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 established behaviors are 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 these circuits could change over time, provided that the network dynamics are contained within a manifold that produces stable behavior. To arbitrate between these possibilities, we recorded single unit activity in motor cortex and striatum continuously for several months 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 and the stochasticity of neural firing. These results suggest that stereotyped behaviors are stored and generated in stable neural circuits.

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 (Figure 1A), where they are stably maintained and can be reliably executed even after months of no intervening practice (Krakauer and Shadmehr, 2006; Melnick, 1971; Park and Sternad, 2015). The neural circuits underlying such motor skills have been the subject of extensive study (Churchland et al., 2012; Haith and Krakauer, 2013; Kawai et al., 2015; Peters et al., 2014; Wolpert and 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 and 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 and Nitz, 2004; Peters et al., 2017; Schoonover et al., 2021). While neural circuits being 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.

Competing theories and predictions

Multiple theories have been put forth to explain the apparent paradox of stable memories in plastic circuits subject to constant remodeling. 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 the behavioral output (Gallego et al., 2020; 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 1B) (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 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).

It has also been hypothesized that changes in neural circuits may be constrained to directions of connectivity space that do not affect the task-associated activity of single units during previously learned behaviors (Duncker et al., 2020; Kao et al., 2021a; Qin et al., 2021). This is similar to approaches for continual learning in the field of machine learning. Here, catastrophic forgetting can be mitigated by reducing learning-related plasticity at synapses important for already acquired skills (Duncker et al., 2020; Kao et al., 2021a; Kirkpatrick et al., 2017), thus minimizing changes in the activity patterns associated with existing behaviors. In this case, rather than drifting neural activity constrained only to a subspace that does not affect motor output, we would expect the activity patterns of task-relevant single neurons to remain constant or highly similar over time (Figure 1B) (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 devoted exclusively to learning and generating their one song, with plasticity largely restricted to a 'critical period' of development (Sizemore and 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.

Experimental challenges

Distinguishing the various possibilities outlined above has been attempted by recording neural activity over time during the performance of a specified behavior, either by means of electrophysiology (Carmena et al., 2005; Chestek et al., 2007; Flint et al., 2016; Fraser and Schwartz, 2012; Ganguly and 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 neural representations (Chestek et al., 2007; Flint et al., 2016; Ganguly and Carmena, 2009; Katlowitz et al., 2018; Stevenson et al., 2011), and others reporting drift in neural activity (Carmena et al., 2005; Liberti et al., 2016; Rokni et al., 2007). In lieu of high-quality long-term recordings of the same neurons, which can be technically challenging, a recent approach considered the stability of neural activity in low-dimensional latent spaces. This was done by applying linear dimensionality reduction to recordings from each experimental session followed by linear alignment of the resultant low-dimensional dynamics (Gallego et al., 2020). While this work suggests that the latent dynamics underlying stable motor behaviors are stable over time, it does not address the source of this stability. In particular, it is unclear whether such 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.

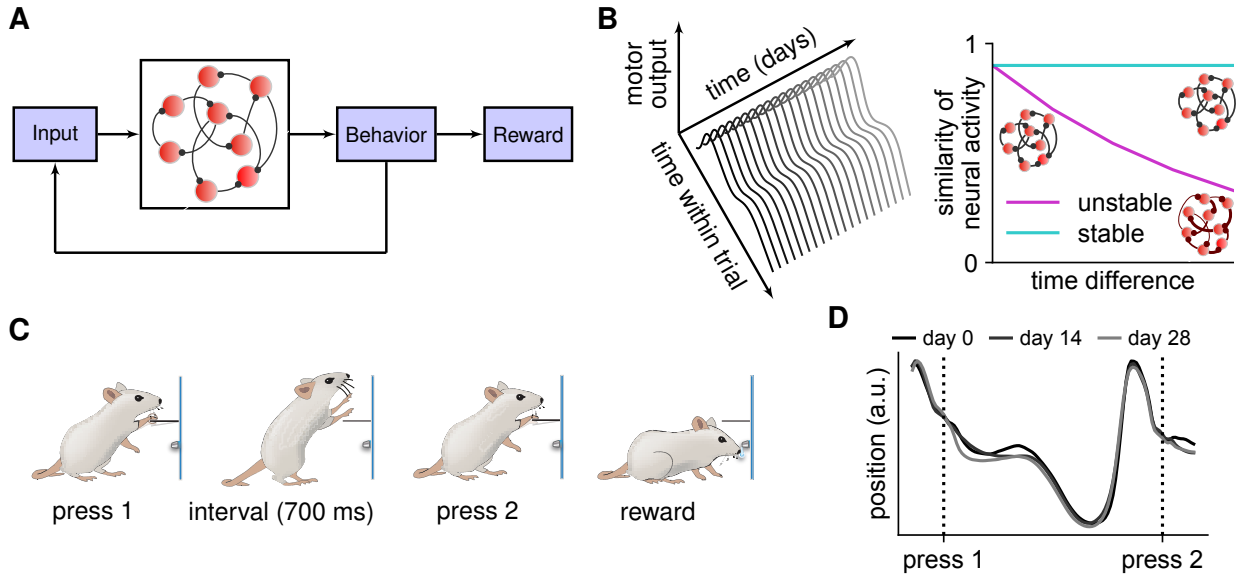

Figure 1: A paradigm for interrogating long-term neural and behavioral stability. (A) Motor circuits in the brain (recurrent network; red) drive behavior, which leads to reward if the task is executed correctly. In addition to local connectivity in the motor circuit, neural and behavioral dynamics are also determined by upstream inputs to the motor system, which can be affected by online behavioral feedback. **(B)** For a constant motor output over time (left), motor circuits driving a behavior can either remain constant or change in a behavioral 'null direction' (right) (Kao et al., 2021b). If neural dynamics are stable over time, the similarity of single unit neural activity for two trials or two sets of trials is independent of the time separating the trials (cyan). This could for example be achieved through stable connectivity (RNN schematic). Conversely, if the neural dynamics driving the behavior change over time, the similarity of task-associated neural activity decreases with increasing time difference (magenta) (Clopath et al., 2017). **(C)** 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. **(D)** Mean forelimb position parallel to the floor across trials for an example rat on three days spaced two weeks apart.

In this work, we first demonstrate – using an artificial neural network – how long-term single-unit recordings during a stably executed behavior can effectively distinguish these scenarios. We then go on to perform such recordings in animals performing stable behaviors – both learned and innate (Figure 1C). To reduce any source of neural variability not directly related to the behavior in question, such as fluctuating environmental conditions or changes in the animal's internal state, we performed our experiments in a highly stable and controlled environment. To account for any changes in task-irrelevant movements, we also recorded the animals' behavior at high spatiotemporal resolution (Figure 1D) (Chestek et al., 2007; Musall et al., 2019). Our combined neural and behavioral recordings revealed that neural circuit dynamics at the level of single neurons are highly stable. The small amount of drift in task-related neural activity seen over several weeks of recordings could be accounted for by a concomitant slow drift in the behavior and the stochasticity of neural firing (Stevenson et al., 2011). These results suggest that stable behaviors are generated by highly stable motor circuits that maintain a fixed mapping between single neuron activity and motor output.

Results

Network models of stable and unstable motor circuits

When analyzing the stability of task-associated neural activity, it is important to distinguish between stability at the population level (e.g. in the form of stable latent dynamics) and stability at the level of single task-associated neurons. We start by highlighting the difference between these scenarios in an artificial model system and demonstrating the necessity of longitudinal single-unit recordings to arbitrate between them. 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 and Buonomano, 2013; Sussillo and Abbott, 2009). Analyzing single unit activity patterns and low-dimensional latent trajectories in this artificial network allowed us to assess the degree to which these proxies for network dynamics can distinguish between circuits with stable and drifting dynamics. The network consisted of 250 fully connected recurrent units, which were subject to Gaussian process noise and projected to 5 linear readout units. We trained the network using gradient descent (Kingma and Ba, 2014) to generate smooth target output trajectories – the simulated control signal (Figure S1; 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 2A).

Importantly, due to the degeneracy of the circuit, multiple distinct networks can achieve the same target output. Thus, we could 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 2B; 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 2B; 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 representational similarity with the similarity of aligned latent dynamics across sessions (Gallego et al., 2020). To do this, we simulated recordings in which a subset of the total population is tracked by randomly sampling 50 neurons. We then used PCA to reduce the dimensionality of each recording from 50 to 10. This was followed by alignment of the resulting latent trajectories using canonical correlation analysis (CCA) on each pair of simulated sessions. We used the resulting canonical correlations as a measure of similarity for the latent dynamics (Methods) (Gallego et al., 2020). In contrast to the single-unit correlations, similarity of the latent dynamics was highly similar between pairs of identical and pairs of different networks (Figure 2B). This reflects the fact that even though distinct networks differ in the activity patterns of individual neurons, they may have similar population level statistics reflecting the conserved nature of the task.

To intuit how activity patterns change over time in a network with slowly drifting connections, 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 a series of RNNs with progressively more dissimilar connectivity, yet which all produced the same output (Figure 2D; Figure S1; Methods) – a phenomenological model of neural drift where the position of a network within such an interpolation series is a proxy for time. We considered 7 networks tiling the space from the original network to 70% along the interpolant (Methods), at which point the parameters had only little correlation with the initial parameters (Figure S1) and neural activity was largely uncorrelated (Figure 2D). This resulted in a 'week' of recording from a recurrent network with

drifting parameters which enabled us 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 this series of 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 2C). 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 2D). 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 'behavioral' output and the single-unit neural activity in this stable network were highly stable over time (Figure 2D).

