning (Kawai et al., 2015) and executing (Dhawale et al., 2021) the stereotyped motor patterns required for mastering the lever-pressing task, the wet-dog shake is generated by circuits downstream of DLS and MC (Bedard & Pycock, 1977). Under the hypothesis that degenerate or redundant circuits exhibit a higher degree of drift associated with a given behavior, we might therefore expect less neural stability for the wet-dog shakes compared to the learned lever-pressing task. Alternatively, if the sensorimotor circuits maintain a stable mapping to behavior more generally, we should expect activity patterns in MC and DLS to be stable also in relation to the WDS behavior, albeit perhaps with less behaviorally time-locked firing since these brain regions are dispensable for the behavior.

Given the stereotyped frequency of the WDS events, it is possible to identify them using an accelerometer attached to the head of each animal (Methods). Each WDS event lasted approximately half a second, and each animal performed on the order of 50 WDS per day. This allowed us to analyze them in a 'trial-like' manner, similar to our analyses of the lever-pressing task. We found that the accelerometer readouts corresponding to WDS events were consistent across trials over long time periods (Figure 6A), and we identified units in both DLS and MC whose activity was locked to the behavior. Consistent with the stable neural activity patterns observed during the learned lever-pressing task, the neurons exhibited qualitatively similar firing patterns over time (Figure 6B), though there was weaker task modulation overall (Extended Data Fig. 11). When computing PETH correlations over time, we also found that they remained stable throughout the period of recording (Figure 6C, Extended Data Figure 3B & 4B), although the baseline trial-to-trial similarity was lower than for the learned motor patterns associated with the timed lever-pressing task. These results are consistent with a lesser (or no) involvement of DLS and MC in the specification and control of WDS (Bedard & Pycock, 1977). Hence, the observed activity patterns in MC and DLS during WDS, and consequently the high degree of stability over time, is likely to reflect the stability in the sensorimotor system as a whole, including in the behaviorally-locked activity of connected areas which presumably process sensory feedback and motor efference (Hatsopoulos & Suminski, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

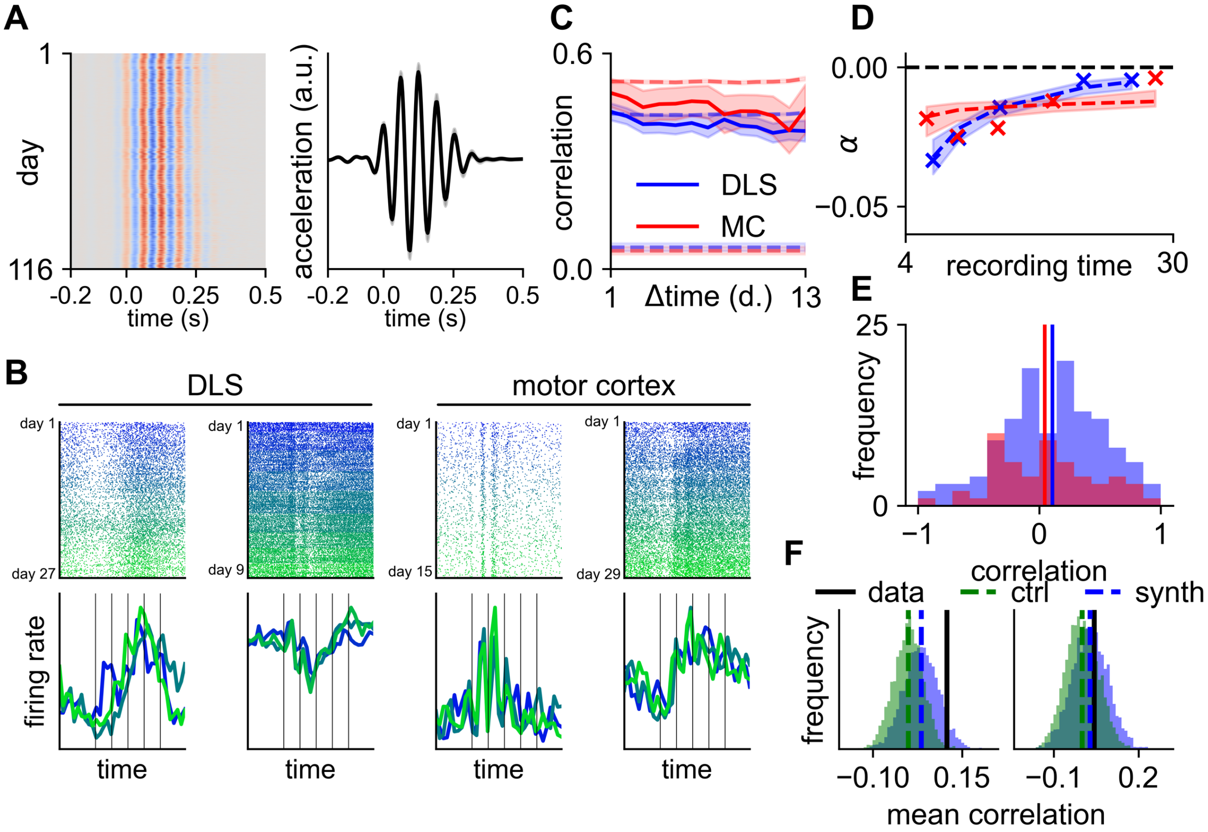


Figure 6: Neural activity is stable during an innate behavior (WDS). (A; left) Vertical acceleration across all 12,775 wet-dog shakes (WDSs) recorded over 116 days in an example animal. Each row corresponds to a single 'trial'. **(A; right)** We computed the mean acceleration across trials for each day. Line and shading indicate the mean and standard deviation across all days as a function of time-within-trial. All kinematics and spike times were linearly time-warped to the median WDS frequency for each animal (Methods; warping coefficient = 1.01 ± 0.07). **(B; top)** Raster plots for two example units in DLS (left) and MC (right) illustrating units with a firing pattern that is time-locked to the behavior over timescales of days to weeks. Color indicates the progression of time from day 1 (blue) to the last day of recording (green). **(B; bottom)** PETHs computed on three different days (early/middle/late) for each of the four example units. Vertical lines indicate peaks in the accelerometer trace. **(C)** Mean value of the correlation between PETHs calculated on separate days, averaged over all units recorded for at least 14 days in MC (red; n = 17) or DLS (blue; n = 66) and plotted as a function of time difference (see Extended Data Figure 4B for other recording thresholds). Shadings indicate standard error across units. Colored dashed lines indicate the similarity between non-identical neurons (lower) and in a resampled dataset with neural activity drawn from a stationary distribution (upper; Methods). **(E)** Rolling median of the stability index (crosses) for units with different recording durations. Bins are overlapping, with each neuron occurring in two bins (see Extended Data Figure 7B for the non-binned data). Dashed lines indicate exponential model fits to the non-binned data, and shadings indicate interquartile intervals from bootstrapping the units included in the model fits (Methods). **(E)** Distribution of the correlation between neural similarity and behavioral similarity on consecutive days for neurons recorded in DLS (blue) or MC (red). Vertical lines indicate the average correlations. **(F)** Mean across units of the correlation between neural and behavioral similarity on consecutive days (black). Green histogram indicates a control distribution from permuting the indices of the days in the behavioral data. Blue histogram indicates the distribution of correlations in synthetic datasets where neural activity is determined entirely by behavior via a GLM. Left panel includes all neurons from DLS, right panel from MC.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To quantify the degree of stability for the population of recorded neurons, we computed stability indices for each neuron. Similar to our observations in the lever-pressing task, the stability indices were centered near zero, indicating largely stable circuits, but with a slow decay over time (DLS: $\alpha_{median} = -0.014$, $\tau_{median} = 70$ days, p < 0.001, n = 180 neurons; MC: $\alpha_{median} = -0.016$, $\tau_{median} = 63$ days, p < 0.001, n = 99 neurons; permutation tests). We expected that this apparent drift would be partly due to our finite recording durations as for the lever-pressing task. Consistent with this hypothesis, fitting an exponential model with bootstrapping to the experimental data and extrapolating to infinite recording durations suggested a much longer median asymptotic timescale of $\tau_\infty = 467$ days for neural drift in DLS and $\tau_\infty = 138$ days in MC (Figure 6D). Similarly, fitting an exponential model with a baseline to the average similarity across neurons suggested a decay to an asymptotic correlation of $\gamma = 0.39$ for DLS and $\gamma = 0.29$ for MC. These results show that the neural activity patterns associated with this innate behavior are also stable over time, similar to our observations for learned motor skills.

Based on our analyses of the lever-pressing task, we wondered whether some of the residual neural drift could be accounted for by changes in the kinematics associated with the WDS. We therefore investigated whether the motor output during the WDS events exhibited a systematic drift over time and found this to be the case for all animals (Extended Data Fig. 8). To query whether this behavioral drift could be linked to the drift in neural activity, we computed the mean correlation between the neural and behavioral drifts and found a weak but significant effect of behavioral drift on the neural drift in DLS (Figure 6E; $\bar{\rho} = 0.11$, p = 0.005; permutation test). We also found a weak correlation in MC, although this did not reach the threshold for 'statistical significance' ($\bar{\rho} = 0.04$, p = 0.24; permutation test). While these correlations are small, we found them to be consistent with the expectation from a synthetic model where neural drift is entirely driven by behavioral drift (Figure 6F). This suggests that the weak effect size could be due to the slow rate of drift in both behavior and neural activity, which leads to a small signal-to-noise ratio. Both here and in the case of the timed lever-pressing task, we thus interpret the observed drift in neural activity to be accounted for, in large part, by slowly changing behavior.

Discussion

We have investigated whether stereotyped motor behaviors are driven by stable single neuron dynamics (Figure 1) in two major nodes of the motor system that are involved in the acquisition of motor skills – MC and DLS. Using an RNN model, we first demonstrated the necessity of long-term single-unit recordings for answering this question (Figure 2). We then performed such recordings in rats trained to generate highly stereotyped task-specific movement patterns (Figure 3) (Dhawale et al., 2017; Kawai et al., 2015). We found that the task-aligned activity of neurons in both MC and DLS was remarkably consistent over time, as expected for a stable control network. Recording single units for long durations was important to reveal this stability and distinguish it from constrained fluctuations on shorter timescales (Figure 4). We did observe a slow drift at the population level, which was accompanied by a concomitant drift in behavioral output (Figure 5). This is similar to previous reports of motor drift in expert performers (Chaisanguanthum et al., 2014). Importantly, the drift in behavior was correlated with the recorded drift in neural activity, suggesting that the neural drift could be explained, in large part, by small but systematic behavioral changes. Finally, we showed that these observations extend to an innate behavior with trial-like structure (Figure 6), suggesting that stable sensorimotor circuits underlie stereotyped behavior, both learned and innate.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Impact of behavioral variability in studies of neural stability

We showed how behavioral changes not fully constrained by the task can lead to the appearance of drift in single-unit neural activity patterns. This is particularly relevant since movements not relevant to the task are known to be strongly represented in cortical activity (Musall et al., 2019). As a result, our reported neural stability in relation to both the learned and innate motor behaviors, and similar reports from other studies, should be seen as lower bounds on the neural stability associated with a hypothetical perfectly stable behavior. Additionally, this observation of correlated neural and behavioral drift highlights the importance of high-resolution behavioral measurements when investigating the stability of neural circuit dynamics, since most tasks studied in neuroscience do not fully constrain behavioral output (Zagha et al., 2022).

