Imperial College London
Department of Bioengineering

Neural population dynamics in motor control and learning

Joanna Chen Hsuan Chang

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
at Imperial College London, December 2025

I hereby declare that to the best of my knowledge all work presented in this thesis was performed by me or was referenced otherwise.

Joanna Chen Hsuan Chang

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND). Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Animals have a remarkable ability to learn and control a wide range of movements. This ability is crucial not only for the survival of an individual over its lifetime, but also for the survival of its species over evolutionary timescales. How the brain achieves both the stability required for precise control and the flexibility required for learning new skills remains an open question. Recent work in motor neuroscience has found that neural population activity is largely confined to a low-dimensional subspace called the neural manifold. Examining the time-varying activity in this space—the population latent dynamics—has provided many insights into the neural underpinnings of movement. In this thesis, we examine the role of latent dynamics in motor control and learning in three parts.

First, we explore the stability of latent dynamics in motor control by asking how different individuals of the same species are able to generate similar behaviors. We hypothesize that evolutionarily conserved latent dynamics allow animals to reliably produce the same adaptive behaviors. Through a combination of experimental recordings and computational modeling, we demonstrate that latent dynamics are consistently stable and preserved across multiple motor domains. Second, we investigate the flexibility of latent dynamics during motor learning. Using recurrent neural networks to model motor cortical dynamics, we examine how population activity acquired through *de novo* skill learning influences the flexibility of subsequent motor adaptation. Third, we explore how animals can flexibly learn new motor tasks without disrupting stable motor control. By modeling a sequential learning experiment, we investigate neural signatures of motor memories and their relationship to continual learning.

Overall, our work establishes the importance of latent dynamics in motor control and motor learning, offering new perspectives on how the brain balances stability and flexibility for behavior, across an animal's lifetime and over the course of a species' evolution.

Acknowledgements

Some say that having the right PhD supervisor is the most important part of a PhD. It would have been a blessing to have one good supervisor, but I hit the jackpot with two. To Claudia and Juan, thank you for being the dream team. Claudia, thank you for your constant support and for giving me the freedom to go on side quests—I can always trust that you'll guide me back on the right path if necessary. You have been a role model in academia and in life. Juan, thank you for the hours-long brainstorming sessions, the food and sci-fi recommendations, and the chats about life. You have always wanted and fought for the best for me, and that has made all the difference. It has been a privilege to learn from both of you on how to do science and how to be kind.

To my collaborators: Matt, Lee, Junchol, and Josh, thank you for your insights and for trusting us with your datasets. Mostafa, thank you for being a great mentor but above all, a dear friend. Thank you all for helping me grow as a scientist.

To my labmates: you have been the best part of this PhD. Bence, Cátia, Cecilia, Ciara, Colleen, Elena, Francesca, Genji, Grace, Jimmie, Klara, Martín, Mostafa, Natalia, Tudor, and many more, thank you for this wild ride. To the plus-ones who I always forget are not actually in the lab—João, Josu, and Viki, thank you for joining along. I know how special it is to have labmates who also become some of your favorite people. Thank you for all the Friday pubs, the Sunday picnics, the board game nights, and everything in between. I have explored London and traveled the world with you. I can't imagine a better group of people to have shared the last four years with.

To my friends: Dani, Jonny, and Mingke, thank you for keeping me sane through Covid lockdowns, for all the Chinese New Year and Chrismas dinners. It has been a pleasure to see us all grow from that horrible little London flat. To Max and Ruth, thank you for being my first friends in London and showing me all the best parts of being British. Thank you all for helping London feel like home. To Allison, thank you for cheering me on from the very beginning. You have been both an anchor and a safe harbor. To Bailey, Brendan, and Hayat, thank you for being with me through all of the ups and downs. You know how much you mean to me.

To my big sister Jackie, thank you for being my biggest supporter, for embarassing me by telling other people how proud you are. I learned to be bold, creative, and determined from you. Thank you for being a reminder of how far we've come.

To my partner and favorite person Andrea, thank you for everything. You were there for every poster, every talk, every breakdown, every celebration. The PhD was hard but you made everything so much easier and brighter. Thank you for all the laughter.

Contents

Abstract **3**

Acknowledgements **3**

1 Introduction **9**
1.1 Motor control . . . 9
1.2 Computational approaches . . . 13
1.3 Overview . . . 17

2 Preserved latent dynamics in motor control **19**
2.1 Abstract . . . 20
2.2 Introduction . . . 21
2.3 Results . . . 22
2.4 Discussion . . . 32
2.5 Methods . . . 34
2.6 Supplementary Figures . . . 51

3 Structured latent dynamics in motor learning **64**
3.1 Abstract . . . 65
3.2 Introduction . . . 66
3.3 Results . . . 69
3.4 Discussion . . . 85
3.5 Methods . . . 90
3.6 Supplementary Figures . . . 102

4 Motor memories in sequential learning **116**
4.1 Abstract . . . 117
4.2 Introduction . . . 118
4.3 Results . . . 121
4.4 Discussion . . . 134
4.5 Conclusion . . . 138
4.6 Methods . . . 139
4.7 Supplementary Figures . . . 149

5 Conclusion **161**
5.1 Summary . . . 161
5.2 Limitations and future work . . . 162
5.3 Conclusion . . . 167

Bibliography **168**

List of Figures

1.1 The motor system . . . 10
1.2 Neural manifold and population latent dynamics framework . . . 14

2.1 Hypothesis schematic . . . 22
2.2 Preserved latent dynamics across animals performing the same behaviour 24
2.3 Preserved latent dynamics in a task with higher behavioural complexity 26
2.4 Preserved latent dynamics extend to basal ganglia and covert behaviour 29
2.5 Similar behavioural output is necessary but not sufficient for the preservation of latent dynamics . . . 31
S2.1 Behavioural and neural data from monkeys performing a centre-out reaching task . . . 51
S2.2 A broad range of techniques uncover preserved latent dynamics that capture dominant features of neural population activity . . . 52
S2.3 Additional examples of canonical correlations, and preserved latent dynamics between pairs of sessions from two different monkeys and mice . 54
S2.4 Preserved latent dynamics across monkeys are behaviourally relevant . . 55
S2.5 Behaviour and neural activity for mice performing a grasping and pulling task . . . 56
S2.6 Additional data for mouse latent dynamics and across-animal decoders . 57
S2.7 Aligning based on the topological organisation of the neural data is not sufficient to uncover preserved latent dynamics across individuals 59
S2.8 Behavioural and neural data from monkeys performing a sequential reaching task, and additional data on the preservation of latent dynamics . . . 60
S2.9 The latent dynamics of two different monkeys performing the same task are more correlated than those of the same monkey performing two related tasks . . . 61
S2.10 Additional data: Behavioural similarity is necessary but not sufficient for the preservation of latent dynamics . . . 63

3.1 Probing *de novo* learning and adaptation with RNNs . . . 70
3.2 Networks that have acquired multiple movements produce more constrained and robust dynamics that are structured in neural space 73
3.3 Networks with larger repertoires can adapt to perturbations more easily 75
3.4 The structure in neural space of multi-movement networks resulting from *de novo* learning is responsible for their patterns in adaptation . . . 76
3.5 The structure in neural space arising from *de novo* learning can facilitate or hinder adaptation . . . 81
3.6 The observed joint influence of motor repertoire and *de novo* learning experience extends to a more complex sinusoidal task . . . 83

S3.1 Networks that can generate multiple movements produce more constrained neural dynamics . . . 102
S3.2 Networks reproduce experimental results . . . 103
S3.3 Trends in variance remain robust for synthetic movements and different measures of variance . . . 104
S3.4 Addition of noise shows underlying structure in multi-movement networks 106
S3.5 Networks can generalize and adapt to perturbations that require movements within a learned range . . . 107
S3.6 Adaptation training with stochastic gradient descent (SGD) is better at overcoming catastrophic forgetting than FORCE . . . 108
S3.7 Explicit preparatory epochs are not necessary for differences in adaptation and structure . . . 109
S3.8 Simulation results reflect experimental monkey data . . . 110
S3.9 Adaptation with fixed input or recurrent weights . . . 111
S3.10 Structure in neural space hinders adaptation when larger changes are required . . . 112
S3.11 Adaptation to visuomotor rotation perturbation is within-manifold . . . 113
S3.12 Main results remain robust to changes in modeling parameters . . . 114

4.1 Potential neural signatures of learning . . . 119
4.2 RNNs recapitulate experimental results in sequential learning . . . 122
4.3 Uniform shifts and memory traces arise from sequential learning without explicit mechanisms . . . 125
4.4 The magnitude of memory traces, but not that of uniform shifts, correlates with retention savings . . . 128
4.5 Experimentally hard-to-learn outside-manifold perturbations lead to larger memory traces and smaller uniform shifts . . . 129
4.6 Context cues alter uniform shifts and memory traces . . . 132
S4.1 Neural activity changes from Task A1 to Task A2 . . . 149
S4.2 Uniform shifts for different simulations . . . 150
S4.3 Memory traces consistently arise and are due to learning Map B . . . 152
S4.4 Relationship between memory trace and retention savings . . . 153
S4.5 The magnitude of memory traces and uniform shifts compared to different measures . . . 154
S4.6 Relationship between memory trace and uniform shift and different factors once initial MSE is fixed . . . 155
S4.7 Learning within- and outside-manifold maps under different training regimes . . . 156
S4.8 Larger memory traces for outside-manifold learning are not due to performance tradeoffs . . . 157
S4.9 Context disrupts the correlation between memory traces and savings . . . 158
S4.10 Manifold overlap for different context magnitudes . . . 159
S4.11 Memory traces and sustained uniform shifts arise when learning is confined to upstream of motor cortex . . . 160

Chapter 1

Introduction

> It is a misconception to think that during evolution humans sacrificed physical skill in exchange for intelligence: wielding one's body is a mental activity.
>
> *Ted Chiang*

1.1 Motor control

The ability to move defines how we explore the world and adapt to its ever-changing environments. From basic actions such as walking or reaching for an object to complex behaviors like speaking or playing an instrument, movement is central to our daily lives. We have an exceptional ability to excise precise, skilled control over a vast repertoire of movements. While we can slowly accumulate different motor skills throughout our lifetimes without significant interference, we can also quickly adapt existing skills to ongoing perturbations. This acquisition and control of motor skills is one of the most sophisticated functions of the brain, and understanding the underlying neural processes remains one of the biggest questions in neuroscience. Given the complexity of these processes, large populations of neurons must coordinate to facilitate learning and movement. This thesis combines experimental recordings and computational modeling to explore how the brain learns and executes movements, particularly through a neural population framework.