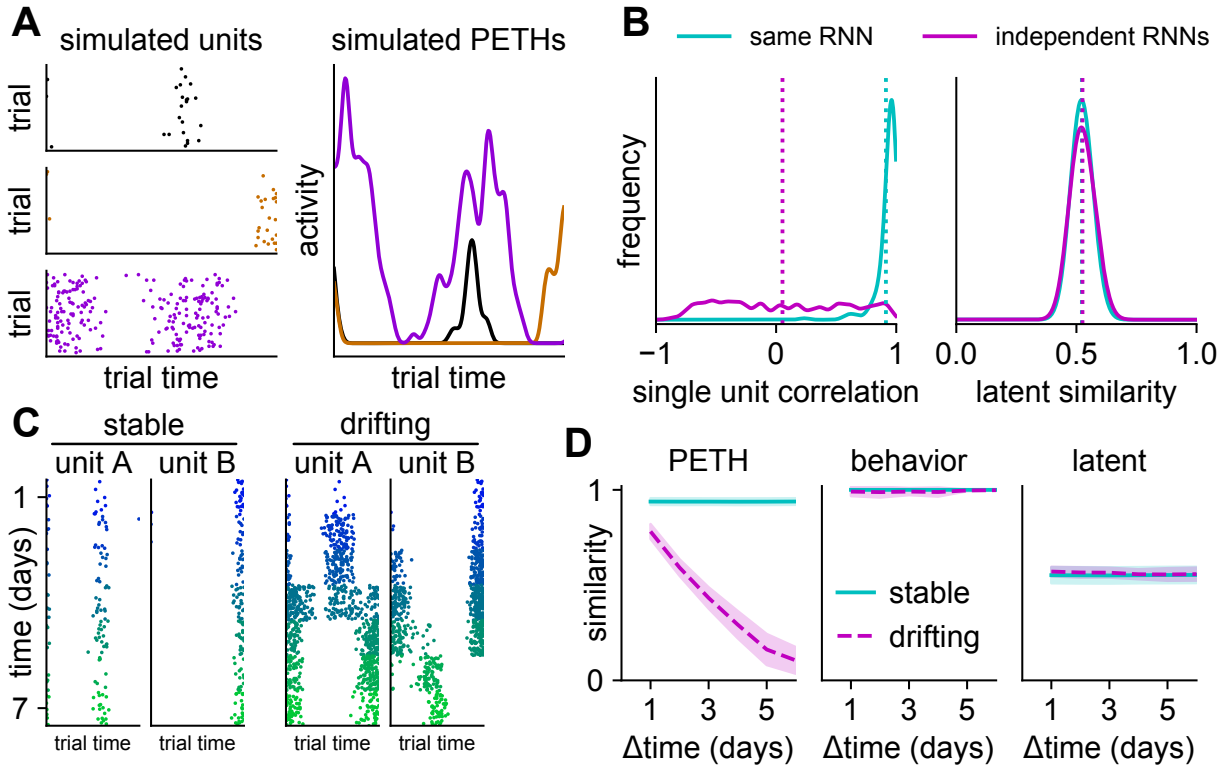

Figure 2: Analyzing neural stability in a recurrent network model. (A) Activity of three example recurrent neurons after training (Methods). (Left) Raster plot of spike times across 100 trials. (Right) Peri-event time histograms (PETHs) computed across the corresponding trials. **(B; left)** Distribution across neurons of the correlation between PETHs computed for two 'sessions' as in (A) (cyan). A second RNN was independently trained to produce the same output, and PETH correlations were approximately uniformly distributed between these distinct networks (magenta). **(B; right)** Distribution of latent similarities between the same (cyan) or different (magenta) networks after alignment using CCA, considering 250 random samples of 50 neurons which were matched across each pair of networks. **(C)** Example raster plots as in (A) across 7 different sessions for a model network exhibiting either stability (left) or representational drift (right). Colors indicate the progression of time from day 1 (blue) to day 7 (green) **(D)** 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 (C)) for the stable RNN (cyan) and the drifting RNN (magenta). Lines and shadings indicate mean and standard deviation across 10 independent networks.

Finally, we again considered how such single-unit analyses differ from approaches that consider the stability of low-dimensional latent trajectories encoded by the neurons (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. We found that the latent dynamics of the RNN with drifting single unit representations did not become more dissimilar over time, reflecting the fact that the network retained similar population statistics despite the changing activity profiles of individual units (Figure 2D).

Our findings in this artificial model system thus highlight 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 level of single units implies stability at the level of latent trajectories, while stable latent dynamics can be driven by either stable or drifting single-unit activity patterns.

Behaviorally locked activity of single neurons in motor cortex and DLS is stable

To investigate the stability of biological motor circuits experimentally, we trained rats (n=6) to perform a timed lever-pressing task in which they receive a water reward for pressing a lever twice with an inter-press interval of 700 ms. Rats learn to solve the task by developing complex stereotyped movement patterns (Dhawale et al., 2021; Kawai et al., 2015). Since the task is kinematically unconstrained (meaning there are many 'motor solutions' for the task) and acquired through trial-and-error, each animal converges on its own idiosyncratic solution (Figure 3A, 3B). However, once acquired, the individually distinct behaviors persist over long periods of time (Figure 3A). To reduce any 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 environment and training protocol (Poddar et al., 2013).

After reaching expert performance, animals were implanted with tetrode drives for neural recordings (Dhawale et al., 2017) (Methods). We targeted two main nodes of the motor system: motor cortex (MC) and dorsolateral striatum (DLS) (Hunnicutt et al., 2016). 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).

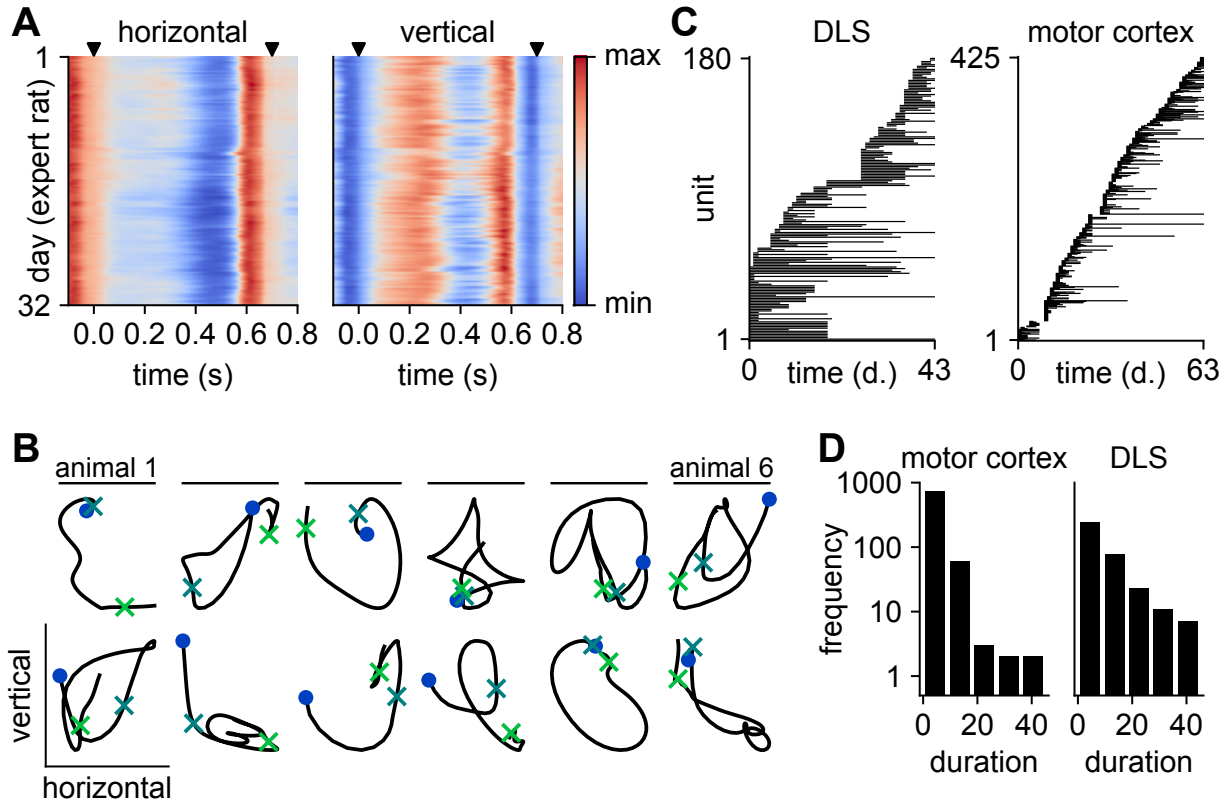

Figure 3: **Experimental recordings of behavior and neural activity. (A)** Right forelimb trajectories in the horizontal and vertical directions (parallel and perpendicular to the floor respectively; c.f. Figure 1C) for an example expert rat. Color indicates forelimb position. Trials were subselected based on task performance, 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 motor cortex (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 motor cortex (right). Note that the data used in this study has previously been analyzed in (Dhawale et al., 2017, 2021).

The combination of stable behavior and continuous neural recordings during the task 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 representational stability at the level of single neurons. We considered the PETHs of all units combined across time and found that units in both MC and DLS fired preferentially during particular phases of the learned behavior (Figure 4A) (Dhawale et al., 2017). Remarkably, 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).

To quantitatively compare neural activity profiles across days, we considered all trials on a particular day and computed a PETH for each neuron (Figure 4B). We then computed the correlation between pairs of PETHs as a function of their time difference, similar to our RNN analyses (Figure 2) and to previous studies in visual and motor circuits (Deitch et al., 2020; Dhawale et al., 2017). We first considered neurons

recorded for at least two weeks and found that the mean PETH similarity remained high in both DLS and motor cortex (Figure 4C). This is consistent with results from the stable RNN model (Figure 2D), and it suggests that learned motor behaviors are driven by circuits that do not change over the duration of our recordings despite the life-long structural and functional plasticity in these circuits (Holtmaat and Svoboda, 2009; Peters et al., 2017; Wolff et al., 2019; Xu et al., 2009). While our similarity analyses focused on the time-varying patterns of neural activity in the form of PETHs, we observed similar results when analyzing task-associated firing rates (Figure S2). This is consistent with our previous report of homeostatically maintained firing rates in DLS and MC over long time periods (Dhawale et al., 2017).