While the observed slow drift in neural and behavioral space in expert animals suggests that the changes in neural circuits occur in directions of neural state space that affect motor output, it remains to be seen whether this behavioral drift constitutes a learning process that optimizes a utility function such as energy expenditure (Srinivasan & Ruina, 2006) or magnitude of the control signal (Todorov & Jordan, 2002). Alternatively, it could reflect a random walk in a degenerate motor space that preserves task performance (Chaisanguanthum et al., 2014; Qin et al., 2021). Previous work has also suggested that motor variability could be explicitly modulated to balance exploration and exploitation as a function of past performance and task uncertainty (Dhawale et al., 2019). If the behavioral drift we observe experimentally reflects such deliberate motor exploration, we might expect neural drift to be biased towards behaviorally potent dimensions to drive the necessary behavioral variability (H. G. Wu et al., 2014). Conversely, if the behavioral drift is a consequence of inevitable drift at the level of neural circuits, neural drift might be unbiased or even preferentially target behavioral null dimensions to minimize the impact on task performance. Future studies will be needed to arbitrate between these possibilities.

Prior studies of neural stability

It is worth noting the contrast between our results and previous studies that found task-associated neural activity in sensory and motor circuits to drift over time (Carmena et al., 2005; Deitch et al., 2021; Liberti et al., 2016; Rokni et al., 2007; Schoonover et al., 2021). Some of these differences could reflect physiological differences between species, circuits, or cell function (Rule & O'Leary, 2022), with recent studies e.g. showing differential stability between hippocampal cells representing time and odor identity in an odor discrimination task (Taxidis et al., 2020). However, they could also reflect differences in methodology. For example, brain computer interfaces (Carmena et al., 2005) circumvent the natural readout mechanism of the brain, which could affect the stability of learned representations. Additionally, different statistical assessments of stability have previously been suggested to underlie discrepancies in the apparent neural stability underlying a primate reaching task (Stevenson et al., 2011). Similarly, we find that accounting for the bias arising from finite recording durations is necessary to reveal the stability of sensorimotor circuits, and that unaccounted behavioral variability can confound analyses of representational drift in neural circuits. Furthermore, electrophysiology and calcium imaging can provide contrasting views on stability as discussed elsewhere (Dhawale et al., 2017; Lütcke et al., 2013). For the behaviors we probed in this study, electrophysiological recordings were essential to resolve neural dynamics on timescales of tens to hundreds of milliseconds (Huang et al., 2021). Moreover, being able to record continuously over many weeks mitigates the need to stitch together separate recording sessions with potential movement of the recording electrodes and changes of spike waveforms between sessions (Deitch et al., 2021; Dhawale et al., 2017; Lütcke et al., 2013). We therefore expect that our understanding of neural stability will benefit further from recent impressive advances in recording technology (Chung et al., 2019; Hong & Lieber, 2019; Steinmetz et al., 2021), especially if such advances can eventually be

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

combined with methods for chronic recordings to track changes in the waveforms of individual neurons (Dhawale et al., 2017).

The finding of stable neural correlates of motor output by us and others (Chestek et al., 2007; Flint et al., 2016; Ganguly & Carmena, 2009; Katlowitz et al., 2018) can also be contrasted with recent work suggesting that neural activity patterns in posterior parietal cortex (PPC) change over a few days to weeks during a virtual navigation task with stable performance (Driscoll et al., 2017). This discrepancy could arise from differences in methodology, recording duration, or limited behavioral constraints as discussed above. It could also reflect the fact that higher cortical regions are more sensitive to internal or external latent processes that lead to the appearance of drift due to an unconstrained environment. However, an alternative explanation is that higher-order brain regions, such as PPC or prefrontal cortex, accommodate drifting representations to allow fast learning processes or context-dependent gating of stable downstream dynamics (Mante et al., 2013; Murray & Escola, 2020; Roxin & Fusi, 2013; Rule et al., 2020). This is consistent with theoretical work on stable decoding from drifting neural codes (Rule et al., 2020; Rule & O'Leary, 2022) as well as a recent hypothesis that piriform cortex implements a 'fast' learning process with drifting representations, which drives a 'slow' learning process of stable downstream representations (Schoonover et al., 2021). In the context of brain-computer interfaces where a stable mapping between measured activity and system output is desirable (Degenhart et al., 2020), these considerations suggest that decoding activity from motor cortex or even subcortical brain regions is preferable to higher-order cortical areas such as prefrontal cortex or PPC.

Maintaining stability in the face of dynamic network changes

Our findings of long-term stability in both MC and DLS raise questions of how this is achieved mechanistically and whether there are active processes maintaining stability of network dynamics. Manipulation studies in both motor and sensory circuits suggest that this might be the case. It has previously been shown that motor circuits can recover their activity and function after invasive circuit manipulations by returning to a homeostatic set-point even in the absence of further practice (Otchy et al., 2015). At the single-neuron level, there are also intrinsic mechanisms keeping the firing rates of neurons in a tight range. For example, an increase in the excitability of individual neurons has been observed following sensory deprivation in both barrel cortex (Margolis et al., 2012) and V1 (Hengen et al., 2013, 2016; Mrsic-Flogel et al., 2007), with V1 also recovering higher-order network statistics after the perturbation (Y. K. Wu et al., 2020). These observations suggest that the brain uses homeostatic mechanisms to overcome direct perturbations. Similar mechanisms may also help explain how memories persist over time in the presence of continual learning and adaptation at the network level (Golowasch et al., 1999; Marder & Goaillard, 2006). Of course, such invasive perturbations are large compared to the changes that occur during normal motor learning, which instead consist of gradual synaptic turnover and plasticity. However, it is plausible that many of the same mechanisms that help restabilize the network following such large-scale perturbations are also involved in maintaining network stability under normal conditions.

Taken together, our results resolve a long-standing question in neuroscience by showing that the single neuron dynamics associated with stereotyped behaviors, both learned and innate, are stable over long timescales. However, they raise another mechanistic question of how new behaviors are learned without interfering with existing dynamics – that is, how does the brain combine long-term single-unit stability with life-long flexibility and adaptability (Benna & Fusi, 2016; Duncker et al., 2020; Kao, Jensen, et al., 2021; Kaplanis et al., 2018; Kirkpatrick et al., 2017; Yang et al., 2009)? This is an essential yet unanswered question for neuroscience, and future work in this area will likely require more elaborate experimental protocols with interleaved learning of multiple tasks coupled with long-term neural recordings and high-

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

resolution behavioral tracking to elucidate the mechanistic underpinnings of the balance between network stability and flexibility.

Acknowledgements

We are grateful to Kiah Hardcastle, Cengiz Pehlevan, Ta-Chu Kao, Guillaume Hennequin, and Marine Schimel for their feedback on the manuscript. This work was supported by a Gates Cambridge scholarship and Nordea-fonden (KTJ); a Helen Hay Whitney postdoctoral fellowship, the Zuckerman STEM Leadership Program postdoctoral fellowship, and the Women in Science Weizmann Institute of Science Award (NKH); a Life Sciences Research Foundation and Charles A. Kings Foundation postdoctoral fellowship (AKD); an EMBO postdoctoral fellowship ALTF1561-2013 and an HFSP postdoctoral fellowship LT 000514/2014 (SBEW); and NIH grants R01-NS099323-01 and R01-NS105349 (BPÖ).

References

Bar-Gad, I., Morris, G., & Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, *71*(6), 439–473. https://doi.org/10.1016/j.pneurobio.2003.12.001

Bedard, P., & Pycock, C. J. (1977). 'Wet-Dog' shake behaviour in the rat: A possible quantitative model of central 5-hydroxytryptamine activity. *Neuropharmacology*, *16*(10), 663–670. https://doi.org/10.1016/0028-3908(77)90117-4

Benna, M. K., & Fusi, S. (2016). Computational principles of synaptic memory consolidation. *Nature Neuroscience*, *19*(12), 1697–1706. https://doi.org/10.1038/nn.4401

Carmena, J. M., Lebedev, M. A., Henriquez, C. S., & Nicolelis, M. A. L. (2005). Stable Ensemble Performance with Single-Neuron Variability during Reaching Movements in Primates. *Journal of Neuroscience*, *25*(46), 10712–10716.

Chaisanguanthum, K. S., Shen, H. H., & Sabes, P. N. (2014). Motor Variability Arises from a Slow Random Walk in Neural State. *Journal of Neuroscience*, *34*(36), 12071–12080.

Chestek, C. A., Batista, A. P., Santhanam, G., Yu, B. M., Afshar, A., Cunningham, J. P., Gilja, V., Ryu, S. I., Churchland, M. M., & Shenoy, K. V. (2007). Single-Neuron Stability during Repeated Reaching in Macaque Premotor Cortex. *Journal of Neuroscience*, *27*(40), 10742–10750. https://doi.org/10.1523/JNEUROSCI.0959-07.2007

Chung, J. E., Joo, H. R., Fan, J. L., Liu, D. F., Barnett, A. H., Chen, S., Geaghan-Breiner, C., Karlsson, M. P., Karlsson, M., Lee, K. Y., Liang, H., Magland, J. F., Pebbles, J. A., Tooker, A. C., Greengard, L. F., Tolosa, V. M., & Frank, L. M. (2019). High-Density, Long-Lasting, and Multi-region Electrophysiological Recordings Using Polymer Electrode Arrays. *Neuron*, *101*(1), 21-31.e5. https://doi.org/10.1016/j.neuron.2018.11.002

Churchland, M. M. (2015). Using the precision of the primate to study the origins of movement variability. *Neuroscience*, *296*, 92–100. https://doi.org/10.1016/j.neuroscience.2015.01.005

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56. https://doi.org/10.1038/nature11129

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Clopath, C., Bonhoeffer, T., Hübener, M., & Rose, T. (2017). Variance and invariance of neuronal long-term representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1715), 20160161. https://doi.org/10.1098/rstb.2016.0161

Degenhart, A. D., Bishop, W. E., Oby, E. R., Tyler-Kabara, E. C., Chase, S. M., Batista, A. P., & Yu, B. M. (2020). Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature Biomedical Engineering*, *4*(7), 672–685. https://doi.org/10.1038/s41551-020-0542-9

Deitch, D., Rubin, A., & Ziv, Y. (2021). Representational drift in the mouse visual cortex. *Current Biology*, *31*(19), 4327-4339.e6. https://doi.org/10.1016/j.cub.2021.07.062

Dhawale, A. K., Miyamoto, Y. R., Smith, M. A., & Ölveczky, B. P. (2019). Adaptive Regulation of Motor Variability. *Current Biology*, *29*(21), 3551-3562.e7. https://doi.org/10.1016/j.cub.2019.08.052

Dhawale, A. K., Poddar, R., Wolff, S. B., Normand, V. A., Kopelowitz, E., & Ölveczky, B. P. (2017). Automated long-term recording and analysis of neural activity in behaving animals. *ELife*, *6*, e27702. https://doi.org/10.7554/eLife.27702

Dhawale, A. K., Wolff, S. B. E., Ko, R., & Ölveczky, B. P. (2021). The basal ganglia control the detailed kinematics of learned motor skills. *Nature Neuroscience*, *24*(9), 1256–1269. https://doi.org/10.1038/s41593-021-00889-3

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., & Harvey, C. D. (2017). Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell*, *170*(5), 986-999.e16. https://doi.org/10.1016/j.cell.2017.07.021

Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M., & Sussillo, D. (2020). Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in Neural Information Processing Systems*, *33*. https://proceedings.neurips.cc/paper/2020/hash/a576eafbce762079f7d1f77fca1c5cc2-Abstract.html

Fletcher, A., & Harding, V. (1981). An examination of the 'wet dog' shake behaviour in rats produced by acute administration of sodium n-dipropylacetate. *Journal of Pharmacy and Pharmacology*, *33*(1), 811–813. https://doi.org/10.1111/j.2042-7158.1981.tb13945.x

Flint, R. D., Scheid, M. R., Wright, Z. A., Solla, S. A., & Slutzky, M. W. (2016). Long-Term Stability of Motor Cortical Activity: Implications for Brain Machine Interfaces and Optimal Feedback Control. *Journal of Neuroscience*, *36*(12), 3623–3632. https://doi.org/10.1523/JNEUROSCI.2339-15.2016

Fraser, G. W., & Schwartz, A. B. (2012). Recording from the same neurons chronically in motor cortex. *Journal of Neurophysiology*, *107*(7), 1970–1978. https://doi.org/10.1152/jn.01012.2010

Fu, M., Yu, X., Lu, J., & Zuo, Y. (2012). Repetitive motor learning induces coordinated formation of clustered dendritic spines in vivo. *Nature*, *483*(7387), 92–95. https://doi.org/10.1038/nature10844

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., & Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, *23*(2), 260–270. https://doi.org/10.1038/s41593-019-0555-4

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978–984. https://doi.org/10.1016/j.neuron.2017.05.025

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ganguly, K., & Carmena, J. M. (2009). Emergence of a Stable Cortical Map for Neuroprosthetic Control. *PLOS Biology*, *7*(7), e1000153. https://doi.org/10.1371/journal.pbio.1000153

Golowasch, J., Casey, M., Abbott, L. F., & Marder, E. (1999). Network Stability from Activity-Dependent Regulation of Neuronal Conductances. *Neural Computation*, *11*(5), 1079–1096. https://doi.org/10.1162/089976699300016359

Haith, A. M., & Krakauer, J. W. (2013). Model-Based and Model-Free Mechanisms of Human Motor Learning. In M. J. Richardson, M. A. Riley, & K. Shockley (Eds.), *Progress in Motor Control* (pp. 1–21). Springer. https://doi.org/10.1007/978-1-4614-5465-6_1

Hatsopoulos, N. G., & Suminski, A. J. (2011). Sensing with the Motor Cortex. *Neuron*, *72*(3), 477–487. https://doi.org/10.1016/j.neuron.2011.10.020

Hengen, K. B., Lambo, M. E., Van Hooser, S. D., Katz, D. B., & Turrigiano, G. G. (2013). Firing Rate Homeostasis in Visual Cortex of Freely Behaving Rodents. *Neuron*, *80*(2), 335–342. https://doi.org/10.1016/j.neuron.2013.08.038

Hengen, K. B., Torrado Pacheco, A., McGregor, J. N., Van Hooser, S. D., & Turrigiano, G. G. (2016). Neuronal Firing Rate Homeostasis Is Inhibited by Sleep and Promoted by Wake. *Cell*, *165*(1), 180–191. https://doi.org/10.1016/j.cell.2016.01.046

Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M., & Miller, K. D. (2018). The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron*, *98*(4), 846-860.e5. https://doi.org/10.1016/j.neuron.2018.04.017

Hennequin, G., Vogels, T. P., & Gerstner, W. (2014). Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. *Neuron*, *82*(6), 1394–1406. https://doi.org/10.1016/j.neuron.2014.04.045

Holtmaat, A., & Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, *10*(9), 647–658. https://doi.org/10.1038/nrn2699

Hong, G., & Lieber, C. M. (2019). Novel electrode technologies for neural recordings. *Nature Reviews Neuroscience*, *20*(6), 330–345. https://doi.org/10.1038/s41583-019-0140-6

Huang, L., Ledochowitsch, P., Knoblich, U., Lecoq, J., Murphy, G. J., Reid, C., de Vries, S. E. J., Koch, C., Zeng, H., Buice, M. A., Waters, J., & Li, L. (2021). Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *ELife*, *10*, e51675. https://doi.org/10.7554/eLife.51675

Hunnicutt, B. J., Jongbloets, B. C., Birdsong, W. T., Gertz, K. J., Zhong, H., & Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization. *ELife*, *5*, e19103. https://doi.org/10.7554/eLife.19103

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. *ArXiv:1605.03170 [Cs]*. http://arxiv.org/abs/1605.03170

Jensen, K. T., Kao, T.-C., Stone, J. T., & Hennequin, G. (2021). Scalable Bayesian GPFA with automatic relevance determination and discrete noise models. *BioRxiv*, 2021.06.03.446788. https://doi.org/10.1101/2021.06.03.446788

Kanwal, J. K., Coddington, E., Frazer, R., Limbania, D., Turner, G., Davila, K. J., Givens, M. A., Williams, V., Datta, S. R., & Wasserman, S. (2021). Internal State: Dynamic, Interconnected Communication

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Loops Distributed Across Body, Brain, and Time. *Integrative and Comparative Biology*, *61*(3), 867–886. https://doi.org/10.1093/icb/icab101

Kao, T.-C., Jensen, K. T., Bernacchia, A., & Hennequin, G. (2021). Natural continual learning: Success is a journey, not (just) a destination. *ArXiv:2106.08085 [Cs, q-Bio]*. http://arxiv.org/abs/2106.08085

Kao, T.-C., Sadabadi, M. S., & Hennequin, G. (2021). Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. *Neuron*, *109*(9), 1567-1581.e12. https://doi.org/10.1016/j.neuron.2021.03.009

Kaplanis, C., Shanahan, M., & Clopath, C. (2018). Continual Reinforcement Learning with Complex Synapses. *ArXiv:1802.07239 [Cs]*. http://arxiv.org/abs/1802.07239

Kargo, W. J., & Nitz, D. A. (2004). Improvements in the Signal-to-Noise Ratio of Motor Cortex Cells Distinguish Early versus Late Phases of Motor Skill Learning. *Journal of Neuroscience*, *24*(24), 5560–5569. https://doi.org/10.1523/JNEUROSCI.0562-04.2004

Katlowitz, K. A., Picardo, M. A., & Long, M. A. (2018). Stable Sequential Activity Underlying the Maintenance of a Precisely Executed Skilled Behavior. *Neuron*, *98*(6), 1133-1140.e3. https://doi.org/10.1016/j.neuron.2018.05.017

Kawai, R., Markman, T., Poddar, R., Ko, R., Fantana, A. L., Dhawale, A. K., Kampff, A. R., & Ölveczky, B. P. (2015). Motor Cortex Is Required for Learning but Not for Executing a Motor Skill. *Neuron*, *86*(3), 800–812. https://doi.org/10.1016/j.neuron.2015.03.024

Keshtkaran, M. R., Sedler, A. R., Chowdhury, R. H., Tandon, R., Basrai, D., Nguyen, S. L., Sohn, H., Jazayeri, M., Miller, L. E., & Pandarinath, C. (2021). A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *BioRxiv*, 2021.01.13.426570. https://doi.org/10.1101/2021.01.13.426570

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*. http://arxiv.org/abs/1412.6980

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kleinrok, Z., & Turski, L. (1980). Kainic acid-induced wet dog shakes in rats. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *314*(1), 37–46. https://doi.org/10.1007/BF00498429

Krakauer, J. W., & Shadmehr, R. (2006). Consolidation of motor memory. *Trends in Neurosciences*, *29*(1), 58–64. https://doi.org/10.1016/j.tins.2005.10.003

Kubota, Y., Liu, J., Hu, D., DeCoteau, W. E., Eden, U. T., Smith, A. C., & Graybiel, A. M. (2009). Stable Encoding of Task Structure Coexists With Flexible Coding of Task Events in Sensorimotor Striatum. *Journal of Neurophysiology*, *102*(4), 2142–2160. https://doi.org/10.1152/jn.00522.2009

Laje, R., & Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, *16*(7), 925–933. https://doi.org/10.1038/nn.3405

Liberti, W. A., Markowitz, J. E., Perkins, L. N., Liberti, D. C., Leman, D. P., Guitchounts, G., Velho, T., Kotton, D. N., Lois, C., & Gardner, T. J. (2016). Unstable neurons underlie a stable learned behavior. *Nature Neuroscience*, *19*(12), 1665–1671. https://doi.org/10.1038/nn.4405

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Lindsay, G. W. (2022). Testing the Tools of Systems Neuroscience on Artificial Neural Networks. *ArXiv:2202.07035 [Cs, q-Bio]*. http://arxiv.org/abs/2202.07035

Lütcke, H., Margolis, D. J., & Helmchen, F. (2013). Steady or changing? Long-term monitoring of neuronal population activity. *Trends in Neurosciences*, *36*(7), 375–384. https://doi.org/10.1016/j.tins.2013.03.008

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Marder, E., & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, *7*(7), 563–574. https://doi.org/10.1038/nrn1949

Margolis, D. J., Lütcke, H., Schulz, K., Haiss, F., Weber, B., Kügler, S., Hasan, M. T., & Helmchen, F. (2012). Reorganization of cortical population activity imaged throughout long-term sensory deprivation. *Nature Neuroscience*, *15*(11), 1539–1546. https://doi.org/10.1038/nn.3240

Marinescu, I. E., Lawlor, P. N., & Kording, K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour*, *2*(12), 891–898. https://doi.org/10.1038/s41562-018-0466-5

Marshall, J. D., Aldarondo, D. E., Dunn, T. W., Wang, W. L., Berman, G. J., & Ölveczky, B. P. (2021). Continuous Whole-Body 3D Kinematic Recordings across the Rodent Behavioral Repertoire. *Neuron*, *109*(3), 420-437.e8. https://doi.org/10.1016/j.neuron.2020.11.016

Martin, W. R., Wikler, A., Eades, C. G., & Pescor, F. T. (1963). Tolerance to and physical dependence on morphine in rats. *Psychopharmacologia*, *4*(4), 247–260. https://doi.org/10.1007/BF00408180

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

Mehler, D. M. A., & Kording, K. P. (2020). The lure of misleading causal statements in functional connectivity research. *ArXiv:1812.03363 [q-Bio]*. http://arxiv.org/abs/1812.03363

Melnick, M. J. (1971). Effects of Overlearning on the Retention of a Gross Motor Skill. *Research Quarterly. American Association for Health, Physical Education and Recreation*, *42*(1), 60–69. https://doi.org/10.1080/10671188.1971.10615036

Miller, E. M., Shankar, M. U., Knutson, B., & McClure, S. M. (2014). Dissociating Motivation from Reward in Human Striatal Activity. *Journal of Cognitive Neuroscience*, *26*(5), 1075–1084. https://doi.org/10.1162/jocn_a_00535

Mrsic-Flogel, T. D., Hofer, S. B., Ohki, K., Reid, R. C., Bonhoeffer, T., & Hübener, M. (2007). Homeostatic Regulation of Eye-Specific Responses in Visual Cortex during Ocular Dominance Plasticity. *Neuron*, *54*(6), 961–972. https://doi.org/10.1016/j.neuron.2007.05.028

Murray, J. M., & Escola, G. S. (2020). Remembrance of things practiced with fast and slow learning in cortical and subcortical pathways. *Nature Communications*, *11*(1), 6441. https://doi.org/10.1038/s41467-020-19788-5