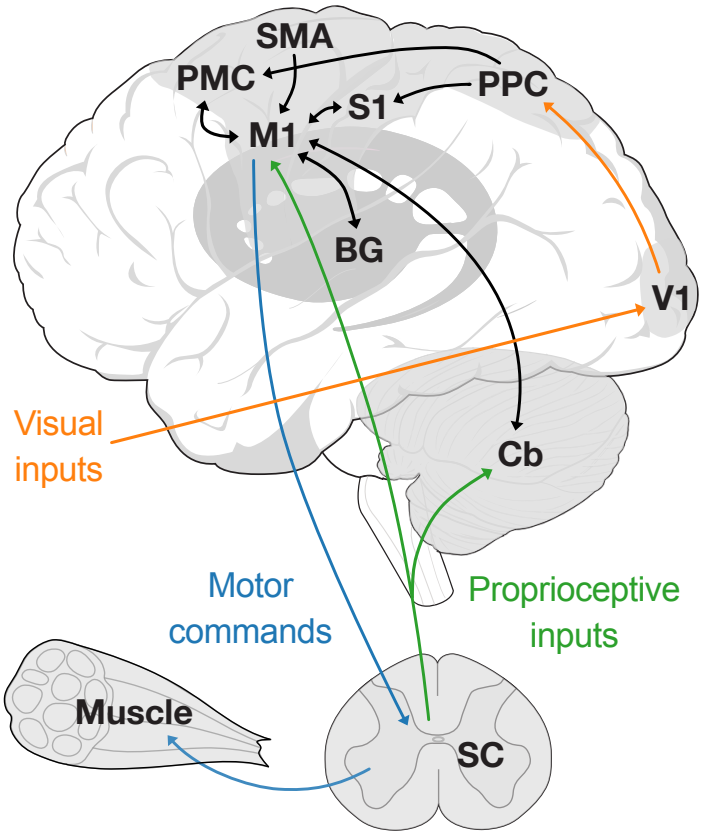


Figure 1.1: The motor system. The brain regions involved in motor control[1,2], including major inputs and outputs of the motor cortex. Primary motor cortex (M1), premotor cortex (PMC), supplementary motor area (SMA), somatosensory cortex (S1), posterior parietal cortex (PPC), visual cortex (V1), basal ganglia (BG), cerebellum (Cb), and spinal cord (SC).

Motor system

The motor cortex is widely studied for its role in producing voluntary movements. The primary motor cortex (M1), an area within the motor cortex, is the main region responsible for sending neural signals to motor neurons in the spinal cord, which ultimately control muscle activation (Figure 1.1). M1 is organized somatotopically, meaning that different areas of M1 correspond to specific body parts, with greater representation for areas like hands and figures that require finer control[3]. Electrical stimulation of specific points in M1 elicits movement in corresponding muscles[3], making M1 directly connected to motor output.

M1 does not act in isolation, however (Figure 1.1). It integrates inputs from other motor cortical areas, the dorsal premotor cortex (PMd) and the supplementary motor area, which are often implicated in movement planning. The motor cortex also receives sensory input via two key pathways. Visual input is received by the primary visual cortex, which sends signals to the posterior parietal cortex, which then sends signals to PMd. Proprioceptive feedback is received more directly via the primary somatosensory cortex, which projects to M1. Additionally, the motor cortex has crucial loops with several subcortical structures, notably the cerebellum and the dorsal striatum of the basal ganglia. The cerebellum has been found to play a critical role in motor adaptation[4–6], whereas the dorsal striatum is involved in controlling movement kinematics[7,8].

The following chapters will mainly focus on examining the motor cortex——in particular M1——since it is the region most directly responsible for motor output. Across Chapters 2 to 4, we will examine motor cortical neural populations through experimental studies in monkeys and mice in tandem with computational models. To generate movement, the motor cortex must also integrate inputs from other brain regions and external sensory cues with its internal neural dynamics. In Chapters 3 and 4, we explore how different inputs that provide task-specific and context-specific cues to the motor cortex may alter motor learning and adaptation.

Motor learning

Animals possess a remarkable capacity to learn and refine motor skills over both short and long timescales. Long-term *de novo* learning allows the development of new motor skills from scratch and requires many hours and hours of practice[9]. In contrast, short-term adaptation enables rapid adjustments of a familiar skill to temporary changes in the environment[9]. For example, after learning to walk, individuals can quickly adapt to walking on slippery surfaces in the rain or uneven terrain on a hike.

Long-term motor learning is associated with neural changes, often observed at the level of synaptic connections between neurons[10–12]. Modifications in synaptic connectivity

can reshape the neural activity patterns generated by a neural population[13]. In contrast, motor adaptation appears to occur with little to no synaptic changes in motor cortical populations[14,15]. Given that motor adaptation must be rapid and transient, experimental studies suggest that it may leverage pre-existing neural activity patterns rather than forming entirely new ones[16,17]. However, how adaptation builds upon prior neural activity patterns established through long-term learning remains an open question. While several psychophysical studies have investigated *de novo* learning in humans[18–20], relatively few have examined this process using electrophysiological recordings. Experimentally, *de novo* learning is challenging to study due to the difficulty of maintaining continuous neural recordings in animals throughout the learning process. An even greater challenge is assessing how motor repertoires developed through long-term learning influence short-term adaptation, as capturing and quantifying an animal's entire repertoire remains nearly impossible. To circumvent this experimental challenge, Chapter 3 explores the interplay between long-term learning and short-term adaptation using computational models.

While motor learning often leads to changes in neural activity, the brain is remarkably capable of acquiring numerous skills over time with minimal interference. In some cases, acquiring a new task can lead to forgetting of a familiar task (retrograde interference), while prior knowledge of a familiar task can sometimes hinder the learning of a new one (anterograde interference)[9]. These instances show that neural changes associated with different skills have the potential to conflict. Despite this, motor learning remains highly robust: individuals continually learn new skills, building rich motor repertoires throughout their lifetimes without catastrophic forgetting of old skills. This process, called continual learning, is a capability that artificial learning systems still struggle to replicate[21,22]. How the brain can accommodate multiple motor memories of different motor tasks is an area of ongoing research. In Chapter 4, we examine how two potential signatures of motor memories identified from experimental studies relate to continual learning.

1.2 Computational approaches

Neural manifolds

Classically, the neural control of movement has been mainly examined on the single-neuron level[23–25]. Movement parameters like muscle force and limb velocity were correlated with, and thought to be coded in, the activity of single neurons[24,26,27]. However, these correlations were often inconsistent, with supposed relationships to movement parameters disappearing under varying conditions such as changes in posture or external load[28]. Additionally, many neurons showed no correlation with any movement parameters at all[28]. Together, these findings suggested that studies on single neurons were not sufficient to explain the neural basis of movement[28–31].

Rather, focus has shifted to using a neural population view to provide insight into the neural underpinnings of movement. Recent advancements in electrophysiology have made this possible by allowing simultaneous recordings from hundreds to thousands to neurons[32]. In this neural population view, the computations for movement are generated through the coordinated activity of a network of interconnected neurons (Figure 1.2). Since the neurons are connected, the activity of one neuron is correlated with the activity of other neurons. As a result, the underlying connectivity of the network constrains the possible patterns of activity that can be produced to a lower-dimensional subspace called the neural manifold[13] (Figure 1.2). Mathematically, this neural manifold can be estimated by applying dimensionality reduction methods[33,34] to extract the dominant patterns or axes of neural co-variation—the neural modes—that span the manifold (Figure 1.2). The time-dependent activation of these patterns are then the latent dynamics of the network.

Examining population latent dynamics through this manifold framework has shed light on many aspects of motor control and motor learning. Latent dynamics have been able to provide insight on neural principles underlying various domains ranging across covert[35–37] and overt behavior[38,39], sensory[40,41] and cognitive integration[42], and behavioral stability[43]

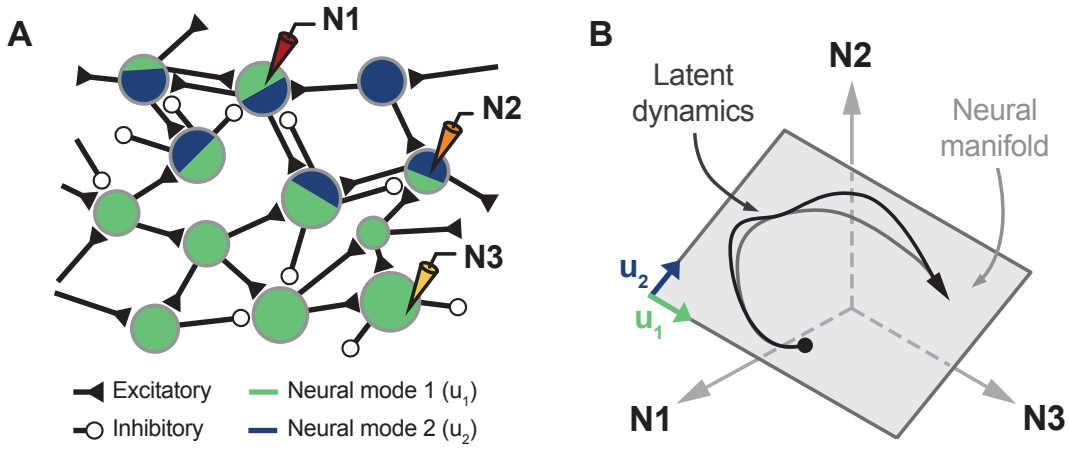


Figure 1.2: Neural manifold and population latent dynamics framework. **A**) Neurons in a network are synaptically connected to each other. The underlying connectivity determines the main activity patterns (neural modes) that the network can create. **B**) The neural activity space of a neural population can be geometrically represented with each axis corresponding to the activity of one neuron. The population activity is restricted to a lower-dimensional subspace called the neural manifold. The time-varying population activity, the population latent dynamics, is mostly confined to the manifold. Figure adapted from Gallego et al[13].

and flexibility[44,45]. Learning also seems to be constrained by neural manifolds and their dynamics[16,17,46]. The ability for latent dynamics to capture so many functional processes suggest that they are not only an important tool for scientists to understand behavior, but may also be an important tool for the brain to generate behavior. In Chapter 2, we explore how latent dynamics are preserved to generate similar behavior across different individuals. In Chapters 3 and 4, we use latent dynamics to explore motor skill learning and adaptation.