In contrast to our RNN model, the experimental data contained neurons that were recorded for different durations (Figure 3D). This introduces additional variability and emphasizes neurons recorded for longer durations if data is naively pooled across units. To combine information across neurons, we instead considered the PETH similarity as a function of the time difference between PETHs for each neuron individually. We then fitted an exponential model $\rho = \beta e^{\alpha\,\delta t}$ to the Pearson correlation ρ as a function of time difference δt for each neuron (Methods; see Figure S3 for example fits and model errors). We refer to the exponential parameter α with units of inverse time as a 'stability index'. Note that the stability index $\alpha = -\tau^{-1}$ corresponds to the (negative) inverse time constant in an exponential decay model, and that fitting this inverse parameter is numerically stable even for slow decays where the time constant approaches infinity.

We then considered the distribution of such stability indices across 'task-associated' neurons ($\beta > 0.2$; Methods) recorded for more than 3 days. In a null-model where mean neural activity remains constant, the PETH similarity should be independent of the time difference for all units (c.f. Figure 2D). 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. When performing this analysis, we found that the population-level similarity distributions were indeed centered near zero (Figure 4D). 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: $\alpha_{median} = -0.014$, $\tau_{median} = 69$ days, $p < 0.001$; MC: $\alpha_{median} = -0.029$, $\tau_{median} = 34$ days, $p < 0.001$; permutation test).

At first glance this appears to suggest that neural representations drift on a timescale of several weeks. However, previous work has highlighted how intrinsic neural noise and measurement noise can bias estimates of the stability of neural representations with finite firing rates (Stevenson et al., 2011). We hypothesized that our estimates of stability would similarly be sensitive to recording duration due to the stochastic nature of the neural firing patterns, and that this could lead to systematic underestimation of the neural stability. To further investigate this possibility, we simulated synthetic neural populations with trial structure matched to the experimental data and a range of different firing rates (Methods). We found that for all firing rates, the apparent stability of the neural population increased with recording duration, even in this synthetic population with no systematic drift (Figure 4E). To probe whether this was also the case in the experimental data, we pooled data across all animals for each region of recording, again considering neurons recorded for at least two weeks. We then subsampled this data in different recording windows to simulate the effect of shorter recording durations. Finally, we computed the mean stability index for each unit and recording duration, z-scored the results across durations for each unit, and calculated the average stability as a function of recording duration across units. Similar to the synthetic data, we found a strong effect of recording duration on apparent neural stability, suggesting that our initial estimates of the timescales of neural drift underestimate the actual stability of the task-associated neural representations (Figure 4F).

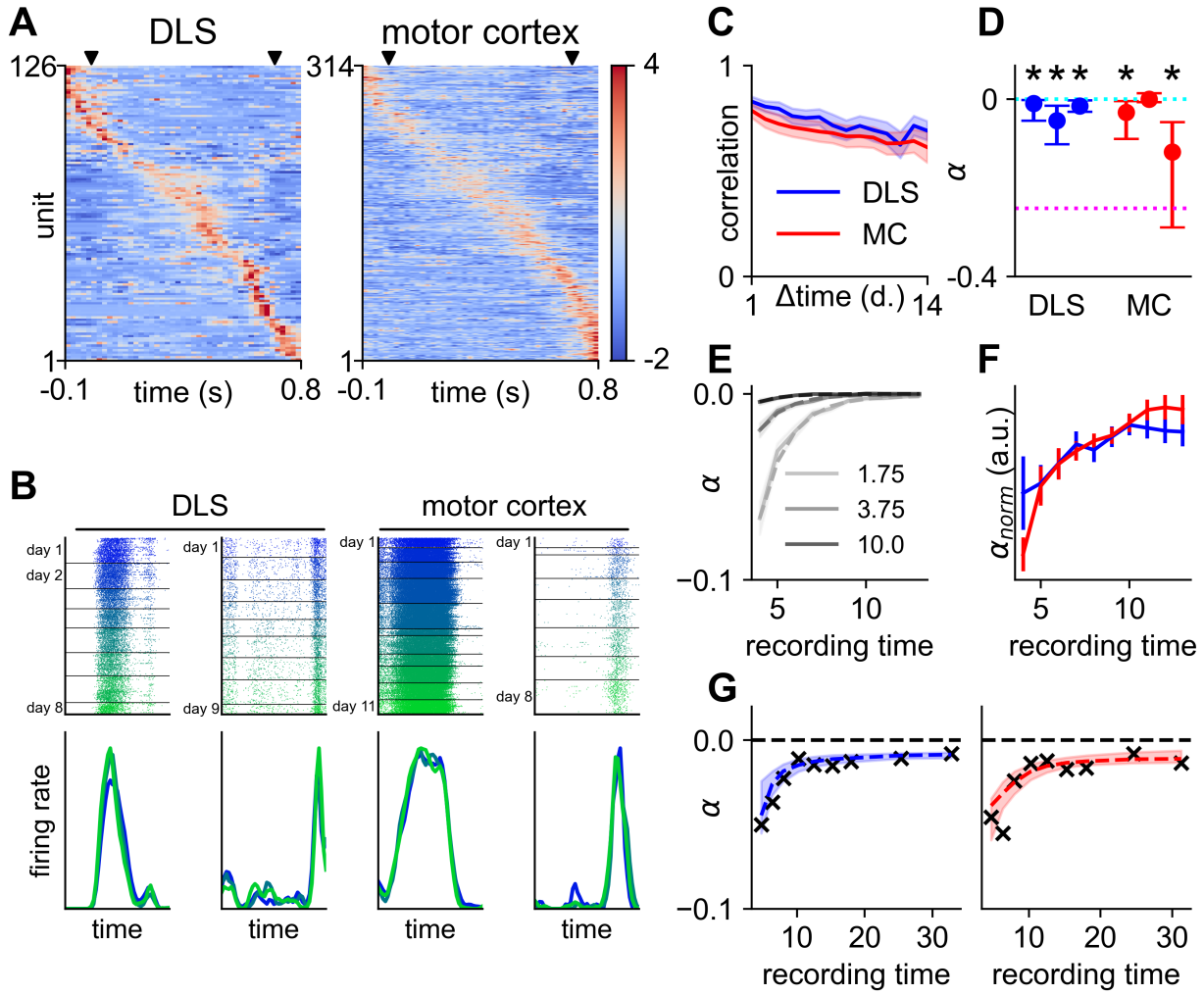

Figure 4: Single-unit activity is stable over time in DLS and motor cortex. (A) z-scored PETHs across trials and sessions for all units firing at least 100 spikes during the lever-pressing task for two example rats, sorted according to the activity peak from a set of held-out trials. x-axis indicates 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 motor cortex (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 examples 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 motor cortex (red) or DLS (red) and plotted as a function of time between days (n = 39 neurons for DLS; n = 27 neurons for motor cortex). Shaded regions indicate standard error across units. **(D)** Quartiles of the distribution of stability indices for each animal (see main text for details). Asterisks indicate $p < 0.05$ for the median stability index being different from zero (permutation test; Methods; n = [90, 22, 13] neurons for DLS; [121, 6, 7] for motor cortex). Horizontal dashed lines indicate median stability indices of the drifting (magenta) and stable (cyan) RNN models (Figure 2D). **(E)** Stability index for stable synthetic neural populations with increasing firing rates (light to dark; legend) simulated for different recording durations. Solid lines and shadings indicate mean and standard error across 10 simulations. Dashed lines indicate exponential model fits (Methods). **(F)** Mean and standard error of the stability index as a function of subsampled recording duration for experimental units recorded for at least 14 days (see main text for details). Results were z-scored for each unit prior to averaging. **(G)** Rolling median of the stability index (crosses) for units recorded for different recording durations (x axis), plotted together

with an exponential model fit to the raw data (dashed line; Methods). Shadings indicate interquartile intervals from bootstrapping the units included in the model fit.

To better estimate drift in the recorded population, we binned the stability indices of all neurons by the duration of the recordings. 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 4G). To estimate the stability of these neural populations in the limit of long recording durations (or equivalently in the absence of noise), we proceeded to fit an exponential model to the data of the form $\alpha = -a - b\exp(-c\,t)$, which we found empirically to fit both the synthetic and experimental data well (Figure 4E, 4G). In this model, the parameter $\tau_\infty = a^{-1}$ provides an estimate of the asymptotic stability of the population (Methods). We fitted the model 5,000 times with bootstrapping of the data and found a median asymptotic stability of $\tau_\infty = 134$ days for DLS (interquartile range 102-173) and $\tau_\infty = 94$ days for motor cortex (interquartile range 76-194).

In summary, we thus found that neural activity associated with the learned motor skill was qualitatively very stable over periods of several days and weeks (Figure 4B, 4C). Extrapolating to longer recording durations, we found evidence for only a slow drift on a timescale of ~100 days after accounting for the systematic bias resulting from finite recording durations (Figure 4G). This suggests that motor memories are retained by maintaining a stable circuit with stable activity patterns over several months.

Neural drift is correlated with behavioral changes

In the previous section, we demonstrated the importance of accounting for the stochasticity inherent in neural activity when assessing the stability of neural representations. However, the observed drift in neural activity, as slow as it is, is still likely to be an underestimate of the true stability, even after accounting for such noise. In particular, 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). Indeed, even after performance saturates and stabilizes in a motor task, humans and animals 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 tasks we analyze, it might explain the slow drift we observed in neural activity (Figure 4). This would in turn suggest a more stable circuit linking neural activity to behavior than revealed by our analyses of neural data alone. To probe the degree to which the neural drift we see could 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 drifted over time.