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686. https://doi.org/10.1038/s41593-019-0502-4

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Oorschot, D. E. (1996). Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: A stereological study using the cavalieri and optical disector methods. *Journal of Comparative Neurology*, *366*(4), 580–599. https://doi.org/10.1002/(SICI)1096-9861(19960318)366:4<580::AID-CNE3>3.0.CO;2-0

Otchy, T. M., Wolff, S. B. E., Rhee, J. Y., Pehlevan, C., Kawai, R., Kempf, A., Gobes, S. M. H., & Ölveczky, B. P. (2015). Acute off-target effects of neural circuit manipulations. *Nature*, *528*(7582), 358–363. https://doi.org/10.1038/nature16442

Park, S.-W., Dijkstra, T., & Sternad, D. (2013). Learning to never forget—Time scales and specificity of long-term memory of a motor skill. *Frontiers in Computational Neuroscience*, *7*. https://doi.org/10.3389/fncom.2013.00111

Park, S.-W., & Sternad, D. (2015). Robust retention of individual sensorimotor skill after self-guided practice. *Journal of Neurophysiology*, *113*(7), 2635–2645. https://doi.org/10.1152/jn.00884.2014

Peters, A. J., Chen, S. X., & Komiyama, T. (2014). Emergence of reproducible spatiotemporal activity during motor learning. *Nature*, *510*(7504), 263–267. https://doi.org/10.1038/nature13235

Peters, A. J., Lee, J., Hedrick, N. G., O'Neil, K., & Komiyama, T. (2017). Reorganization of corticospinal output during motor learning. *Nature Neuroscience*, *20*(8), 1133–1141. https://doi.org/10.1038/nn.4596

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999. https://doi.org/10.1038/nature07140

Poddar, R., Kawai, R., & Ölveczky, B. P. (2013). A Fully Automated High-Throughput Training System for Rodents. *PLOS ONE*, *8*(12), e83171. https://doi.org/10.1371/journal.pone.0083171

Qin, S., Farashahi, S., Lipshutz, D., Sengupta, A. M., Chklovskii, D. B., & Pehlevan, C. (2021). Coordinated drift of receptive fields during noisy representation learning. *BioRxiv*, 2021.08.30.458264. https://doi.org/10.1101/2021.08.30.458264

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496. https://doi.org/10.1126/science.1242072

Rokni, U., Richardson, A. G., Bizzi, E., & Seung, H. S. (2007). Motor learning with unstable neural representations. *Neuron*, *54*(4), 653–666. https://doi.org/10.1016/j.neuron.2007.04.030

Roxin, A., & Fusi, S. (2013). Efficient Partitioning of Memory Systems and Its Importance for Memory Consolidation. *PLOS Computational Biology*, *9*(7), e1003146. https://doi.org/10.1371/journal.pcbi.1003146

Rule, M. E., Loback, A. R., Raman, D. V., Driscoll, L. N., Harvey, C. D., & O'Leary, T. (2020). Stable task information from an unstable neural population. *ELife*, *9*, e51121. https://doi.org/10.7554/eLife.51121

Rule, M. E., & O'Leary, T. (2022). Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences*, *119*(7). https://doi.org/10.1073/pnas.2106692119

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Sadeh, S., & Clopath, C. (2022). *Contribution of behavioural variability to representational drift* (p. 2022.01.02.474731). https://doi.org/10.1101/2022.01.02.474731

Schoonover, C. E., Ohashi, S. N., Axel, R., & Fink, A. J. P. (2021). Representational drift in primary olfactory cortex. *Nature*, *594*, 541–546. https://doi.org/10.1038/s41586-021-03628-7

Sheng, M., Lu, D., Shen, Z., & Poo, M. (2019). Emergence of stable striatal D1R and D2R neuronal ensembles with distinct firing sequence during motor learning. *Proceedings of the National Academy of Sciences*, *116*(22), 11038–11047. https://doi.org/10.1073/pnas.1901712116

Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013). Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience*, *36*(1), 337–359. https://doi.org/10.1146/annurev-neuro-062111-150509

Sheppard, P. A. S., Choleris, E., & Galea, L. A. M. (2019). Structural plasticity of the hippocampus in response to estrogens in female rodents. *Molecular Brain*, *12*(1), 22. https://doi.org/10.1186/s13041-019-0442-7

Sizemore, M., & Perkel, D. J. (2011). Premotor synaptic plasticity limited to the critical period for song learning. *Proceedings of the National Academy of Sciences*, *108*(42), 17492–17497. https://doi.org/10.1073/pnas.1104255108

Srinivasan, M., & Ruina, A. (2006). Computer optimization of a minimal biped model discovers walking and running. *Nature*, *439*(7072), 72–75. https://doi.org/10.1038/nature04113

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., … Harris, T. D. (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

Stevenson, I. H. (2018). Omitted Variable Bias in GLMs of Neural Spiking Activity. *Neural Computation*, *30*(12), 3227–3258. https://doi.org/10.1162/neco_a_01138

Stevenson, I. H., Cherian, A., London, B. M., Sachs, N. A., Lindberg, E., Reimer, J., Slutzky, M. W., Hatsopoulos, N. G., Miller, L. E., & Kording, K. P. (2011). Statistical assessment of the stability of neural movement representations. *Journal of Neurophysiology*, *106*(2), 764–774. https://doi.org/10.1152/jn.00626.2010

Sussillo, D., & Abbott, L. F. (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron*, *63*(4), 544–557. https://doi.org/10.1016/j.neuron.2009.07.018

Taxidis, J., Pnevmatikakis, E. A., Dorian, C. C., Mylavarapu, A. L., Arora, J. S., Samadian, K. D., Hoffberg, E. A., & Golshani, P. (2020). Differential Emergence and Stability of Sensory and Temporal Representations in Context-Specific Hippocampal Sequences. *Neuron*, *108*(5), 984-998.e9. https://doi.org/10.1016/j.neuron.2020.08.028

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235. https://doi.org/10.1038/nn963

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, *43*(1), 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Willett, J. A., Cao, J., Johnson, A., Patel, O. H., Dorris, D. M., & Meitzen, J. (2020). The estrous cycle modulates rat caudate–putamen medium spiny neuron physiology. *European Journal of Neuroscience*, *52*(1), 2737–2755. https://doi.org/10.1111/ejn.14506

Williams, A. H., Poole, B., Maheswaranathan, N., Dhawale, A. K., Fisher, T., Wilson, C. D., Brann, D. H., Trautmann, E. M., Ryu, S., Shusterman, R., Rinberg, D., Ölveczky, B. P., Shenoy, K. V., & Ganguli, S. (2020). Discovering Precise Temporal Patterns in Large-Scale Neural Recordings through Robust and Interpretable Time Warping. *Neuron*, *105*(2), 246-259.e8. https://doi.org/10.1016/j.neuron.2019.10.020

Wolff, S. B. E., Ko, R., & Ölveczky, B. P. (2019). Distinct roles for motor cortical and thalamic inputs to striatum during motor learning and execution. *BioRxiv*, 825810. https://doi.org/10.1101/825810

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11), 1212–1217. https://doi.org/10.1038/81497

Wu, H. G., Miyamoto, Y. R., Castro, L. N. G., Ölveczky, B. P., & Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nature Neuroscience*, *17*(2), 312–321. https://doi.org/10.1038/nn.3616

Wu, Y. K., Hengen, K. B., Turrigiano, G. G., & Gjorgjieva, J. (2020). Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. *Proceedings of the National Academy of Sciences*, *117*(39), 24514–24525. https://doi.org/10.1073/pnas.1918368117

Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., Jones, T., & Zuo, Y. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature*, *462*(7275), 915–919. https://doi.org/10.1038/nature08389

Yang, G., Pan, F., & Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. *Nature*, *462*(7275), 920–924. https://doi.org/10.1038/nature08577

Zagha, E., Erlich, J. C., Lee, S., Lur, G., O'Connor, D. H., Steinmetz, N. A., Stringer, C., & Yang, H. (2022). The Importance of Accounting for Movement When Relating Neuronal Activity to Sensory and Cognitive Processes. *Journal of Neuroscience*, *42*(8), 1375–1382. https://doi.org/10.1523/JNEUROSCI.1919-21.2021

Zheng, T., & Wilson, C. J. (2002). Corticostriatal Combinatorics: The Implications of Corticostriatal Axonal Arborizations. *Journal of Neurophysiology*, *87*(2), 1007–1017. https://doi.org/10.1152/jn.00519.2001

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Data Analysis

Animal training and data acquisition

Female Long Evans rats (n = 6) were trained in an automated home-cage system on a lever-pressing task as described previously (Dhawale et al., 2017; Kawai et al., 2015; Poddar et al., 2013). In short, animals were rewarded for pressing a lever twice with an inter-press interval of 700 ms. Electrophysiological data was recorded from layer 5 of the motor cortex (MC; n = 3) and from the dorsolateral striatum (DLS; n = 3) and spike-sorted as described by Dhawale et al. (Dhawale et al., 2017). Data from all animals have previously been used in (Dhawale et al., 2021). The care and experimental manipulation of all animals were reviewed and approved by the Harvard Institutional Animal Care and Use Committee.

Behavioral tracking

Videos were recorded at 120 Hz during the lever-pressing task from two cameras positioned at the left and right side of the home cage relative to the lever. Automated behavioral tracking was carried out using DeeperCut (Insafutdinov et al., 2016; Mathis et al., 2018). For each of 500 frames from each camera, the corresponding forelimb of the animal was manually labelled. This was used as a training dataset for DeeperCut to generate full trajectories for all trials followed by interpolation with a cubic spline. Clustering of task trials based on behavioral readouts was carried out using forelimb positions as tracked by DeeperCut as well as accelerometer data from an accelerometer attached to the skull of each animal. These features were embedded in a t-SNE space and clustered using density-based clustering (Rodriguez & Laio, 2014). Only trials falling in the largest cluster for each animal (range of 37% to 93% of trials across animals) and with inter-press intervals (IPIs) between 600 ms and 800 ms were included in the analyses to minimize behavioral variability.

Detection and classification of wet dog shakes

To identify wet dog shakes (WDS), accelerometer data was first passed through a 12-20 Hz filter and the magnitude of the response calculated as $m = \sqrt{x^2 + y^2 + z^2}$. A moving average of m was calculated with a window size of 1/6 seconds, and WDS events identified as periods with $m > 0.03$. Peaks were found in a window of 800 ms centered at the middle of each WDS event and identified as local maxima or minima with a prominence of at least 0.07 times the difference between the highest maximum and lowest minimum in each channel. WDS events were aligned to the first positive peak in the vertical (z) channel and time-warped according to the inter-peak separation in this channel. Aligning to either horizontal channel gave similar results, and the vertical channel was preferred to avoid the degeneracy of the horizontal plane.

Time-warping

For all analyses of experimental data, we time-warped neural activity and behavior using piecewise linear warping (Williams et al., 2020) with parameters that aligned the two lever presses across all trials. We did this since neurons in DLS and MC have previously been shown to have activity patterns linked to these events (Dhawale et al., 2017). Time-warping of spike data in the lever-pressing task was carried out by linearly scaling all spike times between the first and second presses by a factor $\rho = \frac{700\ ms}{t_{trial}}$, where t_{trial} is the inter-press interval (IPI) in a given trial. All spike times after the second press were shifted by $700\ ms - t_{trial}$. Warping of behavioral data was carried out by fitting a cubic spline to the trajectories

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and extracting time points at a frequency of 120 Hz prior to the first press, $\rho \times 120$ Hz between the two presses, and 120 Hz after the second press. The warping coefficient ρ had a mean of 1.00 and a standard deviation of 0.07 across all trials and animals.