Brain-computer interfaces

Brain-computer interfaces (BCIs) also provide a powerful tool for probing the neural principles of motor control and learning. Unlike natural movement, where neural activity drives muscle activation, BCIs establish a direct link between neural signals and external devices, such as prosthetic limbs or computer cursors. While BCIs exist in various domains, those designed for motor control typically record from neurons in the motor cortex and decode parameters like cursor velocity from ongoing neural activity[46–48]. In

clinical applications, the decoders are constructed to closely align output parameters with user intent. This enables individuals with paralysis to exert fine control over a cursor, for example. In research, however, the decoders can be manipulated to study the neural bases of movement. Since the decoders provide a complete mapping between neural activity and output, modifying this mapping allows researchers to test different hypotheses and uncover fundamental neural principles.

For instance, a seminal study explored motor learning by perturbing BCI maps to require new activity patterns either within or outside the existing neural manifold[46]. Learning to control the new within-manifold maps could be accomplished within a matter of minutes or hours in a single session[46], whereas learning to control outside-manifold maps required multiple sessions over several days[16]. These experiments demonstrated that the underlying neural manifold constrains the set of activity patterns that can be readily learned and generated. In Chapter 4, we leverage this BCI paradigm to explore how learning new tasks changes the neural activity for a familiar task during sequential learning.

Neural network modelingNeural network modeling

Computational models provide additional insight by modeling the activity of neural populations. Artificial neural networks, particularly recurrent neural networks (RNNs), have been shown to reproduce key features of neural population dynamics found in experimental neural recordings[49–51]. These similarities provide evidence that RNNs can be used to investigate and make predictions about neural activity, particularly under scenarios that are difficult to probe experimentally.

Commonly, rate-based RNN models are used to model the dynamics of a neural population using the equation

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \mathbf{W^{rec}}\mathbf{r} + \mathbf{W^{in}}\mathbf{s},$$

where $\mathbf{r}$ represents the firing rates of the neurons following some nonlinear activation of the hidden states $\mathbf{x}$. The connectivity of the network is denoted by $\mathbf{W^{rec}}$, where each

nonzero entry W_{ij}^{rec} is a synaptic connection between neuron i and neuron j. In addition to the contributions from other neurons within the network, neural activity is also influenced by different inputs $\mathbf{s}$ into the network, with input weights $\mathbf{W^{in}}$. As previously mentioned, these inputs can come from sensory or cognitive signals via upstream brain regions, converying information such as environmental stimuli or task-specific context. Thus, these models assume that the dynamics of a neural population arise from both intrinsic origins, determined by the recurrent connectivity $\mathbf{W^{rec}}$, and extrinsic factors, determined by the input weights $\mathbf{W^{in}}$. Importantly, the recurrent connectivity of RNNs allows them to generate intrinsic dynamics that evolve over time, a feature that distinguishes them from feedforward networks that can only generate input-driven responses. This makes RNNs particularly well-suited for modeling regions like motor cortex, whose neural activity is affected by external inputs but also largely driven by internally generated dynamics[36,52,53].

The generated RNN activity can then be mapped to motor output using the equation

$$\mathbf{z} = \mathbf{W^{out}}\mathbf{r},$$

where $\mathbf{z}$ can be different movement parameters like position or velocity. Similar to BCIs, the output weights $\mathbf{W^{out}}$ provide the exact mapping between firing rates and motor output, and changing this mapping can be used to model different BCI maps.

By fully capturing the relationships between external inputs, internal dynamics, and motor output, these models provide a controlled framework for studying neural population dynamics and their resulting behavior. They enable simulations of learning scenarios that would be challenging or impossible to perform experimentally due to biological and temporal constraints. These models allow us to precisely control and quantify different learning experiences and motor repertoires, which is necessary to examine the effects of *de novo* learning on motor adaptation in Chapter 3. They also allow us to clone the same population of neurons and explore how they respond under different learning conditions, as we do in Chapters 1 through 3. These computational models thus serve as a key tool throughout this thesis to investigate the role of neural population dynamics in motor

control and motor learning.

As with any computational model, the RNNs models used in this thesis are simplified abstractions of the complex neural circuitry and processes that underlie motor control and motor learning. While beyond the scope of this thesis, it is important to note several biological mechanisms that are either not captured or largely simplified by these models. Complex feedback loops with other brain regions, peripheral circuits, and muscles are simplified as direct or feedback inputs into the RNNs. Aspects of biological learning, such as the role of the cerebellum in adaptation, memory consolidation during sleep, or discontinuous improvements through cognitive understanding, are also not captured by these models. Nevertheless, these models can capture general properties of how interconnected neurons generate coordinated dynamics, making them useful tools for probing population-level phenomena.

1.3 Overview

The purpose of this thesis is to understand how neural population dynamics contribute to motor control and motor learning, with a focus on the stability and flexibility of the dynamics. The main part of the thesis is divided into three self-contained chapters, where each chapter is a published or preprinted manuscript.

In Chapter 2, we assess the role that population latent dynamics play in motor control. We address the question of how different individuals of the same species are able to stably produce similar behaviors despite differences in specific brain circuitry. Since the overall organization across brains is maintained, we explore the hypothesis that evolution imposes constraints that lead to preserved latent dynamics across animals to generate the same behavior. Using experimental recordings from mice and monkeys across several tasks along with neural network models, we show that the latent dynamics are largely preserved and behaviorally relevant for multiple behaviors, in the motor cortex and striatum.

After establishing the importance of population latent dynamics in motor control, we

examine how they can help us understand motor learning. While latent dynamics are generally stable across animals for motor control, they still need to be flexible within animals for motor learning. In Chapter 3, we address the question of how long-term *de novo* learning affects short-term motor adaptation. Using RNNs to model motor cortical population activity, we demonstrate how different *de novo* learning experiences influence the underlying structure of the latent dynamics. The structure of the dynamics then determines the flexibility to adapt to different perturbations.

While learning flexibly causes changes in the latent dynamics, animals can still learn many tasks sequentially without interference, leading to stable motor control. In Chapter 4, we address the question of how learning new tasks alters the population activity of familiar tasks without catastrophic forgetting. Using RNNs to model a BCI paradigm, we examine two signatures of motor memories identified from experiments and show how they relate to continual learning and stable behavior.

Chapter 2

Preserved latent dynamics in motor control

We start by exploring how different animals of the same species are able to produce similar behaviors. From salmon swimming upstream to monkeys picking fruit from a tree, animals have a remarkable ability to generate behaviors that are advantageous to their specific environments. In this chapter, we hypothesize that this stability in motor control across animals results from evolutionarily conserved latent dynamics. Using neural population recordings and neural network models, we demonstrate that latent dynamics are indeed preserved across animals, and this preservation is consistently observed in multiple domains.

The work in this chapter is presented in an article format as published in *Nature* in 2023[54]. I would like to thank Mostafa Safaie, the co-first author on this work. Dr. Safaie and I collaborated on analyzing the data and generating the figures. Dr. Safaie performed most of the analyses on the monkey center out task and the mouse reaching and pulling task. I performed the analyses on the monkey sequential reaching task, and trained and analyzed the decoders and neural network models. I would also like to thank Juan A. Gallego for his supervision and Matthew G. Perich for his guidance and inspiration. Furthermore, I would like to thank our experimental collaborators Junchol Park, Lee E. Miller, and Joshua T. Dudman for providing their data and insight.

Preserved neural dynamics across animals performing similar behaviour

2.1 Abstract

Animals of the same species exhibit similar behaviours that are advantageously adapted to their body and environment. These behaviours are shaped at the species level by selection pressures over evolutionary timescales. Yet, it remains unclear how these common behavioural adaptations emerge from the idiosyncratic neural circuitry of each individual. The overall organisation of neural circuits is preserved across individuals[55] due to their common evolutionarily-specified developmental program[56–58]. Such organisation at the circuit level may constrain neural activity[16,59–61], leading to low-dimensional latent dynamics across the neural population[13,62,63]. Accordingly, here we hypothesised that the shared circuit level constraints within a species lead to suitably preserved latent dynamics across individuals. We analysed recordings of neural populations from monkey and mouse motor cortex to demonstrate that neural dynamics in individuals from the same species are surprisingly preserved when they perform similar behaviour. Neural population dynamics were also preserved when animals consciously plan future movements without overt behaviour[64], and enables the decoding of planned and ongoing movement across different individuals. Furthermore, we found that preserved neural dynamics extended beyond cortical regions to the dorsal striatum, an evolutionarily older structure[65,66]. Finally, we used neural network models to demonstrate that behavioural similarity is necessary but not sufficient for this preservation. We posit that these emergent dynamics result from

evolutionary constraints on brain development, and thus reflect fundamental properties of the neural basis of behaviour.

2.2 Introduction

The behaviour of each individual in a species is driven by the coordinated activity of neural populations throughout the brain. This activity emerges from the *latent dynamics*, which are the time-dependent activation of the dominant patterns of neural co-variation[13,63]. These latent dynamics seem to be shaped by circuit and biophysical constraints[16,59–61]. Given the large differences in brain circuitry across individuals from the same species —including in number of type-specific neurons, dendritic morphology, and receptor distribution[67–71]— it remains unclear how similar adaptive behaviours emerge from such idiosyncratic neural circuitry. One possibility is that unique circuits in each individual generate unique latent dynamics that produce the same behaviour. Indeed, the high degrees of freedom of neural activity relative to behaviour[72,73] could allow distinct latent dynamics to produce similar behavioural output. Alternatively, the same behaviour performed by two individuals is produced by preserved latent dynamics. This preservation emerges from the common organisation of neural circuits across individuals resulting from a species-specific developmental program.

Here, we adopt the latter hypothesis: different individuals from the same species engaged in the same behaviour generate preserved latent dynamics. We posit that preserved circuit constraints give rise to a species-wide *neural landscape*, and the individual-specific latent dynamics observed during a behaviour are different instantiations of a common trajectory through this landscape (Figure 2.1). Our hypothesis provides several testable predictions. First, since low-level details of neural circuits are idiosyncratic, they should not be necessary to account for the emergence of species-typical behaviours. Accordingly, different animals of the same species engaged in the same behaviour should exhibit preserved latent dynamics. Second, the extent of preservation of the latent dynamics across individuals should be constrained by the similarity of the behavioural output.