We first computed z-scored forelimb velocities by subtracting the mean velocity across all trials for each time point and normalizing 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). 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 slow but consistent decrease in the similarity of motor output as a function of time difference (Figure 5B; Figure S5). Importantly, such behavioral drift occurred despite stable task performance (Figure S6) and thus took place in a 'task-null' subspace of the behavioral output. 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 is stable, the 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. However, an alternative explanation is that the drifts in neural activity and behavior are both independently driven by the passing of time. We therefore proceeded to control for this confounding factor (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 computed the correlation between these two features across all consecutive days for each neuron, which should be positive if the drift in neural activity is causally related to the drift in motor output (Figure 5C). We repeated this analysis for all units to generate a distribution over correlations between the drift in neural activity and the drift in forelimb kinematics. The mean of this distribution was at $\rho = 0.32$ for DLS and $\rho = 0.26$ for motor cortex, both of which were significantly higher than a null distribution generated by permuting the behavioral data to break any correlations with the neural drift (Figure 5D; $p < 0.001$; permutation test).

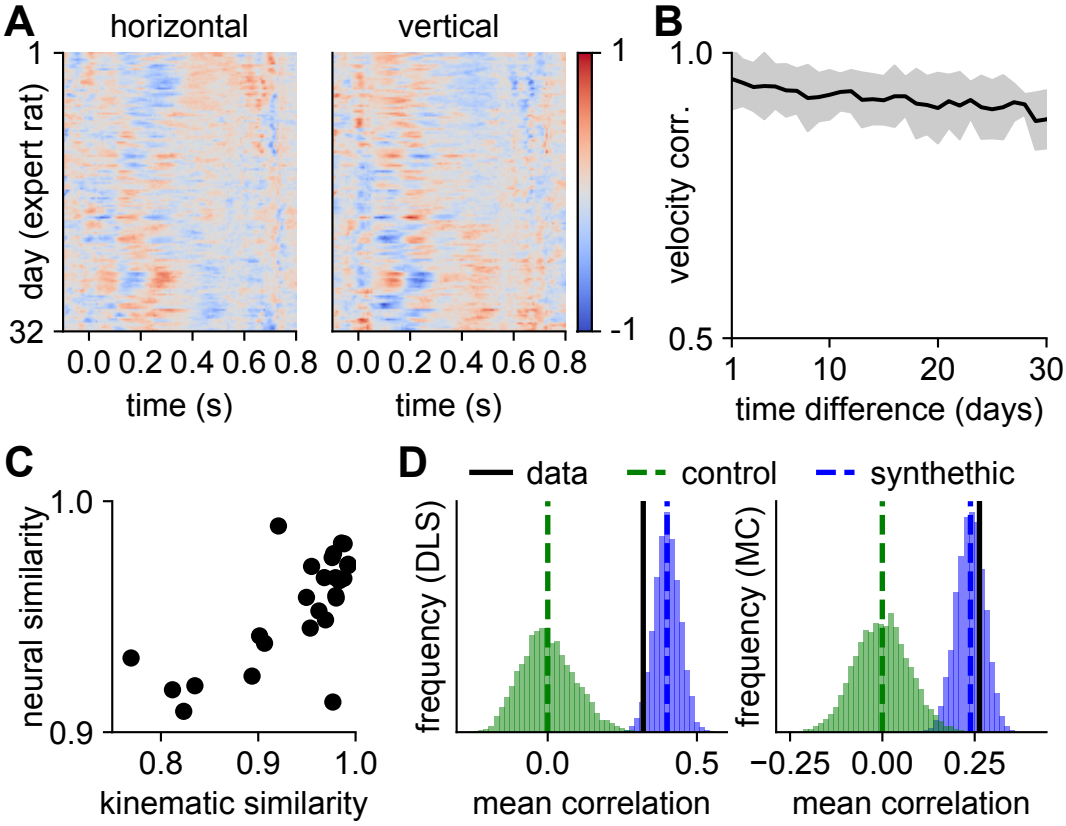

Figure 5: Behavioral drift over months of recording. (A) Forelimb velocities for the example animal from Figure 3A, plotted as z-scores with the mean subtracted. The stereotyped motor sequence masks a behavioral drift across days and weeks. **(B)** Mean and standard deviation of behavioral similarity as a function of time difference, averaged across all pairs of days for the example animal in (A) (see Figure S5 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)** Mean correlation between neural and behavioral similarity across neurons (black). Green histogram indicates a control distribution constructed by 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, which imposes a stable mapping between these two features. Left panel includes all units from DLS, right panel from motor cortex.

We then investigated 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 generated a synthetic dataset by fitting 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. We then proceeded to generate synthetic neural activity on each trial from the recorded behavior and computed the correlation between the simulated neural drift and the experimentally observed behavioral drift. Here, we found an average correlation with behavior of $\rho = 0.40$ for DLS and $\rho = 0.24$ for motor cortex. This is substantially more similar to the correlation values found in the experimental data than to the null distribution with no relation between the drift in neural activity and behavior (Figure 5D; p = 0.03 and p = 0.74 for the experimental average correlation across neurons being larger than or equal to the corresponding synthetic value when resampling datasets from the generative model).

Our behavioral analyses thus confirm that the slow drift in neural activity is driven, at least in part, by a concomitant behavioral drift in a task null-space. Additionally, the correlation between 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 the behavioral changes could account for the majority of the experimentally observed neural drift.

Neural representations are stable during an innate behavior

While the majority of studies on neural stability have considered behaviors that are learned, 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 and Carmena, 2009), or singing a song (Katlowitz et al., 2018; Liberti et al., 2016), 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 and Harding, 1981; Kleinrok and Turski, 1980; Marshall et al., 2021; Martin et al., 1963). Given the stereotyped frequency of these shakes, it is possible to extract instances of such WDS events from an accelerometer attached to the head of each animal (Methods). Each WDS event lasts approximately half a second, and each animal performs on the order of 50 WDS per day. This allowed us to analyze them in a 'trial-like' manner, similar to the lever-pressing task.

We found that the accelerometer readout corresponding to WDS events was consistent across time (Figure 6A), and we identified units in both DLS and motor cortex 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). This suggests a stable circuit linking their activity to the innate WDS behavior. When computing PETH correlations over time, we also found that they remained high throughout the period of recording (Figure 6C), although the baseline trial-to-trial similarity was lower than for the timed lever-pressing task. This likely reflects a lesser involvement of DLS and motor cortex in the specification and control of innate behaviors such as WDS. This would also be consistent with our identification of weaker task modulation for the WDS behavior compared to the lever-pressing task (Figure S7). Hence, the observed activity patterns in motor cortex and DLS during WDS, and consequently any potential drift over time, is likely to reflect ongoing activity in connected areas, presumably those processing sensory feedback and motor efference (Hatsopoulos and Suminski, 2011). Since the activity patterns in neural circuits not controlling the behavior are likely less

constrained relative to the motor output, we might expect a higher degree of representational drift during the WDS behavior than for the lever-pressing task.

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 (Figure 6D). When pooling neurons across experimental groups, we found this drift to be statistically significant for both groups (DLS: $\alpha_{median} = -0.025$, $\tau_{median} = 41$ days, $p < 0.001$; MC: $\alpha_{median} = -0.023$, $\tau_{median} = 43$ days, $p < 0.001$; permutation test). We expected that this apparent drift would be partly driven by the stochasticity of neural activity and 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 $\tau_\infty = 113$ days for neural drift in DLS and $\tau_\infty = 183$ days in motor cortex (Figure 6E). These results suggest that the neural representations of this innate behavior are also highly stable over time, similar to our observations for learned motor skills.

Based on our analyses of the lever-pressing task, we wondered if some of the residual representational drift resulted from 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 (Figure S5). 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 effect of behavioral drift on the neural drift (DLS: $\rho = 0.088$, $p = 0.005$, MC: $\rho = 0.069$, $p = 0.104$; permutation test). While these correlations are quite 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 is largely due to the slow rate of drift in both behavior and neural activity, which leads to a small signal-to-noise ratio.

Under the assumption that the recorded neural circuits drive the WDS behavior, the slow neural drift thus appears to occur at least partially along behaviorally potent directions of neural state space. Conversely, under our favored view of the recorded neural activity reflecting mainly sensory input and motor efference, the change in activity can be partially explained by a change in inputs, suggesting that the sensory representations are highly stable. Both here and in the case of the timed lever-pressing task, we thus interpret the observed slow drift in neural activity to be accounted for, in large part, by slowly changing behavior.