Warping of spike data for the wet dog shakes was carried out by linearly scaling all spike times between a quarter period before the first peak (t_1) and a quarter period after the last peak (t_2) by a factor $\rho = \frac{t_{med}}{t_{trial}}$, where t_{trial} is the period of the oscillation in a given trial and t_{med} is the median period across all trials and sessions for a given animal. All spike times before t_1 were shifted by $t_1 \times (\rho - 1)$ and all spike times after t_2 were shifted by $t_2 \times (\rho - 1)$. Warping of behavioral data was carried out by fitting a cubic spline to the accelerometer data and extracting time points at a frequency of 300 Hz prior to t_1, $\rho \times 300$ Hz between t_1 and t_2, and 300 Hz after t_2. The first detected positive peak was assigned a time of zero for each WDS. The warping coefficient ρ had a mean of 1.01 and a standard deviation of 0.07 across all trials and animals.

Data between 0.1 seconds before the first tap and 0.1 after the second tap was used for all analyses of the lever-pressing task, and data between 0.2 seconds before and 0.5 seconds after the first accelerometer peak was used for all WDS analyses.

Similarity of neural activity

PETHs were calculated for each session by summing the spikes across all trials for each time-within-trial. We convolved the resulting spike counts with a 15 ms Gaussian filter for the lever-pressing task, and with a 10 ms Gaussian filter for the WDS behavior to capture the shorter timescales associated with the 16 Hz oscillations. Pairwise PETH similarities between sessions were calculated as the Pearson correlation between **u** and **v**, where **u** and **v** are vectors containing the PETHs at 20 ms resolution. PETHs were normalized by z-scoring for visualization in Figure 4A for each unit, and by total spike count on each day for the PETHs in Figure 4B and 6B. Neural similarity as a function of time was calculated by computing the pairwise similarity of the PETHs for each unit across every pair of days in which the PETH contained at least 10 spikes. The pairwise similarities for each time difference were averaged across units in Figures 4C and 6C, after first averaging over all PETH pairs separated by the same time difference for each individual unit.

We restricted all analyses to neurons that were ‘task-modulated’. To define task-modulation, we computed a PETH for odd and even trials separately for each recording day and considered the correlation between this pair of PETHs on each day. We then averaged the result across days for each neuron. A neuron was considered task-modulated if this measure of same-day similarity exceeded $\rho_0 = 0.15$. This resulted in 221 of 363 neuron being task-modulated in DLS during lever-pressing task, 446 of 795 in MC during the lever-pressing task, 344 of 1250 in DLS during the WDS behavior, and 372 of 904 in MC during the WDS behavior.

Control analyses for stability vs time

In Figures 4C and 6C, we include a positive and a negative control for the neural similarity as a function of time difference. Here we provide a description of how these were computed. For the negative control, we computed the similarity between non-identical pairs of neurons. This can be seen as the asymptotic similarity in the limit of complete neural turnover but with constant population statistics (i.e. each neuron corresponds to a randomly sampled neuron from the population). This similarity was averaged across 1000 pairs of randomly sampled neurons, with each pair being recorded in a single animal. For the positive control, we resampled the neural activity with replacement on all trials in which it was recorded from the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

total distribution of recorded trials across days. We then computed the similarity as a function of time difference as in the original data. The process of resampling and computing similarity was repeated 100 times, and the figures indicate mean and standard error across samples. This control thus corresponds to the hypothetical similarity in the case where all neurons have a fixed distribution over firing patterns (i.e. neural activity is stable), and where the global distribution of firing patterns is matched to the data.

Alignment of neural dynamics

Aligning neural dynamics using CCA requires simultaneous recording of a large number of neurons. Since our recordings were asynchronous, this was in general not the case (c.f. Figure 3C). For this analysis, we therefore focused on a small subset of the data where 16 neurons were simultaneously recorded for a week and fired at least 10 spikes on each day. This dataset corresponds to days 8-14 of the DLS animal indicated in Figure 3C. We first computed the 'single-neuron similarity' in this dataset by computing the average PETH correlation across all neurons for each pair of days. We then computed the mean and standard deviation of this measure across all pairs of days separated by the same time difference. This provided a measure of the similarity of neural dynamics in a constant coordinate system with the axes aligned to individual neurons. For comparison with this, we also computed the similarity of neural activity when aligning the neural dynamics using CCA for each pair of sessions. To do this, we followed the approach outlined by Gallego et al. (Gallego et al., 2020) to align the dynamics on day 2 to the dynamics on day 1 across all pairs of days. This alignment was carried out at the level of PETHs rather than single trials. For these analyses, we aligned the dynamics across all neurons and considered the average correlation across all the resulting dimensions (i.e. the similarity was the average of all CCs). This addresses the question of whether stability increases if we allow for linear transformations of the coordinate system in which we characterize neural dynamics.

Exponential model fits and stability indices

To assess the stability of neural activity over time, we examined the Pearson correlation ρ between the computed PETHs as a function of the time difference δt between PETHs. We then fitted an exponential model $\hat{\rho} = \beta e^{\alpha\,\delta t}$ to this data for each neuron recorded for at least 4 days. This was done to better quantify the putative drift in neural activity across neurons by learning a parameter α that encompasses the rate of drift for each neuron. Here, β is an intercept describing the expected similarity for two sets of trials recorded on the same day, and α determines the rate of change of neural similarity. For this fit, we constrained β to be between -1 and +1 by passing it through a tanh transfer function since Pearson correlations must fall in this interval. The parameters were optimized to minimize the squared error between the predicted ($\hat{\rho}$) and observed (ρ) PETH correlations. This was done numerically, and the optimization was initialized from a linear fit to the data ($\hat{\rho} \approx \frac{\alpha}{\beta} t + \beta$). We denote the learned parameter α with units of inverse time as a 'stability index'. This is related to the time constant of an exponential decay model via $\alpha = -\tau^{-1}$, with the fitting of α being numerically more stable as it avoids τ approaching infinite values for slow decays. All data points with a time difference of at least 1 day were used to fit the models. The mean error of the model fit was quantified for each neuron as $\frac{1}{N}\sum_{i=1}^{N}|\rho_i - \widehat{\rho_i}|$, where $|\cdot|$ indicates the absolute value, and the sum runs over all N data points. Significance of median stability indices being different from zero was calculated by shuffling the vector of time differences for each unit 2,000 times, each time computing the median of the stability indices across all units and counting the fraction of shuffles where the median stability index was smaller than the experimentally observed median.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For comparison with this single-timescale model, we also considered a model which decayed to a learnable baseline γ: $\hat{\rho} = \beta e^{\alpha\,\delta t} + \gamma$. We did this since the presence of constrained latent processes could lead to a decay in neural similarity to a non-zero asymptotic value at long time differences. Clearly, the single-timescale exponential decay arises as a special case of this model for $\gamma = 0$. However, it is also worth noting that a linear model, commonly used in the literature (Deitch et al., 2021; Dhawale et al., 2017; Gallego et al., 2020; Liberti et al., 2016; Rokni et al., 2007; Schoonover et al., 2021), arises as $\gamma \rightarrow -\infty$. Intuitively, this is the case since any finite region of $\hat{\rho}$ is in the initial linear regime of an exponential that decays to $-\infty$. This model with a baseline thus serves as a generalization of both the linear and exponential models. When fitted to the neuron-averaged data and evaluated using hold-one-out crossvalidation, this model performed comparably to or better than the simple exponential decay across the four datasets (recordings from DLS/MC across the two tasks). Additionally, using this same crossvalidated evaluation metric, the exponential model consistently outperformed a linear model, suggesting that this is a more appropriate single-timescale model.

Stability as a function of recording duration

To extrapolate our stability indices to long recording durations across the population, we fitted a model to the stability index α as a function of recording time T of the form $\hat{\alpha} = -a - b\exp(-c\,T)$. We fitted the model by minimizing the L1 error between the observations and model fit across neurons $\mathcal{L} = \sum_n |\alpha_n - \hat{\alpha}_n|$ and restricted all parameters $\{a, b, c\}$ to be positive. In this model, the asymptotic stability is given by $\tau_\infty = \lim_{T\rightarrow\infty} -\hat{\alpha}^{-1} = a^{-1}$. To construct confidence intervals for this analysis, we subsampled the data points for each neuron $\{\alpha_n, T_n\}$ with replacement and repeated the model fitting procedure. Results are reported as medians and interquartile ranges by considering the 25th, 50th and 75th percentile of the corresponding distribution over τ_∞. While the model itself was fitted to the raw data, we denoised the data for the visualization in Figure 4E by plotting the median stability index across neurons binned by recording duration. The bins were selected with partial overlap (each neuron occurred in two bins), and the x-value indicated for each data point in the figure is the average recording duration for each neuron in the corresponding bin.

To compute the stability as a function of subsampled recording duration in Figure 4F, we used successive maximum time differences from $\delta t_{max} = 3$ to $\delta t_{max} = 13$ days. We then considered the average similarity as a function of time difference in Figure 4C, using only data up to and including δt_{max}. We computed stability indices for these subsets of data as described above and plotted the stability as a function of δt_{max}.

Behavioral similarity

To compute behavioral similarity as a function of time difference, we first extracted instantaneous velocities of both forelimbs in the vertical and horizontal directions as the first derivative of the time-warped cubic spline fitted to position as a function of time. We computed the pairwise behavioral similarity between sessions as the correlation between the mean velocity profiles across all trials from the corresponding sessions, averaged across both forelimbs and the vertical/horizontal directions.

To compute the correlation between neural and behavioral drift rates, we considered the behavioral similarity on pairs of consecutive days together with the neural similarity across the corresponding days, computed as PETH correlations as described above. We then considered the distribution of neural and behavioral similarities across all pairs of consecutive days for each recorded unit and computed the correlation between these two quantities. Finally, we computed the mean of this correlation across the population of units recorded from either DLS or MC. As a control, we permuted the behavioral data across

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

days to break any correlations between the neural and behavioral drift rates and repeated the analysis. In Figure 5E, null distributions are provided across 5,000 such random permutations. For these analyses, we did not include the first day of recording for any unit since this data was used to fit the synthetic control data (see below). Furthermore, we only considered neurons with at least 4 pairs of consecutive recording days (after discarding the first day of recording), such that all correlations were computed on the basis of at least 4 data points.

Stability of population decoding

To investigate the stability of a population decoder, we considered the same week-long subset of data as for the alignment of neural dynamics. We first square root transformed the neural data and convolved it with a 40 ms Gaussian filter similar to previous work. We then trained a crossvalidated ridge regression model to predict the left and right forelimb trajectories from neural activity using data from each single day and tested the model on all other days. Finally, we computed the performance of this decoder as a function of time difference between testing and training. For all decoding, we offset behavior from neural activity by 100 ms to account for the fact that neural activity precedes kinematics, similar to previous work in primates (Jensen et al., 2021; Keshtkaran et al., 2021). To test whether the decoder exhibited a significant decrease in performance as a function of time difference, we performed a bootstrap analysis by resampling with replacement the similarity as a function of time difference (i.e. we resampled 'pairs of days') and computing the slope of a linear fit to the data.