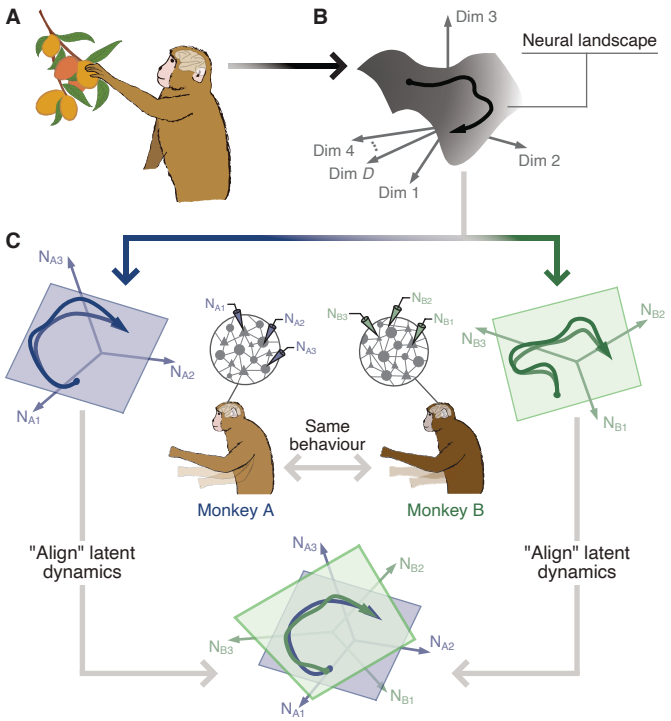


Figure 2.1: Hypothesis. Different individuals from the same species performing the same behaviour (Panel **A**) will generate preserved neural population latent dynamics by instantiating a species-wide 'neural landscape' (Panel **B**). These preserved latent dynamics can be revealed by 'aligning' the individual-specific latent dynamics estimated from their neural population recordings (Panel **C**).

Third, since low-dimensional latent dynamics are found throughout the brain, not just in cortical regions[74–76], we also expect structures that have co-evolved with cortex for hundreds of millions of years such as the basal ganglia[66] to exhibit preserved latent dynamics. Fourth, since covert behaviours seem to be mediated by the same neural circuits as overt behaviours[77], we should find shared latent dynamics across animals performing the same cognitive task.

2.3 Results

Preserved latent dynamics across animals

We tested the four predictions outlined above using neural population recordings from monkeys and mice as they performed upper limb tasks. First, we analysed motor cortical

recordings from monkeys engaged in an instructed-delay centre-out reaching task with eight targets (Figure 2.2A, Figure S2.1A, Methods). All three monkeys were well-trained in the task and exhibited highly stereotyped hand trajectories (mean trajectory correlation for each monkey: $r = 0.89$, 0.90, and 0.92, Figure S2.1B). For each session, we used Principal Component Analysis (PCA) to estimate the latent dynamics underlying overt movement execution by projecting the firing rates of each recorded neuron (or multi-unit, for monkey J) onto the leading ten PCA axes (the *neural modes*; examples in Figure S2.1C; Methods). We then *aligned* the latent dynamics of each pair of experimental sessions from two different animals using Canonical Correlation Analysis (CCA[43]), a method that maximises the correlations between two sets of signals through linear transformations (similar to Ref. 43,78,79).

This linear method revealed that the ostensibly different latent dynamics of two different monkeys are indeed highly preserved (Figure 2.2E). The across-animal correlations greatly exceeded two lower bound controls. The first was established by aligning randomly selected behavioural epochs ('Control' in Figure 2.2F) and the second was based on surrogate data that conserved the statistical structure of the neural activity[80] (TME, Figure S2.2A,B). Most importantly, these correlations were nearly as high as values obtained by aligning two subsets of trials within a single session from the same animal ('Within' in Figure 2.2F; additional examples in Figure S2.3A). This result further held across all pairs of sessions from all three monkeys ($n = 126$ sessions Figure 2.2G). The aligned neural modes captured a large fraction of neural variance (Figure S2.2E), and the results did not depend on the assumed dimensionality of the neural manifold (Figure S2.2F) or the alignment method (Figure S2.2G,H).

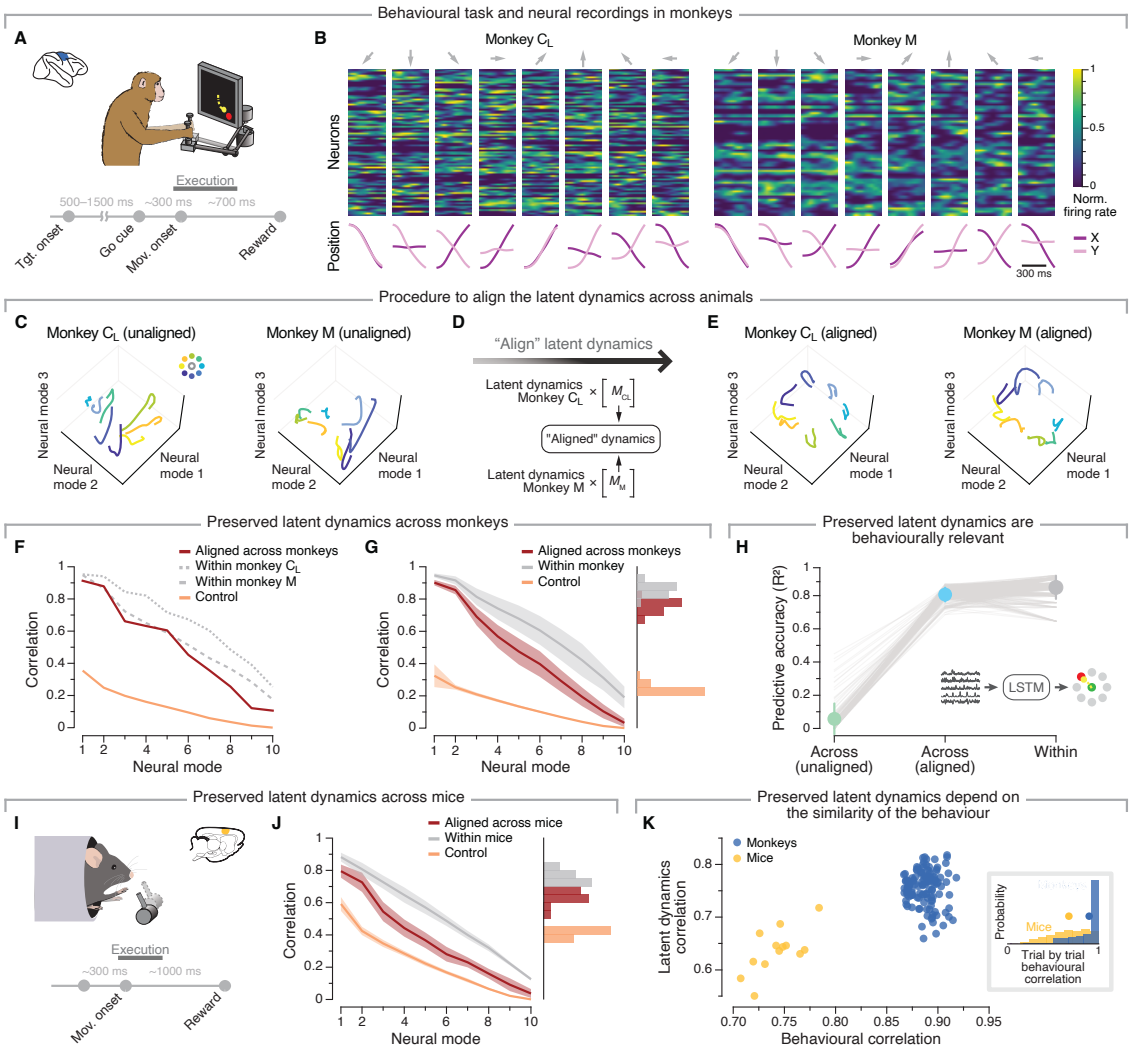


Figure 2.2: Preserved latent dynamics across animals performing the same behaviour. A) Monkeys performed an 8 target instructed-delay centre-out reaching task using a planar manipulandum. **B**) Example normalised neural firing rates aligned to movement onset (top) and hand trajectories (bottom) for two monkeys. Each column shows one reach to the eight targets indicated by the arrows. **C-E**) Three-dimensional representation of the motor cortical latent dynamics for the two monkeys plotted in Panel B before (Panel C) and after (Panel E) alignment with CCA (Panel D). **F**) Correlations of the aligned (red) latent dynamics for the example comparison in Panels C–E compared to within-monkey correlations (grey) and a lower-bound control (orange). **G**) Preserved latent dynamics across all pairs of 21 sessions (126 comparisons) from three different monkeys. Histograms show the mean correlation across the leading four dimensions. Line and shaded area, mean±s.d. **H**) Decoders trained on aligned latent dynamics from one monkey predict movement parameters of a different monkey (blue). Results compared to decoders trained and tested within the same session (grey) and without alignment (green). Data points, individual comparisons ($n = 126$) between sessions from different monkeys. Errorbars, mean±s.d. Comparisons, two-sided Wilcoxon's rank sum test, $P < 0.001$. **I**) Mice grasped and pulled a joystick in two positions (left or right). **J**) Preserved latent dynamics across mice performing the grasping and pulling task. Data include six sessions across four different mice, formatted as in Panel G. **K**) Preserved latent dynamics across animals is related to the similarity of their behaviour. For each session pair, the mean of the top four canonical correlations between the latent dynamics against the mean behavioural correlation. Single dots, pairs of sessions colour coded by species. Inset: Behavioural

correlations for all pairs of trials from different mice and monkeys. Circles, mean.

While we have demonstrated the preservation of latent dynamics across animals, these shared dynamics may not necessarily be relevant to behaviour. To address this, we trained neural network decoders (Long Short-Term Memory networks, or LSTMs) to predict the hand trajectories of one animal[81] and tested their performance on a second animal. The across-animal decoding accuracy approached the upper bound provided by the performance of decoders trained and tested on the same animal (Figure 2.2H; example predictions in Figure S2.4). Thus, the shared motor cortical dynamics across animals do contain detailed information about the ongoing movement kinematics.