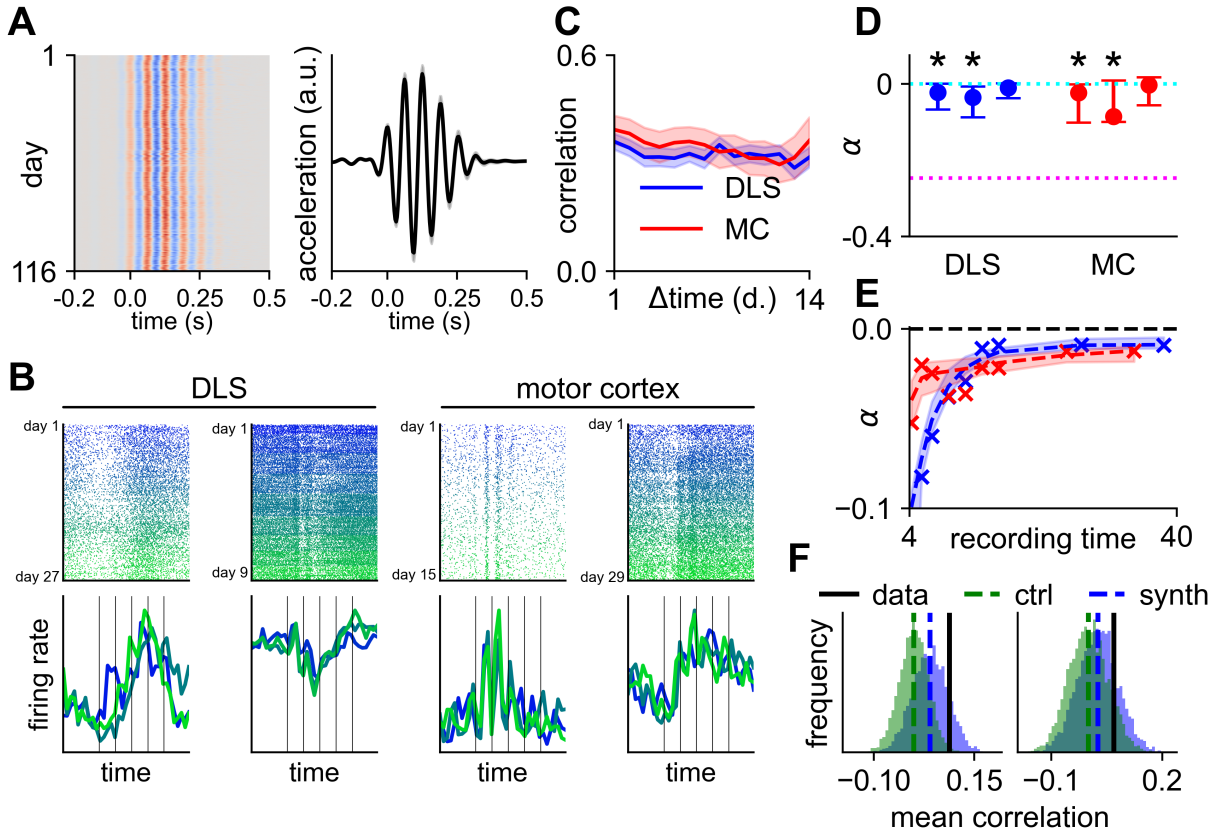

Figure 6: Neural activity is stable during an innate behavior (WDS). (A; left) Acceleration perpendicular to the floor across all wet-dog shakes (WDSs) recorded over 100 days in an example animal. **(A; right)** We computed the mean acceleration across 'trials' for each day. Line and shading indicate 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 motor cortex (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) or DLS (blue) and plotted as a function of time difference. Shadings indicate standard error across units. **(D)** Quartiles of the distribution of stability indices for each animal. Asterisks indicate $p < 0.05$ for the median stability index being different from zero (permutation test; Methods; n = [93, 62, 15] neurons for DLS; [84, 11, 5] for MC). Horizontal lines indicate median stability indices of the drifting (magenta) and stable (cyan) RNN models. **(E)** Rolling median of the stability index (crosses) for neurons recorded for different durations plotted together with an exponential model fit to the raw data (dashed line; Methods). Shadings indicate interquartile intervals from bootstrapping the units included in the model fit. **(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.

Discussion

We have investigated whether stereotyped motor behaviors are driven by stable neural 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 across several months, as expected for a stable control network. Accounting for the stochasticity of neural dynamics was important to reveal this stability, which highlights the importance of recording single units over long durations when assessing stability (Figure 4). We did observe a slow drift at the population level on a timescale of approximately 100 days, which was accompanied by a concomitant slow 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.

Considerations of behavioral variability in studies of neural stability

The 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 but not task performance. It remains to be seen whether this behavioral drift constitutes a learning process that optimizes a non-task-relevant utility function such as energy expenditure (Srinivasan and Ruina, 2006) or magnitude of the control signal (Todorov and Jordan, 2002), or whether it instead reflects a random walk in a degenerate motor space that preserves task performance (Chaisanguanthum et al., 2014; Qin et al., 2021).

Our observation of correlated neural and behavioral drift also highlights the importance of high-resolution behavioral measurements when investigating neural circuit dynamics, since most tasks studied in neuroscience do not fully constrain the behavioral output. When recording from neurons that also encode non-task related or unmeasured variables, the resulting task representations might appear unstable if such auxiliary variables are not sufficiently constrained. This is particularly relevant in the case of kinematic drift since task-unrelated movement is known to be strongly represented in cortical activity (Musall et al., 2019). However, it is also true for other external variables, such as changes in animal weight, time of the year, and measurement variability. As a result, our reported neural stability in 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 and infinite recording durations.

Prior studies of neural stability

It is worth noting the contrast between our results and previous studies that found neural representations in sensory and motor circuits to drift over time (Carmena et al., 2005; Deitch et al., 2020; 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, 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, and importantly, 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 stochasticity 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., 2020; 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 and Lieber, 2019; Steinmetz et al., 2021), especially if such advances can eventually be combined with methods for chronic recordings to track changes in the waveforms of individual neurons (Dhawale et al., 2017).

The finding of stable motor representations by us and others (Chestek et al., 2007; Flint et al., 2016; Ganguly and 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. However, it is also possible 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 and Escola, 2020; Roxin and Fusi, 2013; Rule et al., 2020). This is consistent with 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 motor cortex 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 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 (Wu et al., 2020). This suggests 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 and 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 neural 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 and Fusi, 2016; Duncker et al., 2020; Kao et al., 2021a; 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-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

Barthó, P., Hirase, H., Monconduit, L., Zugaro, M., Harris, K.D., and Buzsáki, G. (2004). Characterization of Neocortical Principal Cells and Interneurons by Network Interactions and Extracellular Features. Journal of Neurophysiology *92*, 600–608.

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

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

Chaisanguanthum, K.S., Shen, H.H., and Sabes, P.N. (2014). Motor Variability Arises from a Slow Random Walk in Neural State. J. Neurosci. *34*, 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., and Shenoy, K.V. (2007). Single-Neuron Stability during Repeated Reaching in Macaque Premotor Cortex. J. Neurosci. *27*, 10742–10750.

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., et al. (2019). High-Density, Long-Lasting, and Multi-region Electrophysiological Recordings Using Polymer Electrode Arrays. Neuron *101*, 21-31.e5.

Churchland, M.M. (2015). Using the precision of the primate to study the origins of movement variability. Neuroscience *296*, 92–100.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Clopath, C., Bonhoeffer, T., Hübener, M., and Rose, T. (2017). Variance and invariance of neuronal long-term representations. Philos Trans R Soc Lond B Biol Sci *372*, 20160161.

Degenhart, A.D., Bishop, W.E., Oby, E.R., Tyler-Kabara, E.C., Chase, S.M., Batista, A.P., and Yu, B.M. (2020). Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nature Biomedical Engineering *4*, 672–685.

Deitch, D., Rubin, A., and Ziv, Y. (2020). Representational drift in the mouse visual cortex. BioRxiv 2020.10.05.327049.

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

Dhawale, A.K., Wolff, S.B.E., Ko, R., and Ölveczky, B.P. (2021). The basal ganglia control the detailed kinematics of learned motor skills. Nat Neurosci *24*, 1256–1269.

Driscoll, L.N., Pettit, N.L., Minderer, M., Chettih, S.N., and Harvey, C.D. (2017). Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. Cell *170*, 986-999.e16.

Duncker, L., Driscoll, L., Shenoy, K.V., Sahani, M., and Sussillo, D. (2020). Organizing recurrent network dynamics by task-computation to enable continual learning. Advances in Neural Information Processing Systems *33*.

Fletcher, A., and 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*, 811–813.

Flint, R.D., Scheid, M.R., Wright, Z.A., Solla, S.A., and Slutzky, M.W. (2016). Long-Term Stability of Motor Cortical Activity: Implications for Brain Machine Interfaces and Optimal Feedback Control. J. Neurosci. *36*, 3623–3632.

Fraser, G.W., and Schwartz, A.B. (2012). Recording from the same neurons chronically in motor cortex. Journal of Neurophysiology *107*, 1970–1978.

Fu, M., Yu, X., Lu, J., and Zuo, Y. (2012). Repetitive motor learning induces coordinated formation of clustered dendritic spines in vivo. Nature *483*, 92–95.

Gallego, J.A., Perich, M.G., Miller, L.E., and Solla, S.A. (2017). Neural Manifolds for the Control of Movement. Neuron *94*, 978–984.

Gallego, J.A., Perich, M.G., Chowdhury, R.H., Solla, S.A., and Miller, L.E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. Nat Neurosci *23*, 260–270.

Ganguly, K., and Carmena, J.M. (2009). Emergence of a Stable Cortical Map for Neuroprosthetic Control. PLOS Biology *7*, e1000153.

Golowasch, J., Casey, M., Abbott, L.F., and Marder, E. (1999). Network Stability from Activity-Dependent Regulation of Neuronal Conductances. Neural Computation *11*, 1079–1096.

Haith, A.M., and Krakauer, J.W. (2013). Model-Based and Model-Free Mechanisms of Human Motor Learning. In Progress in Motor Control, M.J. Richardson, M.A. Riley, and K. Shockley, eds. (New York, NY: Springer), pp. 1–21.

Hatsopoulos, N.G., and Suminski, A.J. (2011). Sensing with the Motor Cortex. Neuron *72*, 477–487.

Hengen, K.B., Lambo, M.E., Van Hooser, S.D., Katz, D.B., and Turrigiano, G.G. (2013). Firing Rate Homeostasis in Visual Cortex of Freely Behaving Rodents. Neuron *80*, 335–342.

Hengen, K.B., Torrado Pacheco, A., McGregor, J.N., Van Hooser, S.D., and Turrigiano, G.G. (2016). Neuronal Firing Rate Homeostasis Is Inhibited by Sleep and Promoted by Wake. Cell *165*, 180–191.

Hennequin, G., Vogels, T.P., and Gerstner, W. (2014). Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. Neuron *82*, 1394–1406.

Hennequin, G., Ahmadian, Y., Rubin, D.B., Lengyel, M., and 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*, 846-860.e5.