For comparison with this decoding model, we also considered decoding performance in an aligned latent space. To do this, we again considered all pairs of days and matched the number of trials on each pair of days to facilitate alignment (i.e. we discarded the later trials on the day with most trials). We then used PCA to reduce the dimensionality of the data from 16 to 10 for each day and trained our crossvalidated ridge regression model to predict behavior from this latent neural activity on the training data. At test time, we aligned the PCs on the test day to the PCs on the training day and predicted behavior from these aligned PCs. This follows the procedure described in previous work (Gallego et al., 2020). Note that alignment was in this case done at the level of single trials rather than trial-averaged PETHs. Finally, we considered the decoding performance as a function of time difference for this aligned decoder.

GLM model fitting and analysis

To investigate the correlation between neural and behavioral drift rates in synthetic data, where neural drift is determined entirely by behavioral drift (Figure 5E), we first fitted a linear-nonlinear Poisson GLM to the first day of recording for each neuron. This model took the form $\boldsymbol{y}_{t=1} \sim Poisson(\exp[\boldsymbol{W}\boldsymbol{x}_{t=1}])$, where $\boldsymbol{y}_t$ are the observed spike counts on day t across time bins (here a concatenation of trials and bins within each trial), $\boldsymbol{x}_t$ is a set of input features, and $\boldsymbol{W}$ is a weight matrix that is learned by maximizing the log likelihood of the data. As input features, we used the velocity of both forelimbs in the x-y plane for the lever-pressing task, and the accelerometer readout in 3 dimensions for the WDS task. In both cases, we included a 200 ms window of kinematics surrounding each 20 ms bin of neural activity in the feature vector. After fitting the model to data from day 1, we proceeded to generate synthetic neural activity by drawing spikes from the model $\tilde{\boldsymbol{y}}_{t>1} \sim Poisson(\exp[\boldsymbol{W}\boldsymbol{x}_{t>1}])$ for all subsequent days using the recorded behavior $\boldsymbol{x}$. We then constructed PETHs for each unit and session, as described for the experimental data, and repeated the analysis correlating behavioral similarity with neural similarity on consecutive days for this synthetic dataset. We repeated the sampling and analysis process 5,000 times to generate a distribution of neural-behavioral correlations from this synthetic model and computed p values as the fraction of synthetic correlation values that were smaller than the experimentally observed value. When performing these analyses, we discarded the first day of recording in both the synthetic and experimental

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data since this was used to fit the GLM.

Recurrent network modelling

Network architecture and training

The RNNs used in Figure 2 consisted of 250 recurrently connected units and 5 readouts units, which were simulated for 250 evenly spaced timesteps to generate 5 target outputs drawn from a Gaussian process with a squared exponential kernel that had a timescale of $\tau = \frac{250}{6}$. The RNN dynamics were given by

$$\boldsymbol{x}_{t+1} = [\boldsymbol{x}_t + \tau^{-1}(-\boldsymbol{x}_t + \boldsymbol{W}_{rec}\boldsymbol{x}_t + \boldsymbol{\epsilon})]_+$$

$$\boldsymbol{\epsilon} \sim N(0, 0.2\boldsymbol{I})$$

$$\boldsymbol{y}_t = \boldsymbol{W}_{out}\ \boldsymbol{x} + \boldsymbol{b}$$

$\boldsymbol{W}_{rec}$, $\boldsymbol{W}_{out}$, $\boldsymbol{b}$, and $\boldsymbol{x}_0$ were optimized using gradient descent with Adam (Kingma and Ba 2014) to minimize the loss function

$$\mathcal{L} = \sum_{i,t} \left({y_{i,t}}^{output} - {y_{i,t}}^{target} \right)^2 + 10^{-4} \left(\sum_{ij} \left| W_{rec,ij} \right|^2 + \sum_{ij} \left| W_{out,ij} \right|^2 \right)$$

We used a learning rate of 0.0005 and batch size of 20 to train all networks.

Similarity measures

100 instances of each network were run to constitute a set of trials (a 'session'). Observation noise was added to all neural activities x by drawing spikes from a Poisson noise model $s \sim Poisson(\lambda x)$, where λ is a constant scaling factor for each session used to scale the mean activity to 6.25 Hz. PETHs were constructed by averaging the activity of each unit across all trials for a given network. PETH similarity was computed as the Pearson correlation between PETHs as for the experimental data. Behavioral similarity was computed as the mean RNN output correlation across pairs of trials for each pair of sessions. Latent similarity was computed by first convolving the single-trial activity with a 30 ms Gaussian filter. The activities of non-overlapping groups of 50 neurons were then concatenated into 50xT matrices for each session to simulate different simultaneously recorded populations of neurons. Here, T is the number of time bins per trial (250) times the number of trials per session (100). The 50xT matrices were reduced to 10xT matrices by PCA, and the resulting matrices were aligned by CCA across networks. The CCA similarity for a pair of networks and group of neurons was computed as the mean correlation of the top 4 CCs. This procedure was intended to mirror the analysis by Gallego et al. (Gallego et al., 2020).

Interpolating networks

To interpolate the networks in Figure 2E & 2F, two networks were first trained independently to produce the target output, generating two sets of parameters

$\theta_1 = \{\boldsymbol{W}_{rec}^1, \boldsymbol{W}_{out}^1, \boldsymbol{b}^1, \boldsymbol{x}_0^1\}$ and $\theta_2 = \{\boldsymbol{W}_{rec}^2, \boldsymbol{W}_{out}^2, \boldsymbol{b}^2, \boldsymbol{x}_0^2\}$

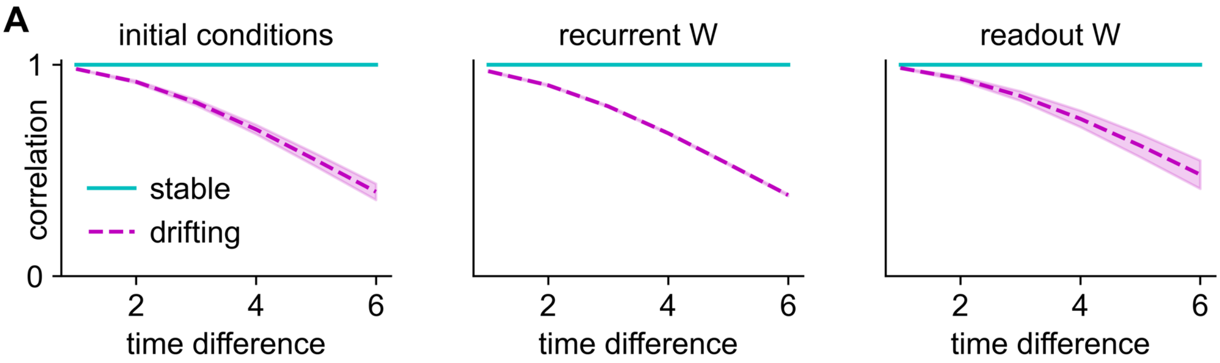
Seven new parameter sets θ^{dt} were generated by linear interpolation between $\theta_1$and $(0.3\theta_1 + 0.7\theta_2)$, or equivalently by considering seven networks spanning the first part of a linear interpolation between θ_1

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and θ_2. We chose not to consider the full interpolation series since neural activities became uncorrelated before the parameters were fully uncorrelated (Figure 2), and we were interested in the range of parameters where neural activity drifted. For each interpolated parameter set, $\boldsymbol{W}_{out}^{dt}$ was fixed and the network was retrained to optimize all other parameters. Note that this procedure is merely used to generate a phenomenological model of a motor circuit with drifting connectivity and stable output, and it should not be interpreted as a mechanistic model. For the control network, the same interpolation and re-optimization procedure was carried out, but in this case interpolating between $\theta_1$and θ_1 (i.e. itself), such that the only differences between networks were fluctuations around the original connectivity. The whole procedure of training two initial networks and interpolating was repeated 10 times, and results in Figure 2F are reported as the mean and standard deviation across these repetitions.

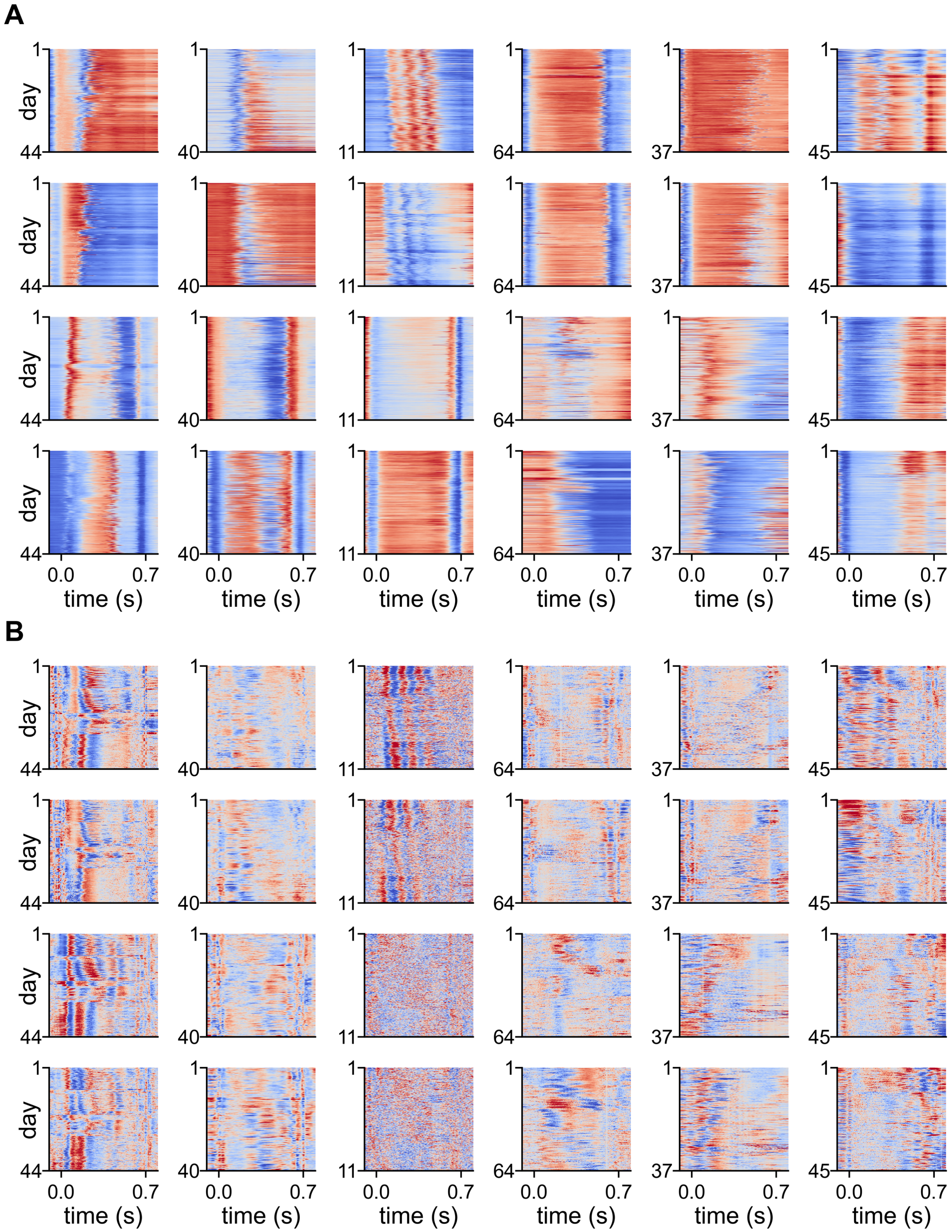
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extended Data figures



Extended Data Fig. 1: RNN parameter interpolation. (A) Mean correlation between the initial conditions (left), recurrent weight matrix (center), and readout weight matrix (right) of the simulated RNNs as a function of time difference for the stable and unstable networks. Shading indicates standard deviation across 10 repetitions of training and interpolation.