We then analysed data from four mice trained to perform a reaching and pulling joystick task (Figure 2.2I and Figure S2.5). We found that both within and across individuals, the behavioural output was less similar from trial to trial than for the monkey dataset (compare Figure S2.1A,B to Figure S2.5A,B). As predicted, our alignment procedure revealed that motor cortical latent dynamics were largely preserved across mice both when reaching to two different targets (Figure 2.2J Figure S2.3B, Figure S2.2C; example in Figure S2.3E), and when subsequently pulling at two different force levels (Figure S2.6A). Yet, the extent of these correlations was lower than those of monkeys' (compare Figure 2.2G to Figure 2.2J), which directly impacted our ability to decode movement kinematics across animals (Figure S2.6B). Much of this difference could be explained by the more highly stereotyped behaviour of the monkeys compared to the mice (Figure 2.2K-inset). Comparing between species confirmed that behavioural stereotypy was indeed well associated with both the preservation of the latent dynamics across individuals (Figure 2.2K) and across-animal decoding accuracy (Figure S2.6F). These results demonstrate in two evolutionarily divergent species that there is a direct correspondence between the similarity of behavioural output and the preservation of motor cortical latent dynamics across individuals.

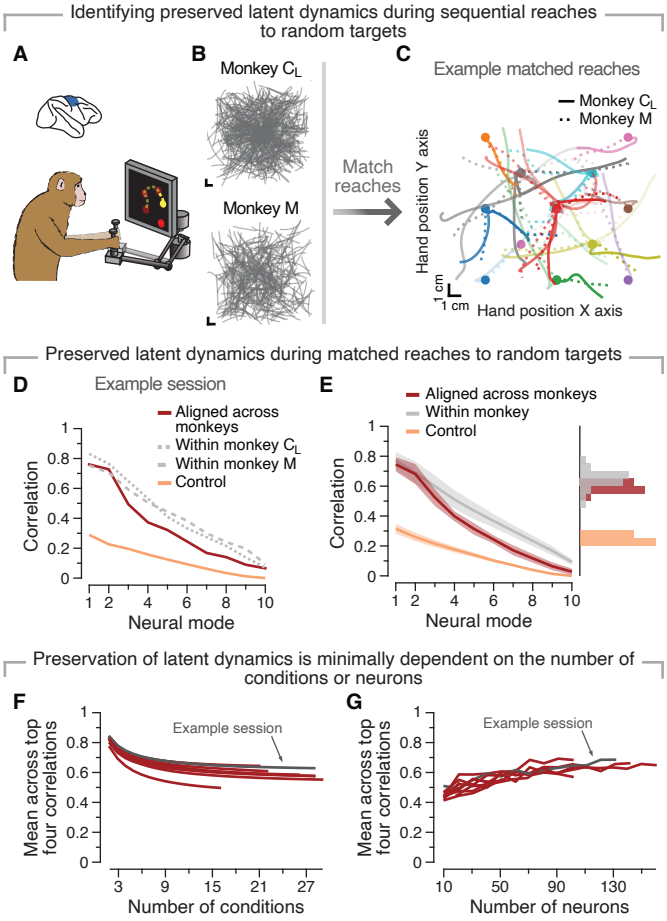


Figure 2.3: Latent dynamics are preserved in a sequential reaching task with higher behavioural complexity. A) Monkeys were trained to perform sequences of four reaches to randomly-placed target locations using the planar manipulandum. **B**) Example hand positions in the workspace for all reaching movements made during a session from two example monkeys. **C**) We matched similar movements between the two monkeys based on their starting location, duration, and direction. Each color represents a different representative condition, and matched reaching trajectories are overlaid for the two monkeys. **D**) Preservation of motor cortical latent dynamics across monkeys performing matched reaches for the example sessions in Panel C. Note that the correlation between aligned latent dynamics across monkeys was quite close to the within-monkey correlations (grey), and largely exceeded our lower-bound control (orange). **E**) Preservation of motor cortical latent dynamics across all pairs of sessions from three different monkeys. Data include 11 sessions and $n = 10$ comparisons, and are presented as in Figure 2.2G. Line and shaded area, mean$\pm$s.d. **F**) The preservation of latent dynamics holds for a wide range of conditions. The plot shows, for each pair of sessions, the mean of the top four canonical correlations between the latent dynamics as a function of the number of conditions subsampled from the total available in each session. Black trace denotes the example session from Panels B-D. **G**) Preserved latent dynamics can be uncovered from even sparsely sampled neural populations. The plot shows, for each pair of sessions, the mean of the top four canonical correlations between the latent dynamics as a function of the number of neurons subsampled from the total available in each session. Black trace denotes the example session from Panels B- D.

Necessity of behavioural similarity

In the preceding analyses, we studied tasks comprising a small number of conditions that inadvertently imposed a topological structure in the produced movements[43,64]. We sought to establish that preserved latent dynamics do not merely reflect this structure. First, we tested whether this topological structure is sufficient to produce preserved latent dynamics and found that the preservation across individuals was significantly impaired (Figure S2.7). We then further demonstrated the existence of preserved latent dynamics during a continuous and less structured task in which monkeys rapidly generated sequences of random reaches[82,83] (Figure 2.3A). The produced movements were highly varied with little organisation. (Figure 2.3B, S2.8A). To facilitate alignment, we parcellated the workspace to match movements across individuals based on the initial hand position and reach direction in order to generate as many as 29 similar reaching conditions (Figure 2.3C). Despite the dramatic increase in number of conditions and behavioural variability, we could still uncover preserved latent dynamics across individuals (Figure 2.3D,E, Figure S2.2D). The increased complexity of behavioural output in this task allowed us to directly study the relationship between the number of conditions and the preservation of latent dynamics. We subsampled the conditions and found that latent dynamics across individuals were preserved for the entire range (up to 29) considered (Figure 2.3F, S2.8B) even when as few as 20 neurons are considered (Figure 2.3G, S2.8C). The preservation of latent dynamics across individuals decreased when we shuffled the conditions (Figure S2.8D), thereby reducing behavioural similarity. This agrees with the previous comparison of mice and monkeys: monkeys performing a more complex task (8 conditions) than mice (2–4 conditions) had higher preservation of latent dynamics due to the more stereotyped behavioural output.

We performed several control analyses to confirm that this close relationship between behavioural stereotypy and preservation of latent dynamics across individuals from the same species is not a trivial consequence of our methodology. First, we compared our results with a previous study that investigated the dynamics of motor cortical activity

within an individual across different but related wrist manipulation and reach-to-grasp tasks[44] (Figure S2.9). During each task, individuals activated the same muscles in a slightly different manner (Figure S2.9B-F). Accordingly, the latent dynamics of the same monkey performing two distinct but related behaviours are much less preserved than those of different monkeys performing the same behaviour (Figure S2.9G). Lastly, as we have previously shown, preserved latent dynamics cannot be explained by stable movement tuning[43], nor do they persist following nonlinear transformations[43]. These analyses demonstrate that the alignment method alone is not sufficient to uncover preserved latent dynamics, even within individuals.

Preserved dynamics in dorsal striatum

Given that the motor cortex is the main cortical output to the spinal circuits that generate movement, the close correspondence between motor cortical latent dynamics and behavioural output may uniquely result from the architecture and projections of this region. To test whether preserved latent dynamics exist across the brain, we studied the subcortical nuclei of basal ganglia, which do not directly project to spinal cord but are crucial for various aspects of behaviour[7,84–87]. We predicted that basal ganglia latent dynamics would also be preserved across animals performing the same task. Replicating our alignment analysis on neural population recordings from mouse dorsolateral striatum during a reaching and pulling task (Figure 2.4A) revealed preserved latent dynamics across individuals (Figure 2.4B and Figure S2.3C; example in Figure S2.3G).

Moreover, despite the vast differences in circuit and cellular architecture between motor cortex and striatum[87,88], both the across-animal correlations (compare Figure 2.4B and Figure 2.2J) and the across-animal decoding performance of hand trajectories (Figure S2.6F) were equally large for both regions (Figure S2.6C-E). Therefore, stable latent dynamics across animals performing the same behaviour are not confined to motor cortex—they extend to different regions throughout the brain, including an evolutionarily older structure that is shared among all vertebrates[65].

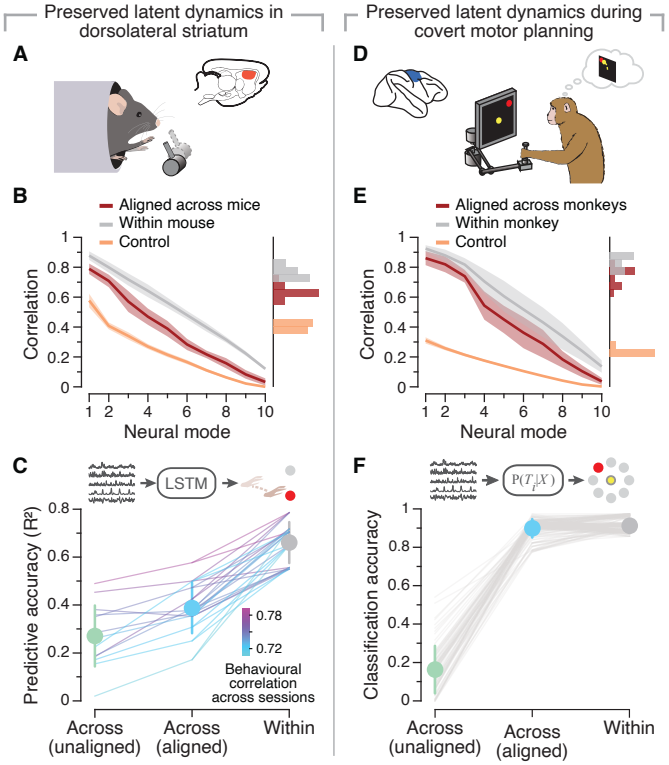


Figure 2.4: Preserved latent dynamics extend beyond cortical regions to the evolutionarily older basal ganglia, and are also present during covert behaviour. A) We studied whether shared latent dynamics can also be found in subcortical regions by analysing recordings from mouse dorsolateral striatum during a reaching and pulling task. **B)** Preservation of striatal latent dynamics across mice performing the same grasping and pulling task. Data include six sessions $n = 6$ comparisons across four mice and are presented as in Figure 2.2G. Line and shaded area, mean±s.d. **C)** Hand trajectories can be decoded across mice using their shared striatal latent dynamics. Data presented as in Figure 2.2H. Errorbars, mean±s.d. For all decoder comparisons, two-sided Wilcoxon's rank sum test, $P < 0.001$. **D)** To investigate the preservation of latent dynamics during covert behaviour, we examined the preparatory period preceding movement execution in the monkey centre-out task. **E)** Preservation of motor cortical latent dynamics when preparing to reach to a target. Data include 18 sessions and $n = 72$ comparisons across two different monkeys, and are presented as in Figure 2.2G. Line and shaded area, mean±s.d. **F)** The preserved latent dynamics during covert behaviour contain behaviourally-relevant information. Naïve Bayes classifiers trained on the aligned latent dynamics (blue) predicted the intended target in a different monkey virtually as well as classifiers trained and tested on the same monkey (grey), while classification was poor without alignment (green). Errorbars, mean±s.d. For all classifier comparisons, two-sided Wilcoxon's rank sum test, $P < 0.001$.