Holtmaat, A., and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. Nature Reviews Neuroscience *10*, 647–658.

Hong, G., and Lieber, C.M. (2019). Novel electrode technologies for neural recordings. Nat Rev Neurosci *20*, 330–345.

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., et al. (2021). Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. ELife *10*, e51675.

Hunnicutt, B.J., Jongbloets, B.C., Birdsong, W.T., Gertz, K.J., Zhong, H., and Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization. ELife *5*, e19103.

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

Jensen, K.T., Kao, T.-C., Stone, J.T., and Hennequin, G. (2021). Scalable Bayesian GPFA with automatic relevance determination and discrete noise models. BioRxiv 2021.06.03.446788.

Kao, T.-C., Jensen, K.T., Bernacchia, A., and Hennequin, G. (2021a). Natural continual learning: success is a journey, not (just) a destination. ArXiv:2106.08085 [Cs, q-Bio].

Kao, T.-C., Sadabadi, M.S., and Hennequin, G. (2021b). Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. Neuron *109*, 1567-1581.e12.

Kaplanis, C., Shanahan, M., and Clopath, C. (2018). Continual Reinforcement Learning with Complex Synapses. ArXiv:1802.07239 [Cs].

Kargo, W.J., and Nitz, D.A. (2004). Improvements in the Signal-to-Noise Ratio of Motor Cortex Cells Distinguish Early versus Late Phases of Motor Skill Learning. J. Neurosci. *24*, 5560–5569.

Katlowitz, K.A., Picardo, M.A., and Long, M.A. (2018). Stable Sequential Activity Underlying the Maintenance of a Precisely Executed Skilled Behavior. Neuron *98*, 1133-1140.e3.

Kawai, R., Markman, T., Poddar, R., Ko, R., Fantana, A.L., Dhawale, A.K., Kampff, A.R., and Ölveczky, B.P. (2015). Motor Cortex Is Required for Learning but Not for Executing a Motor Skill. Neuron *86*, 800–812.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [Cs].

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA *114*, 3521–3526.

Kleinrok, Z., and Turski, L. (1980). Kainic acid-induced wet dog shakes in rats. Naunyn-Schmiedeberg's Arch. Pharmacol. *314*, 37–46.

Krakauer, J.W., and Shadmehr, R. (2006). Consolidation of motor memory. Trends in Neurosciences *29*, 58–64.

Kubota, Y., Liu, J., Hu, D., DeCoteau, W.E., Eden, U.T., Smith, A.C., and Graybiel, A.M. (2009). Stable Encoding of Task Structure Coexists With Flexible Coding of Task Events in Sensorimotor Striatum. Journal of Neurophysiology *102*, 2142–2160.

Laje, R., and Buonomano, D.V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. Nat Neurosci *16*, 925–933.

Liberti, W.A., Markowitz, J.E., Perkins, L.N., Liberti, D.C., Leman, D.P., Guitchounts, G., Velho, T., Kotton, D.N., Lois, C., and Gardner, T.J. (2016). Unstable neurons underlie a stable learned behavior. Nature Neuroscience *19*, 1665–1671.

Lütcke, H., Margolis, D.J., and Helmchen, F. (2013). Steady or changing? Long-term monitoring of neuronal population activity. Trends in Neurosciences *36*, 375–384.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Marder, E., and Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. Nature Reviews Neuroscience *7*, 563–574.

Margolis, D.J., Lütcke, H., Schulz, K., Haiss, F., Weber, B., Kügler, S., Hasan, M.T., and Helmchen, F. (2012). Reorganization of cortical population activity imaged throughout long-term sensory deprivation. Nature Neuroscience *15*, 1539–1546.

Marinescu, I.E., Lawlor, P.N., and Kording, K.P. (2018). Quasi-experimental causality in neuroscience and behavioural research. Nat Hum Behav *2*, 891–898.

Marshall, J.D., Aldarondo, D.E., Dunn, T.W., Wang, W.L., Berman, G.J., and Ölveczky, B.P. (2021). Continuous Whole-Body 3D Kinematic Recordings across the Rodent Behavioral Repertoire. Neuron *109*, 420-437.e8.

Martin, W.R., Wikler, A., Eades, C.G., and Pescor, F.T. (1963). Tolerance to and physical dependence on morphine in rats. Psychopharmacologia *4*, 247–260.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci *21*, 1281–1289.

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*, 60–69.

Mrsic-Flogel, T.D., Hofer, S.B., Ohki, K., Reid, R.C., Bonhoeffer, T., and Hübener, M. (2007). Homeostatic Regulation of Eye-Specific Responses in Visual Cortex during Ocular Dominance Plasticity. Neuron *54*, 961–972.

Murray, J.M., and Escola, G.S. (2020). Remembrance of things practiced with fast and slow learning in cortical and subcortical pathways. Nat Commun *11*, 6441.

Musall, S., Kaufman, M.T., Juavinett, A.L., Gluf, S., and Churchland, A.K. (2019). Single-trial neural dynamics are dominated by richly varied movements. Nat Neurosci *22*, 1677–1686.

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

Park, S.-W., and Sternad, D. (2015). Robust retention of individual sensorimotor skill after self-guided practice. Journal of Neurophysiology *113*, 2635–2645.

Park, S.-W., Dijkstra, T., and Sternad, D. (2013). Learning to never forget—time scales and specificity of long-term memory of a motor skill. Front. Comput. Neurosci. *7*.

Peters, A.J., Chen, S.X., and Komiyama, T. (2014). Emergence of reproducible spatiotemporal activity during motor learning. Nature *510*, 263–267.

Peters, A.J., Lee, J., Hedrick, N.G., O'Neil, K., and Komiyama, T. (2017). Reorganization of corticospinal output during motor learning. Nature Neuroscience *20*, 1133–1141.

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

Poddar, R., Kawai, R., and Ölveczky, B.P. (2013). A Fully Automated High-Throughput Training System for Rodents. PLOS ONE *8*, e83171.

Qin, S., Farashahi, S., Lipshutz, D., Sengupta, A.M., Chklovskii, D.B., and Pehlevan, C. (2021). Coordinated drift of receptive fields during noisy representation learning. BioRxiv 2021.08.30.458264.

Rodriguez, A., and Laio, A. (2014). Clustering by fast search and find of density peaks. Science *344*, 1492–1496.

Rokni, U., Richardson, A.G., Bizzi, E., and Seung, H.S. (2007). Motor learning with unstable neural representations. Neuron *54*, 653–666.

Roxin, A., and Fusi, S. (2013). Efficient Partitioning of Memory Systems and Its Importance for Memory Consolidation. PLOS Computational Biology *9*, e1003146.

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

Rule, M.E., Loback, A.R., Raman, D.V., Driscoll, L.N., Harvey, C.D., and O'Leary, T. (2020). Stable task information from an unstable neural population. ELife *9*, e51121.

Schoonover, C.E., Ohashi, S.N., Axel, R., and Fink, A.J.P. (2021). Representational drift in primary olfactory cortex. Nature *594*, 541–546.

Sheng, M., Lu, D., Shen, Z., and Poo, M. (2019). Emergence of stable striatal D1R and D2R neuronal ensembles with distinct firing sequence during motor learning. PNAS *116*, 11038–11047.

Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical Control of Arm Movements: A Dynamical Systems Perspective. Annu. Rev. Neurosci. *36*, 337–359.

Sizemore, M., and Perkel, D.J. (2011). Premotor synaptic plasticity limited to the critical period for song learning. PNAS *108*, 17492–17497.

Srinivasan, M., and Ruina, A. (2006). Computer optimization of a minimal biped model discovers walking and running. Nature *439*, 72–75.

Steinmetz, N.A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science *372*, eabf4588.

Stevenson, I.H., Cherian, A., London, B.M., Sachs, N.A., Lindberg, E., Reimer, J., Slutzky, M.W., Hatsopoulos, N.G., Miller, L.E., and Kording, K.P. (2011). Statistical assessment of the stability of neural movement representations. Journal of Neurophysiology *106*, 764–774.

Sussillo, D., and Abbott, L.F. (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. Neuron *63*, 544–557.

Taxidis, J., Pnevmatikakis, E.A., Dorian, C.C., Mylavarapu, A.L., Arora, J.S., Samadian, K.D., Hoffberg, E.A., and Golshani, P. (2020). Differential Emergence and Stability of Sensory and Temporal Representations in Context-Specific Hippocampal Sequences. Neuron *108*, 984-998.e9.

Todorov, E., and Jordan, M.I. (2002). Optimal feedback control as a theory of motor coordination. Nature Neuroscience *5*, 1226–1235.

Vyas, S., Golub, M.D., Sussillo, D., and Shenoy, K.V. (2020). Computation Through Neural Population Dynamics. Annual Review of Neuroscience *43*, 249–275.

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., et al. (2020). Discovering Precise Temporal Patterns in Large-Scale Neural Recordings through Robust and Interpretable Time Warping. Neuron *105*, 246-259.e8.

Wolff, S.B.E., Ko, R., and Ölveczky, B.P. (2019). Distinct roles for motor cortical and thalamic inputs to striatum during motor learning and execution. BioRxiv 825810.

Wolpert, D.M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nat Neurosci *3*, 1212–1217.

Wu, Y.K., Hengen, K.B., Turrigiano, G.G., and Gjorgjieva, J. (2020). Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. PNAS *117*, 24514–24525.

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

Yang, G., Pan, F., and Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. Nature *462*, 920–924.

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 in (Dhawale et al., 2017). Data from all animals have previously been used in (Dhawale et al., 2021). Projection neurons and interneurons were clustered on the basis of spike durations as previously described (Figure S4) (Barthó et al., 2004; Dhawale et al., 2017). 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 smoothing 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 and 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 motor cortex 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 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 convolving the concatenated spike times across trials 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. PETHs were normalized by z-score 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 6D, after first averaging over all PETH pairs separated by the same time difference for each individual unit.