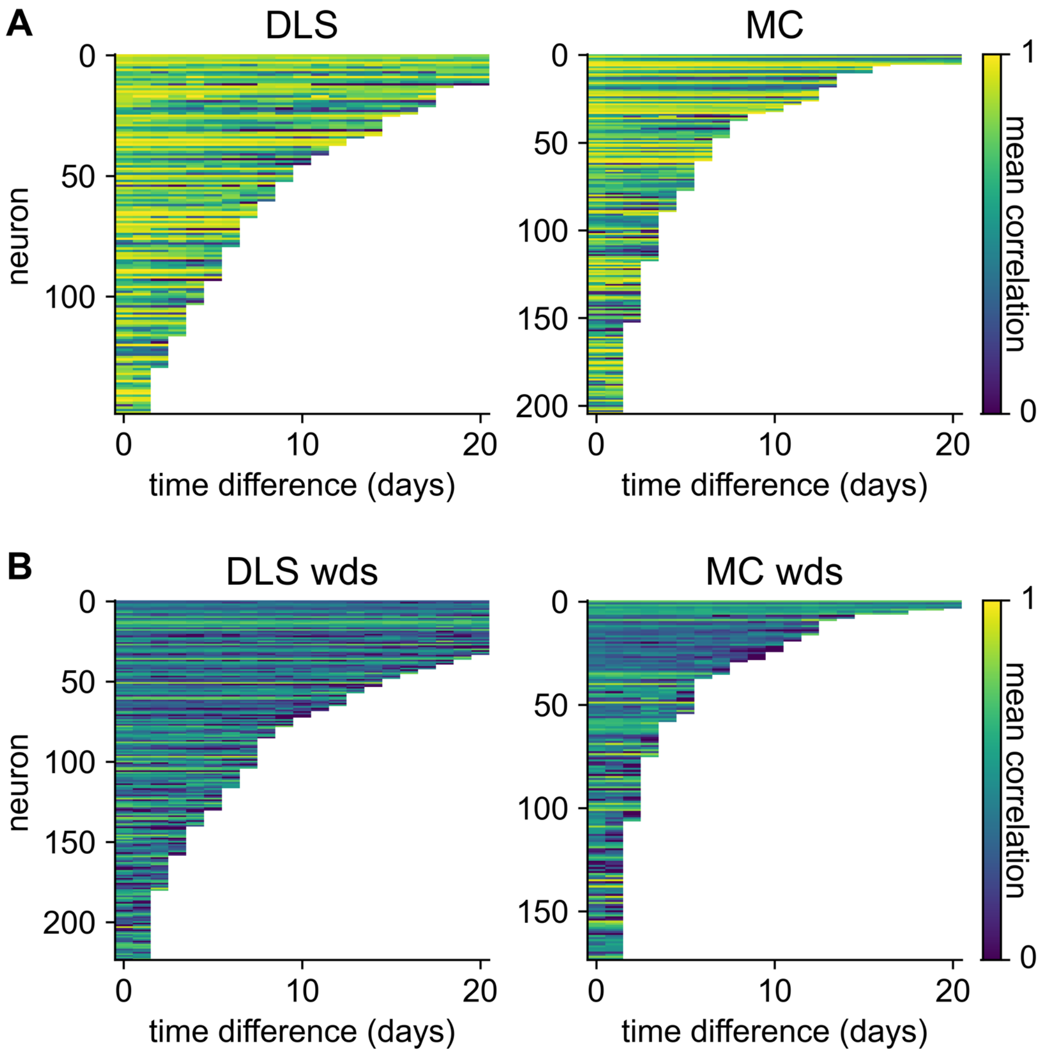
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Data Fig. 2: Kinematics of all animals. (A) Heatmaps showing the paw trajectories of each animal on every trial across all days. x axes indicate time within trial and y axes indicate trial number from first (top) to last (bottom). Each column corresponds to a single animal (first three: DLS, last three: MC). The rows illustrate the trajectories of

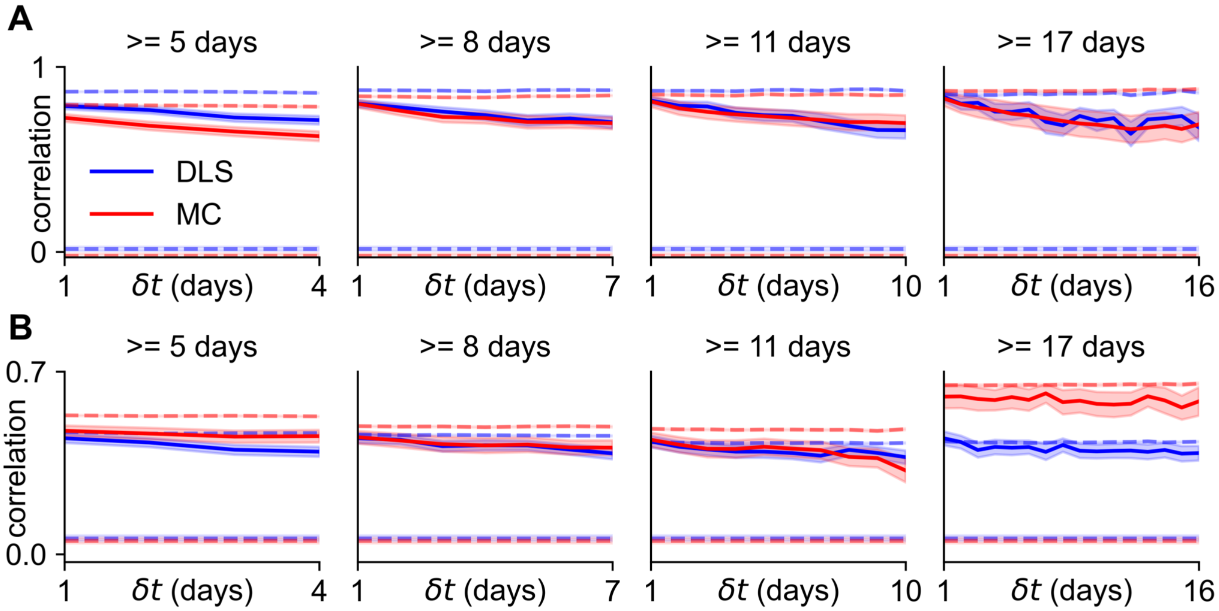
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the right paw parallel and perpendicular to the floor, followed by the left paw parallel and perpendicular to the floor. The second animal from the left corresponds to the example used in Figures 3A and 5A-B. **(B)** Heatmaps showing the z-scored velocity of each animal on every trial across all days for the animals in (A). The rows illustrate the velocity of the right paw parallel and perpendicular to the floor followed by the left paw parallel and perpendicular to the floor.

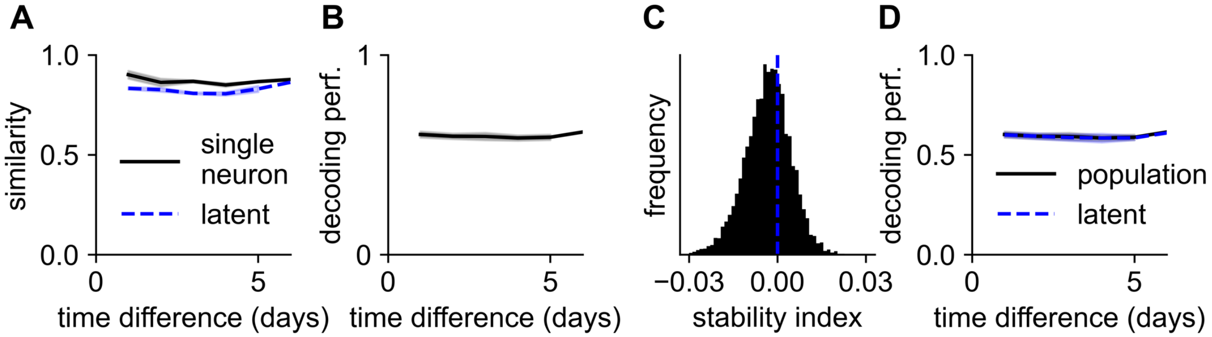


Extended Data Fig. 3: Similarity versus time difference for all neurons. We computed the PETH correlation as a function of time difference for all neurons, taking the average across all pairs of days separated by the same time difference for each neuron. This figure shows the average similarity as a function of time difference for neurons recorded in DLS (left) or MC (right) during the lever-pressing task **(A)** and the wet-dog shake behavior **(B).** Neurons were sorted by recording duration.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



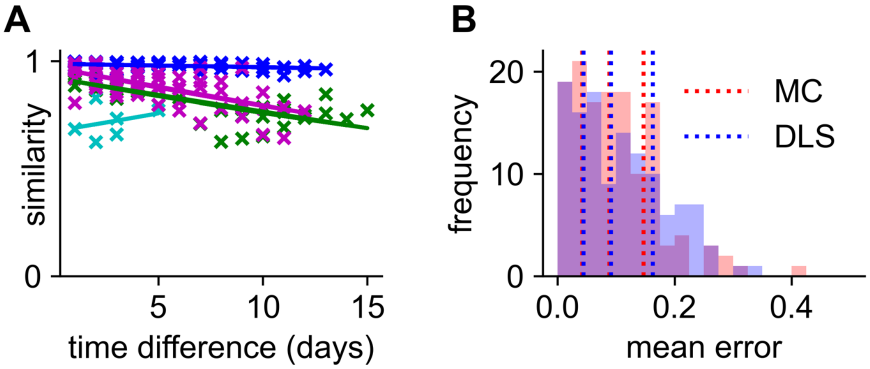
Extended Data Fig. 4: Stability as a function of time difference for different recording durations. (A) We performed analyses as in Figure 4C, plotting the neural similarity as a function of time difference for neurons recorded for at least N days with N ranging from 5 to 17 (c.f. N = 14 in Figure 4C). Dashed colored lines indicate controls as in Figure 4C. **(B)** As in (A), now for the wet-dog shake behavior instead of the lever-pressing task (c.f. Figure 6C).