Preserved dynamics during covert behaviour

We have shown the preservation of latent dynamics across brain regions during active, overt behaviour. However, behaving animals also engage in a variety of covert behaviours such as deliberation and planning. These processes require neural activity that is predomi-

nantly generated internally by the brain. Given that such covert behaviours involve brain regions that also mediate overt behaviours[37,77], we predicted that the latent dynamics underlying these more cognitive processes would also be shared across individuals of the same species. We tested this prediction by analysing motor cortical activity as monkeys planned an upcoming movement prior to executing it (Figure 2.4D). The latent dynamics during the instructed delay period were highly correlated across animals (Figure 2.4E and Figure S2.3D; example in Figure S2.3G), and were virtually identical to those during overt reaching behaviour (*cf*. Figure 2.2G). Moreover, these aligned latent dynamics were also predictive of behaviour: Bayesian models predicting the upcoming reaching target based on the aligned latent dynamics from one monkey generalised to another (Figure 2.4F; Methods). Thus, different individuals use preserved latent dynamics not only to execute the same movement but also to perform the same covert mental process. This result also strengthens our previous observation of preserved latent dynamics during overt behaviour. Afferent feedback arriving at the motor cortex[89,90] could partially explain the observed similarity in latent dynamics during overt movement, yet the latent dynamics are entirely internally generated during covert processes such as movement planning.

Behavioural similarity is not sufficient

Our hypothesis requires that behavioural similarity is necessary but not sufficient to allow for alignment of latent dynamics. To test this, we trained recurrent neural networks (RNNs) to produce similar behavioural output while generating distinct latent dynamics. We devised an RNN simulation where we had control over the degree of preservation of latent dynamics by varying a parameter of the cost function, α (Figure 2.5A, S2.10). We thus created pairs of models that generated highly similar behaviour (Figure 2.5B-D) but exhibited distinct latent dynamics as evidenced by the relative lack of alignment (Figure 2.5E). We predicted that this decrease in preservation of latent dynamics would be driven by differences in underlying circuit properties. When we reverse-engineered the weights of the different networks, we found that more dissimilar latent dynamics corresponded to larger changes in the variance and dimensionality of the weights (Figure 2.5F). Thus,

preservation of latent dynamics is not just a trivial consequence of behavioural similarity; instead, it also likely reflects fundamental organisation and constraints in the underlying circuit implementation.

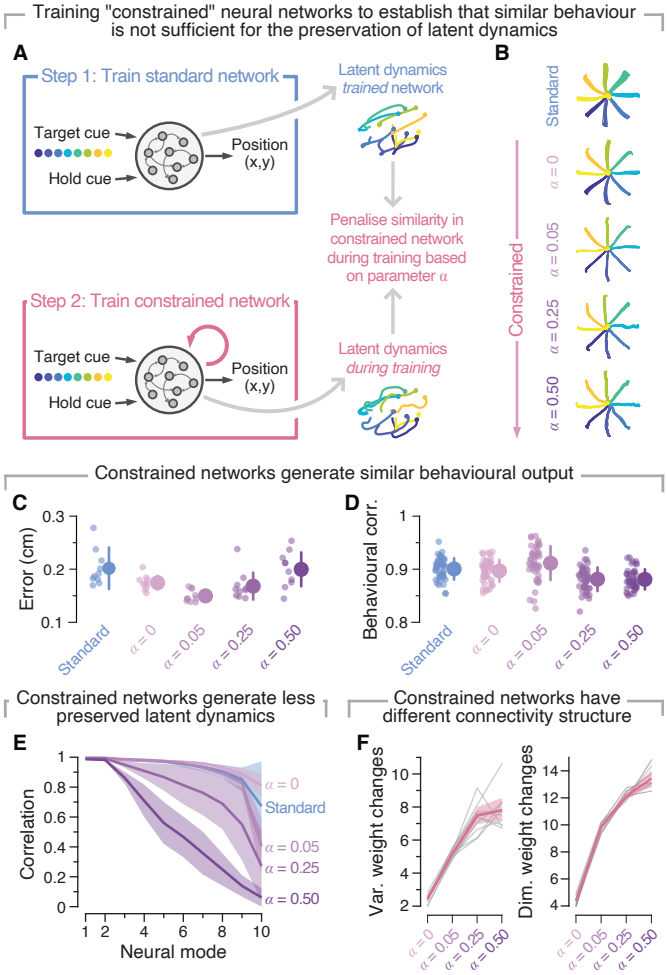


Figure 2.5: Similar behavioural output is necessary but not sufficient for the preservation of latent dynamics A) We trained two sets of recurrent neural network models to perform the monkey centre-out reaching task. The first set of models were standard recurrent neural networks, while the second were constrained to produce distinct latent dynamics from the standard networks. The weight of this additional constraint was controlled by parameter α. **B**) Example movement trajectories for standard and constrained networks. **C-D**) Both standard and constrained networks produced similar behavioural output, as shown by the error (Panel C) and the correlation between 'reach trajectories' across networks (Panel D). Data show 10 networks initialized from different random seeds for each type of network, 45 comparisons within each network condition. Errorbars, mean±s.d. **E**) Increasing the value of α decreased the preservation of latent dynamics, establishing that behavioural similarity is necessary but not sufficient to have preserved latent dynamics. Line and shaded area, mean±s.d. **F**) Networks with distinct latent dynamics show differences in underlying connectivity. The variance dimensionality and the dimensionality of the weight changes increased as the latent dynamics were constrained to be more distinct from those of standard networks by increasing α. Variance values are scaled by 10^5. Line and shaded area, mean±s.d.

2.4 Discussion

Neural population latent dynamics have been proposed as first-level explainers of behavioural and cognitive functions[13,63], a view that has shed light onto the neural basis of numerous phenomena, such as processes underlying covert[37,77,91] and overt behaviour[44,92,93], how learning may happen in neural circuits[14,59,94], and how information may flow between different brain regions[90,95,96]. Here, we extend recent works[72,97–99] to show that latent dynamics are shared across different individuals engaged in the same behaviour for a range of behavioural complexity. The discovery of preserved latent dynamics across individuals will impact both fundamental and applied neuroscience, in particular the development of brain-controlled devices such as neuroprosthetics[100,101]: with proper alignment, decoders trained on one participant could be readily translated to other individuals to minimise training and deployment time[102–106].

We studied two regions (motor cortex and dorsolateral striatum) whose functions are tightly linked to the production of motor output. Accordingly, the similar biomechanical properties of limbs across individuals provide additional constraints that help to enforce the preservation of latent dynamics. These constraints, together with the temporally smooth and relatively constrained movements we studied, allowed us to demonstrate preserved latent dynamics using a relatively small number of dimensions. Highly sensory-driven regions with rapidly changing inputs may require a higher level of granularity in the analysis, but we expect to see similar preservation of latent dynamics across individuals. In contrast, 'higher' brain regions (e.g., frontal cortex) that serve more abstract cognitive functions may show less preservation of latent dynamics across animals due to differences in their internal states (attention, motivation, satiation, etc.). Even if the most task-relevant aspects of the latent dynamics were preserved across individuals, the influence of these more abstract internal states[107,108] could alter the overall latent dynamics, making the activity of these regions less amenable to alignment. Moreover, different animals may also use different covert processes to solve the same cognitive task based on their biases acquired through learning and past experience[109]. Yet, we expect that the neural population latent

dynamics would still be preserved if it were possible to appropriately match the internal states or strategies across individuals.

The influence of learning and experience on neural circuit organisation and the resulting latent dynamics is a fascinating open question. Throughout an individual's lifetime, new skills are acquired through practice, which are consolidated through changes in neural circuitry [110,111]. What, if any, differences should exist in the latent dynamics produced by the motor cortices of, say, a virtuoso guitarist and a first-year guitar student? Regardless of whether those circuits were shaped by practice or even development, we propose that neural circuits are tuned to produce latent dynamics as a solution for behavioural output. In this way, preserved latent dynamics across individuals reflect a fundamental property of the neural basis of behaviour.

Here, we introduced two competing possibilities to explain how similar behaviour can emerge from different neural circuitry. In the first hypothesis, each individual's brain must learn during development to produce latent dynamics within their idiosyncratic neural circuits to enable the desired behaviour, with no guarantee that both individuals arrive at the same 'solution'. In the second hypothesis, the generated latent dynamics for the desired behaviour are constrained by the genetically-specified organisation at the circuit level shared by each individual of a given species. Our results support this second hypothesis, and raise an intriguing possibility: given that in higher vertebrates the genome does not specify the implemented architecture in great detail, e.g., to the level of the synapse[56–58], the genome may provide a 'generative model' that is instantiated by each individual's brain. This generative model may constrain low-level details of the neural circuitry such that the appropriate neural population latent dynamics required for the behavioural repertoire emerge throughout development.