To compute the similarity of firing rates (Figure S2) between two sessions with numbers of trials t_1 and t_2 and spike counts s_1 and s_2, we considered the probability that the difference in firing rates would be greater than the observed difference $\tilde{\delta}_r = r_1 - r_2 = \frac{s_1}{t_1} - \frac{s_2}{t_2}$ under a Poisson model. We first fitted the Poisson model to the two sessions by setting the rate parameter per trial (λ) to the total firing rate $\lambda = \underset{\lambda}{\operatorname{argmax}}[\log p(s_1, s_2 | t_1, t_2, \lambda)] = \frac{s_1+s_2}{t_1+t_2}$. We then numerically estimated

$$p\left(\delta_r \geq \tilde{\delta}_r | t_1, t_2, \lambda\right) = \sum_{s_1=0}^{\infty} p(s_1|t_1, \lambda) \left[\sum_{s_2=t_2\left(\frac{s_1}{t_1}+\tilde{\delta}_r\right)}^{\infty} p(s_2|t_2, \lambda) + \sum_{s_2=0}^{t_2\left(\frac{s_1}{t_1}-\tilde{\delta}_r\right)} p(s_2|t_2, \lambda) \right]$$

This analysis was repeated for every pair of sessions, and the similarity was plotted as a function of time between sessions as for the PETH-based similarity measure.

Estimation of stability indices

The Pearson correlation ρ between PETHs as a function of time difference δt between PETHs was fitted for each neuron recorded for at least 4 days using an exponential model $\hat{\rho} = \beta e^{\alpha\,\delta t}$. 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, and only neurons recorded on at least 4 separate days were included in the analyses. Furthermore, we only included neurons with an intercept of $\beta \geq 0.2$ in our analyses, since neurons with no correlation on consecutive days will trivially not show a decrease in correlation over time and thus appear 'stable' despite not having task-modulated activity. The mean error of the model fit was quantified for each neuron as $\frac{1}{N}\sum_{i=1}^{N}|\rho_i - \hat{\rho}_i|$, 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 1,000 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. When comparing two distributions (Figure S4), significance was calculated with a two-sided t-test.

Stability as a function of recording duration in subsampled data

To investigate how the apparent stability of a neuron depended on the recording duration, we used the neurons recorded for at least 14 days in both motor cortex and DLS. We then considered progressively longer recording durations ranging from 4 days to 13 days and subsampled data from each neuron in every consecutive window of the corresponding recording duration. We used this subsampled data to compute a subsampled stability index and considered the mean stability index for each recording duration for further analysis. Finally, we z-scored the apparent stability index across recording durations and averaged the result across all units from each brain region to consider how the apparent stability depended on recording duration across neurons.

Stability as a function of recording duration across neurons

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 $\tilde{\alpha} = -a - b\exp(-c\,T)$. We fitted the model by minimizing the L1 error between the observations and model fit across neurons $\mathcal{L} = \sum_n |\alpha_n - \tilde{\alpha}_n|$ and restricted all parameters $\{a, b, c\}$ to be positive. In this model, the asymptotic stability is given by $\tau_\infty = \lim_{T\to\infty} -\tilde{\alpha}^{-1} = a^{-1}$. 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 4G by plotting the median stability index across neurons binned by recording duration.

Behavioral similarity

To compute behavioral similarity as a function of time difference, we first extracted instantaneous velocities 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.

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 motor cortex. As a control, we permuted the behavioral data across days to break any correlations between the neural and behavioral drift rates and repeated the analysis. In Figure 5D, 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.

Computational 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 in (Gallego et al., 2020).

Interpolating networks

To interpolate the networks in Figure 2C & 2D, 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 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. The seven interpolated parameter sets were each interpreted as corresponding to a different day of recording when computing stability indices, in correspondence with the experimental data. 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 2D are reported as the mean and standard deviation across these repetitions. The RNN stability indices used for comparison with the experimental data were computed as the mean across repetitions of the median stability index across neurons.

Stability as a function of recording duration in synthetic data

To assess how neural stability depended on recording duration in a synthetic dataset (Figure 4E), we first generated 4 kinematic output features by drawing samples from a Gaussian process with a timescale of 100 ms. We then generated neural activity on each trial by adding noise to the behavior and drawing $\boldsymbol{y} \sim Poisson(\boldsymbol{\lambda}\, \boldsymbol{W}\, \widetilde{\boldsymbol{b}})$, where $\boldsymbol{W} \sim N(0, \boldsymbol{I})$ is a weight matrix drawn randomly at the beginning of the simulation and $\widetilde{\boldsymbol{b}}$ is the noisy behavior. $\boldsymbol{\lambda}$ is a vector of parameters used to scale the mean firing rate of each neuron to either 1.75 Hz, 3.75 Hz or 10 Hz for the three analyses in Figure 4E. Importantly, the noise was drawn identically and independently on each trial such that there was no systematic drift in the generated behavior or neural activity. We simulated 269 synthetic trials on each 'day' of recording to match the mean number of trials per day in the experimental data. We then computed the stability index for each neuron by fitting an exponential model to the PETH correlation as a function of time difference as described for the experimental data. This procedure was repeated for simulated recording durations ranging from 4 days to 14 days, and median stability indices across neurons were averaged over 10 independent simulations for the results reported in Figure 4E. Finally, exponential models of the form $\tilde{\alpha} =$

$-a - b\exp(-c\,T)$ were fitted to the averaged data, similar to the procedure used for the experimental data.

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 5D), 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 $\widetilde{\boldsymbol{y}}_{t>1} \sim Poisson(\exp[\boldsymbol{W}\boldsymbol{x}_{t>1}])$ 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 data since this was used to fit the GLM.

Supplementary Figures

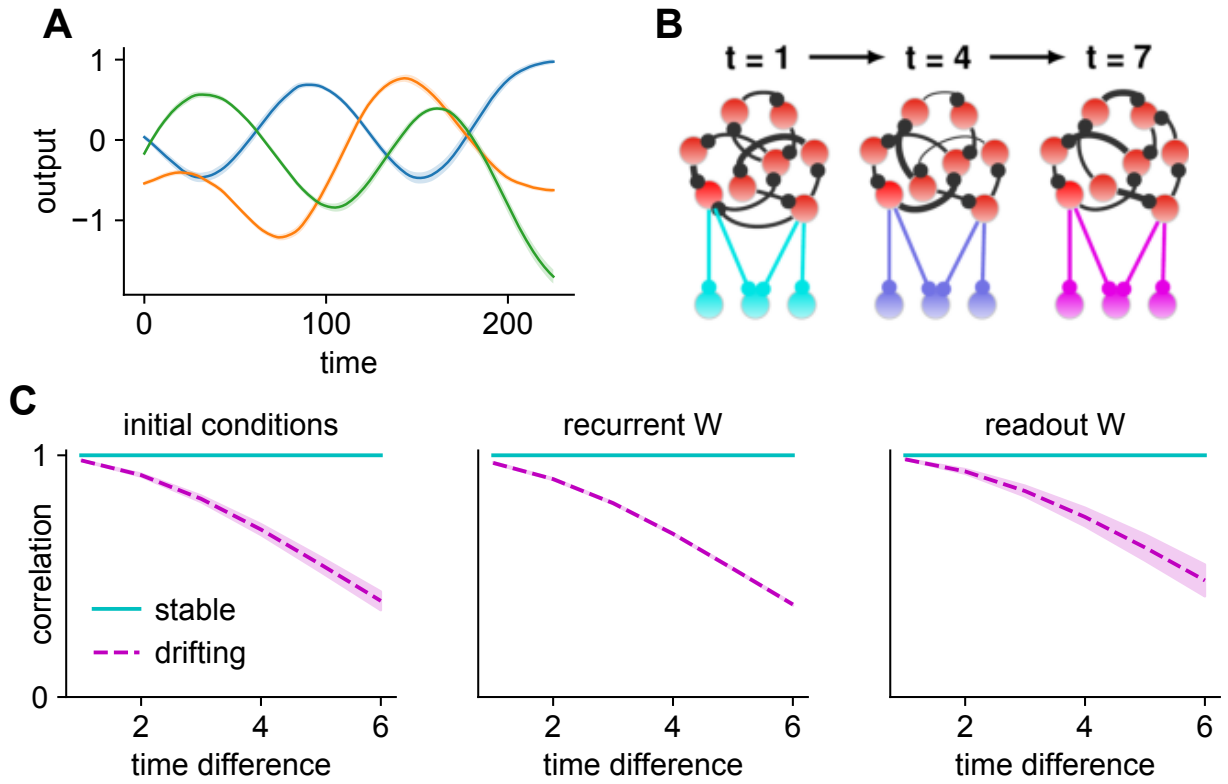

Figure S1: RNN output and interpolation. (A) Mean and standard deviation (shading) of the RNN output across 100 trials for three example output units. The target outputs were drawn from a Gaussian process (Methods). **(B)** Schematic illustrating the RNN interpolation procedure. Two RNNs were trained independently to produce the same output (A). All parameters were then linearly interpolated between the two networks, and the readout were weights frozen for each interpolated network while the recurrent weights and initial conditions were re-optimized to ensure stable performance. We let each interpolated network correspond to a single 'day' of recording and considered 7 networks corresponding to a 'week' of simulated recordings. These networks uniformly tiled the range from the original network to 70% towards the independently trained network, at which point neural activity was largely uncorrelated with the original activity (Figure 2D). This resulted in a series of RNNs that transitioned from the original connectivity to a network with parameters that had only little correlation with the original RNN (C). **(C)** 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.