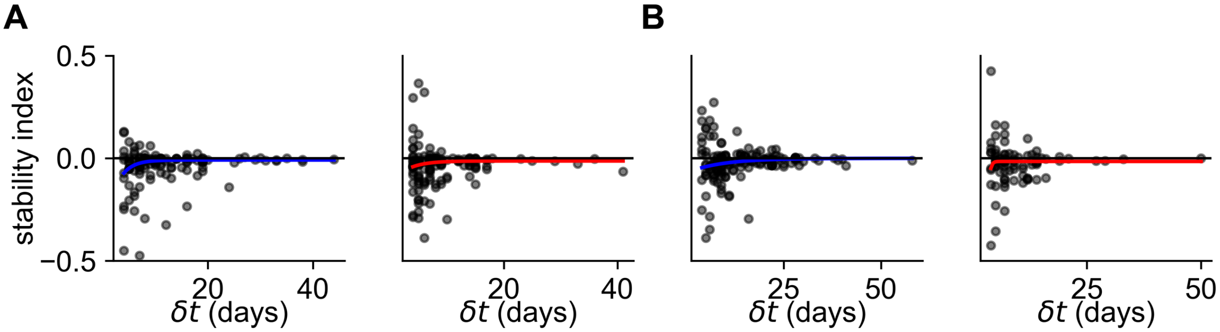
Extended Data Fig. 5: Investigating the population-level stability of neural data. It has previously been reported that stable neural activity can be identified in a common latent space even when there is a turnover of recorded neurons (Gallego et al., 2020). As we show in Figure 2, this can be consistent with either stable or drifting single unit activity. While we have already shown a high degree of similarity for single neurons, here we investigate whether 'aligning' the neural activity between sessions can identify a common subspace with even higher similarity. These analyses require simultaneous recording of a large population of neurons, which in general was not the case for our dataset (c.f. Figure 3C). Instead, we considered a single week of recording in a single animal with recordings from DLS (day 8-14 in Figure 3C), where we simultaneously recorded 16 neurons firing at least 10 spikes during the task on each day. **(A)** We first computed the similarity as a function of time difference as the correlation between single-neuron PETHs, averaged across neurons (black line). We then proceeded to align the neural activity on each pair of days using CCA and compute the similarity in the resulting aligned space as the average correlation across all dimensions. This CCA-aligned similarity was generally lower than the similarity averaged over individual neurons, suggesting that the neuron-aligned coordinate system is more stable than the CCA-aligned alternative (note that CCA performs a greedy alignment rather than finding the optimal alignment, which would provide an upper bound on the single-neuron similarity). **(B)** We proceeded to consider population decoding of behavior from neural activity, using the same data as in (A). We fitted a linear model to predict behavior from neural activity on each day using

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

crossvalidated ridge regression, and we tested the models on data from all other days. Here, we plot the performance as a function of time difference. Line and shading indicate mean and standard error across pairs of days with a given time difference. **(C)** We proceeded to compute stability indices for the data in (B) to see whether there was a significant negative trend. We bootstrapped the individual datapoints (before taking the mean) 10,000 times and estimated stability indices from each surrogate dataset. The distribution over the resulting stability indices was not significantly different from 0 (p = 0.34). **(D)** While the analysis in (A) suggests that the single neurons provide a good coordinate system for stable representations, it does not address the question of whether an aligned low-dimensional manifold can provide better decoding (Gallego et al., 2020). We therefore proceeded to train a population decoding model as in (B), but where the decoder was trained on the top 10 PCs from a single day and tested on the top 10 PCs from every other day after alignment via CCA (Gallego et al., 2020) (blue dashed line). We found that decoding performance from this aligned latent space was almost identical to the decoding performance from raw neural activity (black line). This provides further evidence that the stable aligned dynamics identified in previous work are the result of stable single-unit tuning curves.

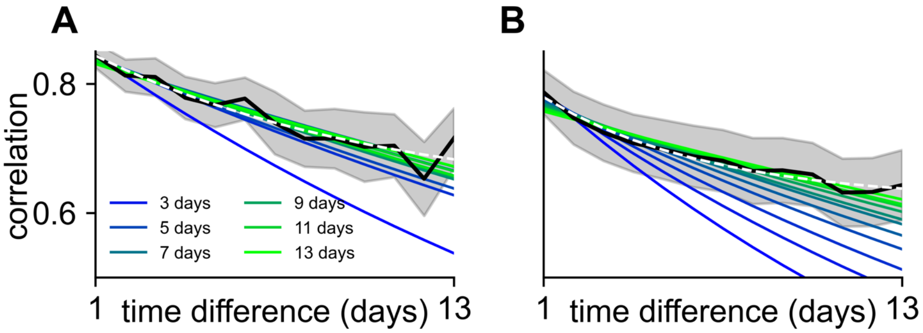


Extended Data Fig. 6: Exponential model fits. (A) Plots of PETH similarity against time difference for four example units (colors) together with exponential fits illustrating a range of different decay rates, baseline similarities, and durations of recording. Note that one of these example units (cyan) exhibits an apparent increase in stability over time due to the noisy nature of the data. Indeed, in a perfectly stable model (such as the stable RNN in Figure 2F), neurons will be equally likely to exhibit such an increase as they are to exhibit a decrease in similarity over time, leading to a median stability index of 0. Such noise is mitigated by increasing recording durations. **(B)** Distribution of the mean error of each model fit across the population of neurons recorded from MC (red) or DLS (blue). Vertical dashed lines indicate quartiles of the distributions.

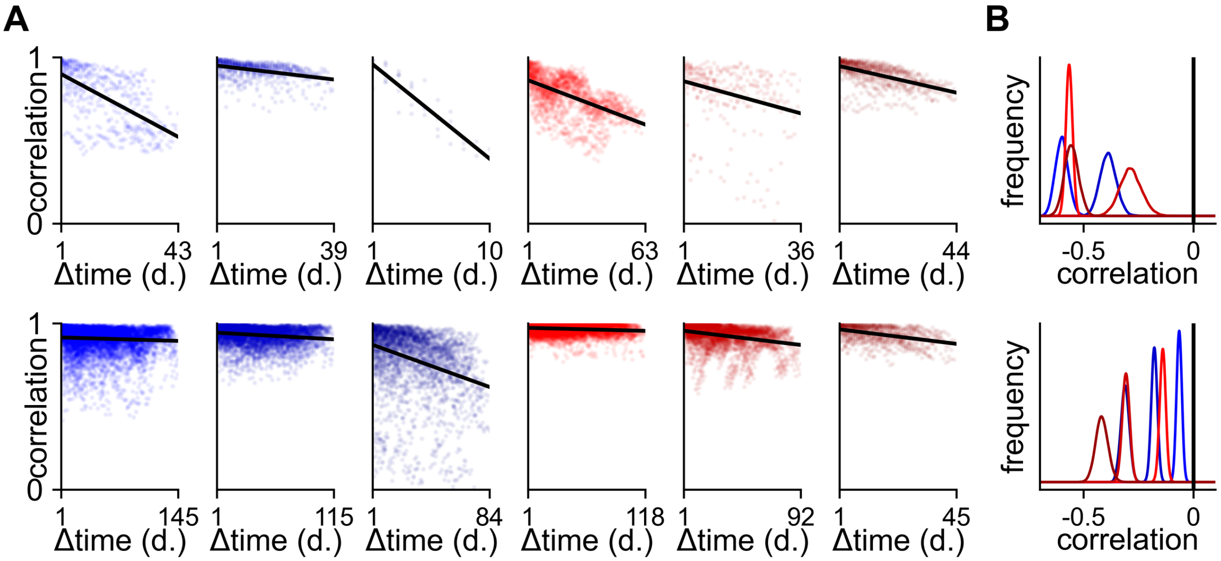


Extended Data Fig. 7: Stability indices for all neurons. (A) Stability indices for all neurons recorded from DLS (left) or MC (right) during the lever-pressing task. Solid lines indicate exponential fits as in Figure 4F. As the time difference increases, the variance decreases (due to the increase in data), and the median stability index gradually increases (c.f. solid lines). **(B)** As in (A) for the WDS behavior.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

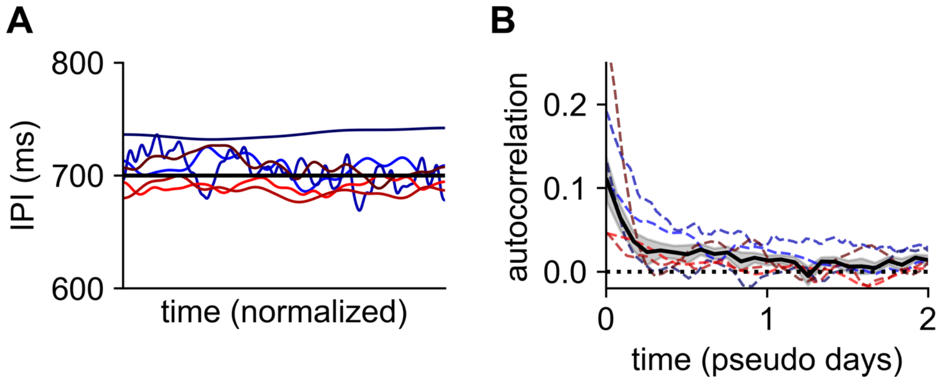


Extended Data Fig. 8: Exponential fits for different subsampled recording durations. We fitted exponential models to the average data across neurons recorded for at least 14 days (Figure 4C), but considering only data up to and including increasing time differences (legend). As the subsampled 'recording duration' increases, so does the stability index learned in the exponential model. White dashed lines indicate exponential model fits with a learnable baseline, which better capture the decreasing rates of decay with increasing recording duration.

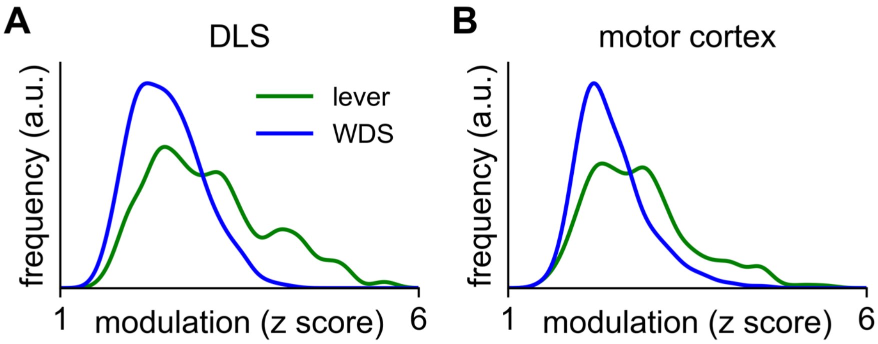


Extended Data Fig. 9: Behavioral drift across animals. (A) Scatter plots of the correlation between mean velocity profiles and time difference for all pairs of days in each animal. Top row: lever-pressing task; bottom row: wet dog shakes. Blue indicates animals with recordings from DLS, red from MC. **(B)** Distribution of correlations between time difference and behavioral similarity across all animals, generated by a bootstrap analysis of the data in (A). All animals exhibit a significant negative correlation between behavior and time difference in both the lever-pressing task and wet dog shake behavior ($p < 0.001$; bootstrap test).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465945; this version posted March 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Data Fig. 10: Inter-press intervals and autocorrelations. (A) Inter-press interval (IPI) for each animal, convolved with a 200-trial Gaussian filter. Time is normalized from 0 to 1 for each animal (n = 9365 ± 6886 trials). Black horizontal line indicates 700 ms. **(B)** We computed the IPI autocorrelation as a function of trial number and normalized time by the average number of trials per day for each animal (colored lines). Black line and shading indicate mean and standard error across animals. Task performance is only correlated over short timescales of 0.5-1 days despite behavioral drift on timescales of weeks (Extended Data Fig. 9). This suggests that behavioral changes are predominantly along 'task-null' directions that do not affect performance.



Extended Data Fig. 11: Task-modulation of neurons in the lever-pressing task and wet-dog shake behavior. (A) A PETH was computed across all trials for each neuron in 20 ms bins, and the time bin identified with the maximum deviation from the mean across all time bins. The corresponding z-score was computed, and the distribution of absolute values of these z-scores plotted across all DLS neurons for the lever-pressing task (green) and wet-dog shake behavior (blue). **(B)** As in (A), now for neurons recorded from MC.