2.5 Methods

Subjects and behavioural tasks

Monkeys

We trained four monkeys (monkeys C, M, J: male, *Macaca mulatta*; monkey T: male, *Macaca fascicularis*; aged 6-10 years) to sit in a primate chair and make reaching movements using a customised planar manipulandum. The movement of a cursor on a computer screen was mapped to the motion of the handle of the manipulandum, and the behavioral task was run through custom software in Matlab (The Mathworks, Inc.). Monkeys C, M, and J were trained to perform a two-dimensional centre–out reaching task for at least several months prior to the neural recordings, ensuring they had reached expert performance. Monkeys C, M, and T were trained on a more complex random target sequential reaching task. In the centre-out task, the monkey moved its hand to the centre of the workspace to begin each trial. After a variable waiting period, the monkey was presented with one of eight outer targets, equally spaced in a circle and selected randomly with uniform probability. Then, an auditory go cue signalled the animals to reach to the target. Monkeys were required to reach the target within 1 s after the go cue and hold for 0.5 s to receive a liquid reward, except for monkey J, who was trained without the instructed delay period or the 0.5 s target hold time and therefore made larger movements (Figure S2.1A, right). For the centre-out task, there were 12 sessions for monkey C, six sessions for monkey M, and three sessions for monkey J. In the random target task, the monkeys made four consecutive reaches to random targets within a 10 cm x 10 cm workspace in each trial. They received a liquid reward during a short break after each successful sequence of four random target acquisitions. There was no explicit auditory go cue and only a brief hold period within the target (100 ms) and then a brief delay period (100 ms) before the next target was presented. These short constraints helped to enforce that the monkeys made separate, directed movements but did not require that the monkey necessarily stop

between movements. For the random target task, there was one 'reference' session for monkey C, six sessions for monkey M, and four sessions for monkey T. As the monkeys performed these tasks, we recorded the position of the endpoint at a sampling frequency of 1 kHz using encoders in the joints, and digitally logged the timing of task events, such as the go cue. Portions of the centre-out reaching data have been previously published and analysed in Refs. 14,43,79,112. Portions of the random target data have been previously published and analysed in Refs. 82,83.

Mice

Four 8–16 week old mice were trained to perform a forelimb reaching and pulling task (similar to Refs. 87,113) for approximately one month, following habituation to head-fixation and the recording setup. In each trial, mice had to reach and pull a joystick positioned approximately 1.5 cm away from the initial hand position. The joystick appeared, without any cue, in one of two positions (left or right, < 1 cm apart). Mice could then self-initiate a reach to the joystick and pull it inwards to get a liquid reward. The joystick was weighted with either a 3 g or a 6 g load (light or heavy), making up a total of 4 trial types (2 joystick positions $\times$ 2 loads). Each trial type was repeated 20 times before task parameters were switched to the next trial type without any cue. Each session consisted of two repetitions of each set of four trial types presented in the same order, making up $2 \times 4 \times 20 = 160$ trials. Trials with incorrect responses (e.g., pushing the joystick past a threshold, 5mm) or time-out (the lack of pull or push for 10 s) were marked as unsuccessful. All joystick operations were programmatically controlled using a custom-written open source Python package: (`https://github.com/janelia-pypi/mouse_joystick_interface_python`). Mice were maintained on a 12-12 h (8am - 8pm) light/dark cycle and recordings were made between 9am and 3pm. The holding room temperature was maintained at 21$\pm$1°C with a relative humidity of 30% to 70%.

There were two sessions for mouse 38, one session for mouse 39, two sessions for mouse 40, and one session for mouse 44. Movement kinematics were tracked using markerless

video-based pose estimation. Annotation of behavior was accomplished using Janelia Automatic Animal Behavior Annotator[114] (JAABA). Briefly, behavior was recorded using two synchronised high-speed (500 frames/s), high resolution monochrome cameras (Point Grey Flea3; 1,3 MP Mono USB3 Vision VITA 1300; Point Grey Research Inc., Richmond, BC, Canada) with 6-mm to 15-mm (f/1.4) lenses (C-Mount; Tokina, Japan), placed perpendicularly in front and to the right of the animal. A custom-made near-infrared light-emitting diode light source was mounted on each camera. Video was recorded using custom-made software developed by the Janelia Research Campus Scientific Computing Department and IO Rodeo (Pasadena, CA). This software controlled and synchronized all facets of the experiment. For the main analyses, light and heavy trials were pooled together, since we focused on the reaching phase of the task and the location of the joystick does not depend on its weight. Note that in Figure S2.6A we repeated the main analysis to demonstrate preserved latent dynamics during the pulling phase, considering all four conditions.

Neural recordings

Monkeys

All surgical and experimental procedures were approved by the Institutional Animal Care and Use Committee of Northwestern University under protocol #IS00000367. We implanted 96-channel Utah electrode arrays in the primary motor cortex (M1) or dorsal premotor cortex (PMd) using standard surgical procedures. Throughout the paper, neural recordings from these two subregions were pooled together and denoted as motor cortex. This allowed us to ensure we could evaluate overt and covert dynamics within the same population. Implants were done in the opposite hemisphere of the hand animals used in the task. Monkeys M and T received two arrays in M1 and PMd simultaneously. Monkey J received a single array in M1. Monkey C received two sets of implants: one array in the right M1 while performing the task using the left hand, and following removal of this original implant, two arrays simultaneously in the left M1 and PMd while using the right hand (respectively, monkeys $\mathrm{C_R}$ and $\mathrm{C_L}$ in our previous work[43]). Note that for all

across-individual analyses, $\mathrm{C_R}$ and $\mathrm{C_L}$ are considered the same animal.

Neural activity was recorded during the behaviour using a Cerebus system (Blackrock Microsystems). The recordings on each channel were band-pass filtered (250 Hz–5 kHz) and then converted to spike times based on threshold crossings. The threshold was set to 5.5 $\times$ the root-mean square activity on each channel. We also manually spike sorted the recordings from monkeys C and M to identify putative single neurons. Monkey J had fewer well-isolated single units than the other monkeys, so rather than spike sorting we directly applied the multi-unit threshold crossings acquired on each electrode. However, it has been shown that the latent dynamics estimated from multi-unit and single neuron activity are similar[115], an observation that holds true for aligning latent dynamics with CCA[43] (note that we refer to both single neurons and multi-units simply as units). We included multiple experimental sessions from each monkey: for the centre-out reaching task, eight from monkey $\mathrm{C_L}$, four from monkey $\mathrm{C_R}$, six from monkey M, and three from monkey J; for the random target task (example data in Figure S2.1), one 'reference session' from monkey C, six from monkey M, and four from monkey T (example data in Figure S2.8). These experimental sessions were chosen based on the high number of units or trials and blind to the behaviour of the animal. For the centre-out reaching task, the average number of units included for each monkey was: monkey $\mathrm{C_L}$, 277 ± 14 (mean$\pm$s.e.m.; range, $210 - 345$); monkey $\mathrm{C_R}$, 85 ± 4 (range, $73 - 92$); monkey M, 117 ± 4 (range, $106 - 130$); and monkey J, 63 ± 9 (range, $54 - 81$). For the random target task, the average number of units included was: monkey $\mathrm{C_L}$, 280 (one session only); monkey M, 127 ± 9 (range, $101 - 153$); and monkey T, 49 ± 8 (range, $30 - 66$). A more detailed description of the behavioural and neural recording methods is presented in Ref.[43].

Mice

All surgical and experimental procedures were approved by the Institutional Animal Care and Use Committee of Janelia Research Campus. A brief ($<$2 h) surgery was first performed to implant a 3D-printed headplate[116]. Following recovery, the water consumption of

the mice was restricted to 1.2 ml per day, in order to train them in the behavioural task. Following training, a small craniotomy for acute recording was made at 0.5 mm anterior and 1.7 mm lateral relative to bregma in the left hemisphere. A Neuropixels probe was centred above the craniotomy and lowered with a 10 degree angle from the axis perpendicular to the skull surface at a speed of 0.2 mm/min. The tip of the probe was located at 3 mm ventral from the pial surface. After a slow and smooth descent, the probe was allowed to sit still at the target depth for at least 5 min before initiation of recording to allow the electrodes to settle.

Neural activity was filtered (high-pass at 300 Hz), amplified ($200\times$ gain), multiplexed and digitised (30 kHz), and recorded using the SpikeGLX 3.0 software (`https://github.com/billkarsh/SpikeGLX`). Recorded data were pre-processed using an open-source software KiloSort 2.0 (`https://github.com/MouseLand/Kilosort`) and manually curated using Phy (`https://github.com/cortex-lab/phy`) to identify putative single units in each of the primary motor cortex and dorsolateral striatum. A total of six experimental sessions (from four mice, see Figure S2.5) with simultaneous motor cortical and striatal recordings were included in this work. The average number of motor cortical units included for each mouse was: mouse 38, 98 ± 4 (range, $95 - 102$); mouse 39, 64; mouse 40, 75 ± 5 (range, $70 - 80$); and mouse 44, 55. The average number of striatal units included for each mouse was: mouse 38, 100 ± 13 (range, $87 - 112$); mouse 39, 108; mouse 40, 74 ± 5 (range, $69 - 79$); and mouse 44, 110.

Data analysis

We used a similar approach for both monkey and mouse data. In all the analyses, we only considered the trials in which the animal successfully completed the task within the specified time and received a reward, which we concatenated—that is, no trial-averages were taken. For the monkey centre-out reaching task and the mouse reaching and pulling task, an equal number of trials to each target was randomly selected (eight targets for the monkeys and two targets for mice, except in Figure S2.6A, where four targets where

considered). Trial order was randomised to eliminate the possible effect of the passage of time. Within each trial, we isolated a window of interest that captured most of the movement, starting 50 ms before movement onset and ending 400 ms after movement onset. To analyse covert behaviour in monkeys, we used a window that spanned the movement planning period, which started 400 ms before movement onset and ended 50 ms after movement onset. Importantly, all of our results held when changing the analysis windows within a reasonable range.

For the monkey random walk task, each reach could start and end anywhere within the workspace. To define movements (conditions) that could be matched across animals, we first segmented the workspace into 12 circular subsections. Each subsection was then divided into six equal sectors, and targets in the same angular sector were grouped together, creating 72 possible target conditions. We separated the sequences of four consecutive reaches and considered each reach as a separate movement. To assign each movement to a target condition, we first assigned each movement to one of the subsections based on the starting position of the given movement, excluding movements that started more than 2 cm from the center of the subsection. We then re-centered the movements so that they started in the center of each subsection, and reached outwards towards their target. The movement was then assigned to a sector and target condition based on the angle of target. To study the preservation of latent dynamics across monkeys performing similar behaviour, we needed to match movements (reach conditions) across sessions for different monkeys. To maximise the number of matched movements, we compared all sessions for Monkey M and Monkey T against a reference session for Monkey $\mathrm{C_L}$ that had the greatest number of successful trials. We matched movements in each pair of sessions by minimizing the mean-squared error (MSE) between pairs of movements, excluding matches that had MSEs above the threshold of 2% of MSEs calculated for all possible pairs of movements. If the matched movements had different corresponding target conditions, we used the target condition label from the reference session. After this process was completed, we excluded target conditions with less than six matched movements, such that paired sessions had up to 29 shared target conditions. Since these movements were more ballistic than in the

other tasks, for this task we examined a window starting at movement onset and ending 350 ms after movement onset.