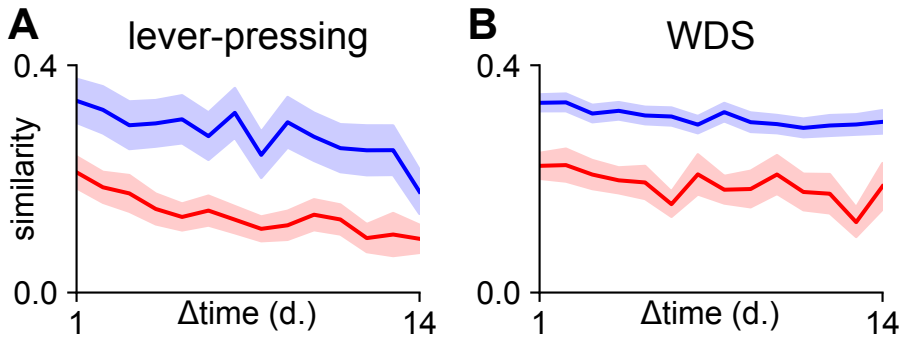

Figure S2: Stability of firing rates over time. We computed the similarity in firing rates between two sessions as the probability under a Poisson model of observing a difference in firing rates greater than the observed difference (Methods). We then computed the similarity as a function of time difference across neurons recorded for at least 14 days from motor cortex (red) and DLS (blue) in either the lever-pressing task **(A)** or the wet dog shake behavior **(B)**.

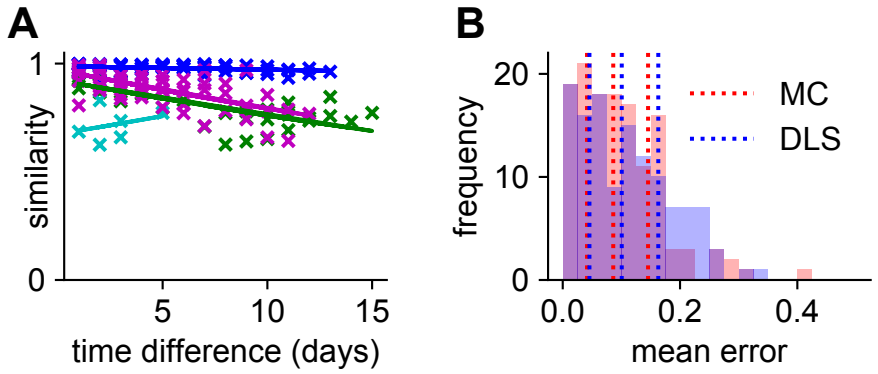

Figure S3: Exponential model fits. (A) Plots of PETH similarity against time difference for four example units (colors) together with exponential fits illustrating a range of different decay rates, baseline similarities and durations of recording. **(B)** Distribution of the mean error of each model fit across the population of neurons recorded from MC (red) or DLS (blue). Vertical dashed lines indicate quartiles of the distributions.

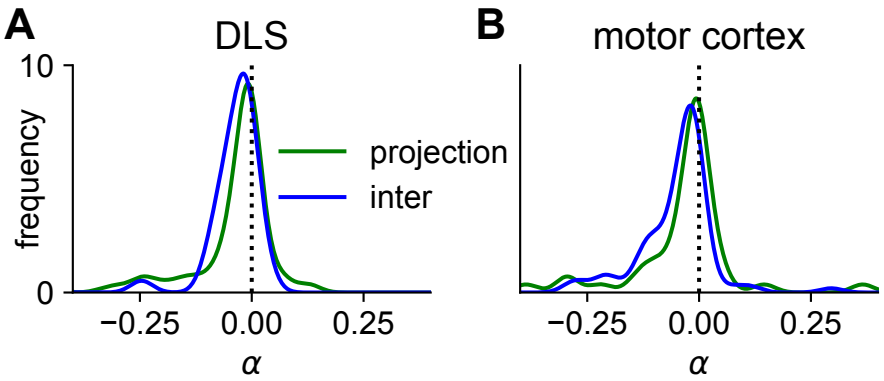

Figure S4: Comparison of interneurons and projection neurons. (A) Density plot of the stability indices for projection neurons and interneurons recorded from DLS (p = 0.29 for the two populations having different means, two-sided t-test; n = [94, 31] neurons). **(B)** As in (A) for recordings from motor cortex (p = 0.11, two-sided t-test; n = [48, 86] neurons). Note the difference from previous observations in songbird HVC where interneurons recorded using electrophysiology were found to exhibit more stable firing patterns than projection neurons recorded using calcium imaging (Liberti et al., 2016). Since we did not observe a significant difference in the stability of interneurons and projection neurons, we combined both classes of neurons for all analyses.

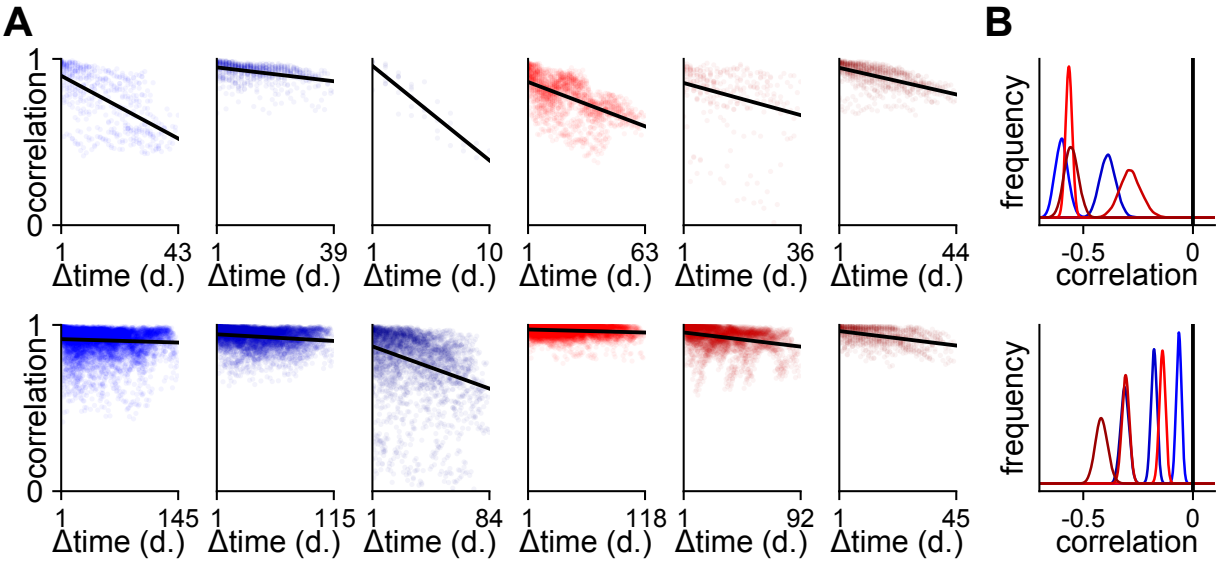

Figure S5: Behavioral drift across animals. (A) Scatter plots of the correlation between mean velocity profiles and the time between days for all pairs of days in each animal. Top row: lever-pressing task; bottom row: wet dog shakes. Blue indicates animals with recordings from DLS, red from motor cortex. **(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).

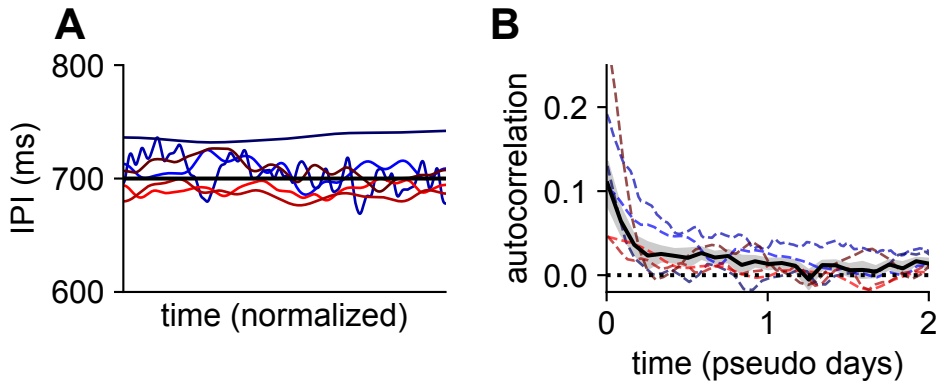

Figure S6: Inter-press intervals and autocorrelations. (A) Inter-press interval (IPI) for each animal convolved with a 200-trial Gaussian filter. Time is normalized from 0 to 1 for each animal (n = 9365 ± 6886 trials). Black horizontal line indicates 700 ms. **(B)** We computed the IPI autocorrelation 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 (Figure S5). This suggests that behavioral changes are predominantly along 'task-null' directions that do not affect performance.

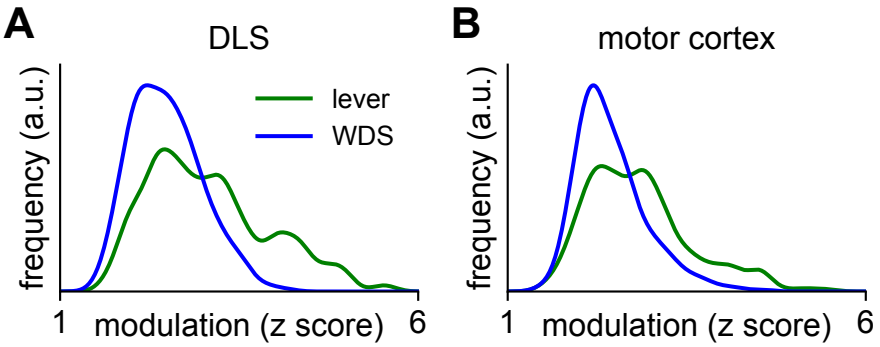

Figure S7: Task-modulation of neurons in the lever-pressing task and wet-dog shake behavior. (A) A PETH was computed across all trials for each neuron in 20 ms bins, and the time bin identified with the maximum deviation from the mean across all time bins. The corresponding z-score was computed, and 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 motor cortex.