All the analyses were implemented in Python using open-source packages such as `numpy`, `matplotlib`, `sci-kit`, `scipy`, and `pandas`[117–121] and custom code. As we were analyzing existing datasets on an individual basis, no explicit planning of sample size, group randomization, or blinding was performed.

Behavioural correlation

To assess the behavioural stereotypy of a given animal, we calculated hand trajectory correlations (Pearson's correlation) of every pair of trials within a session (Figure S2.1B, S2.5B). The distributions in Figure 2.2K-inset illustrate these correlations pooled across all the monkey centre-out and mouse reaching and pulling sessions included in this work. To determine the behavioural similarity across pairs of sessions from different monkeys or mice (Figure 2.2K), we similarly calculated correlations to compare all pairs of trials from the two sessions.

Neural population latent dynamics

To estimate the latent dynamics associated with the recorded neural activity in each session for both mice and monkeys, we computed a smoothed firing rate as a function of time for each unit. We obtained these smoothed firing rates by applying a Gaussian kernel ($\sigma = 50$ ms) to the binned square-root transformed firings (bin size$= 30$ ms) of each unit. We excluded units with a low mean firing rate (< 1 Hz mean firing rate across all bins), but we did not perform any further exclusions, e.g., based on lack of modulation or behavioural tuning. For each session, this produced a neural data matrix X of dimension n by T, where n is the number of recorded units and T the total number of time points from all concatenated trials on a given day; T is thus given by the number of targets per day $\times$ number of trials per target $\times$ number of time points per trial. We performed this

concatenation as described above after randomly sub-selecting the same number of trials for all targets for each animal (15 trials for monkey center-out, six for monkey random walk, 22 for mouse reaching and pulling). For each session, the activity of n recorded units was represented as a neural space—an n-dimensional sampling of the space defined by the activity of all neurons in that brain region. In this space, the joint recorded activity at each time bin is represented as a single point, the coordinates of which are determined by the firing rate of the corresponding units. Within this space, we computed a low-dimensional neural manifold by applying PCA to X. This yielded n PCs, each a linear combination of the smoothed firing rates of all n recorded units. These PCs are ranked based on the amount of neural variance they explain. We defined an m-dimensional, session-specific manifold by only keeping the leading m PCs, which we referred to as neural modes. We chose a manifold dimensionality $m = 10$, based on previous studies examining motor cortical recordings during upper limb tasks[14,43,59]. Across all datasets, a ten-dimensional manifold explained $\sim 60\%$ of the neural variance for each of the monkey motor cortex (Figure S2.1), mouse motor cortex, and mouse striatum (Figure S2.5). Note, however, that our results held within a reasonable range of dimensionalities, similar to[14,43,44] (see Figure S2.2F and Figure S2.4B). We computed the latent dynamics within the manifold by projecting the time-varying smoothed firing rates of the recorded neurons onto the $m = 10$ PCs that span the manifold. This produced a data matrix L of dimensions m by T.

Aligning latent dynamics through CCA

We addressed our hypothesis that different animals performing the same behaviour would share preserved latent dynamics by aligning the dynamics using Canonical Correlation Analysis (CCA)[43,122]. CCA was applied to the latent dynamics of each pair of sessions after randomising the temporal order of trials and matching the number of trials to each target. For details on using CCA to align latent dynamics, see Ref. 43.

We measured the similarity in latent dynamics across animals by computing the across-animal correlations as the canonical correlations (CCs) across all pairs of sessions from

any two different monkeys or mice. To establish the strength of the across-animal correlations, we computed an upper bound defined by the within-animal correlations, which we calculated as the 99th percentile of the correlations between two randomly selected subsets of trials within any given session over 1000 samples. The 'control' correlations represent a lower bound for the CCs. We computed these by shuffling the targets across the two sessions and using a randomly selected control window (more details in the 'Control analyses' section below) in each trial, rather than the movement or preparatory epochs.

Note that to summarise each comparison to a single datapoint (e.g., in Figure 2.2K, Figure S2.2H, S2.6D, S2.7D), we computed the mean of the top four canonical correlations of the latent dynamics[43]. In Figure 2.2K, we used this approach to establish a relationship between the strength of preservation of the latent dynamics and the consistency of behaviour, quantified as the mean trajectory correlation of all possible pairs of trials across two animals. Furthermore, when showing pairs of 'aligned' trajectories across animals, such as in the right hand side of Figure 2.2C and Figure S2.3, the CCA axes were orthogonalised using singular value decomposition for visualisation purposes.

Finally, we showed that preserved latent dynamics could be uncovered across a broad range of manifold dimensionalities. In Figure S2.2F we repeated the alignment analysis for manifold dimensionalities $m = 2 - 19$.

Decoding analysis

To test whether the aligned latent dynamics maintain movement-related information, we built standard decoders to predict hand trajectory during overt behaviour. If the aligned latent dynamics across different animals were behaviourally relevant, they would allow predicting time-varying hand trajectories even if the methods used to identify them (PCA and CCA) are not supervised, that is, they do not attempt to optimise decoding performance. We compared the predictive accuracy of three different types of decoders: 1) a within-animal decoder trained and tested (using ten-fold cross-validation) on two non-overlapping subsets of trials from each session of each animal; 2) an across-animal

'aligned' decoder that was trained on the aligned dynamics from one animal and tested on another, a comparison we performed on each pair of sessions from two different animals; 3) an across-animal 'unaligned' decoder that was trained on the latent dynamics from one animal and tested on another without aligning the dynamics using CCA. We also performed a similar analysis to predict the upcoming target during covert movement preparation in monkeys (Figure 2.4C).

Hand trajectory decoders were LSTM models with two LSTM layers, each with 300 hidden units, followed by a linear output layer. The models were implemented with Pytorch[123] and trained for 10 epochs with the Adam optimiser, with a learning rate of 0.001. Upcoming target classifiers were Gaussian Naïve Bayes models[64] (the `GaussianNB` class in Ref. 119). We included three bins of recent latent dynamics history, for a total of 90 ms, in the input of both the decoders and the classifiers. These additional neural inputs incorporate information about intrinsic neural dynamics and account for transmission delays. The R^2 value, defined as the squared correlation coefficient between actual and predicted hand trajectories, was used to quantify decoder performance. Moreover, in Figure S2.4D we verified that our choice of across-animal decoder accuracy metric did not influence the observation that preserved latent dynamics are informative about behaviour by also computing a variance explained metric, defined as:

$$\mathrm{VAF} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i represents the actual value of the predicted variable, $\hat{y}_i$ its predicted value, and $\bar{y}$ its mean. For this analysis, we normalised hand trajectories by the length of the reaches (determined by the 99^{th} percentile of their hand positions along each axis) since monkeys had workspaces of different sizes.

Hand trajectory was a two-dimensional signal in monkeys and a three-dimensional signal in mice. We built separate decoders to predict hand trajectories along the X, Y (and Z

for mice) axes. We then reported the average performance across all axes. For target classification, we reported the mean accuracy of the classifier (the `score()` method).

Furthermore, to test how many dimensions of the aligned latent dynamics were needed for accurate across-animal decoding of behaviour, we repeated the decoding analysis in the monkey centre-out dataset for manifold dimensionalities $m = 1, 2, ..., 14$ (Figure S2.4B).

Finally, we performed a control analysis to ensure our across animal decoding results were not biased by sharing similar trials for both alignment and decoder training. We split the full dataset of one animal into three non-overlapping sets: one to align the latent dynamics, one to train the decoder, and one to test the performance across animals. Figure S2.4C shows the result of this analysis for the monkey centre-out data. Despite having aligned the latent dynamics only using half of the data, the impact on decoding performance is negligible.

Control analyses

Alignment of latent dynamics with random behavioural windows

To establish a 'behaviourally irrelevant' window as control, we randomly selected windows of similar length to our behavioural windows (450 ms) along the entire duration of the inter-trial and trial periods combined. This ensured we had samples of dynamics in the neural population with realistic statistics but that were not directly coupled to shared behaviour across individuals. We used this window to provide a lower-bound control for the alignment of neural population latent dynamics (Figure 2.2F,G,J, 2.3D,E, 2.4B,E, S2.3, S2.2G, S2.2B-D, S2.8D).

Aligning latent dynamics through Procrustes analysis

We used CCA to align the latent dynamics in all the analyses. However, to ensure that our results hold regardless of the specific method used for alignment, we replicated the main

result using Procrustes analysis[124]. Procrustes finds the best transformation that minimises the sum of the squares of the differences between the two input datasets. Following a procedure identical to the CCA analysis, we aligned the dynamics from two different datasets using Procrustes analysis (the `scipy.spatial.procrustes` class in Ref. 120) and then correlated the aligned dynamics to yield a metric comparable to that of the CCA (Figure S2.2G,H). Note that the degree of preservation of latent dynamics obtained with CCA and Procrustes analysis is largely similar.

Neural variance explained by aligned latent dynamics

We measured the percentage of neural variance explained by the preserved latent dynamics using a method we devised in Ref. 44. Briefly, we 'reconstructed' the preserved neural activity by projecting the aligned latent dynamics along the CC axes back to the PCA space and then to the original neural state space. We then measured the difference between the total neural variance and the variance of these reconstructed signals using an approach similar to that in Ref. 125. By repeating this procedure iteratively for an increasing number of manifold dimensions m, we measured the neural variance explained by each dimension of the aligned latent dynamics. Using this approach we found that preserved latent dynamics explain a significant fraction of the neural population variance (Figure S2.2E).

Surrogate datasets with Tensor Maximum Entropy

We established a lower bound control by aligning the latent dynamics from randomly selected windows sampled across different task conditions and behavioural epochs (see above). In addition to this control, we also used Tensor Maximum Entropy (TME) to generate surrogate neural data as another lower-bound control[80]. TME produces surrogate data that preserves the second order statistics of the actual neural data (i.e., covariance across time, across neurons, or across experimental conditions) but are otherwise random (Figure S2.2A). Aligning these surrogate data through the same procedure as the original

data shows significantly lower correlations for monkey centre-out task, random-walk task and mouse reaching and pulling task (Figure S2.2B-D).

Aligning topological structure in neural population activity

To test whether the topological structure in the produced movements is sufficient to produce preserved latent dynamics, we quantified the degree of similarity in latent dynamics across individuals that could be uncovered when aligning the static, topological features of the neural population activity, rather than the dynamics of the movements, using a technique developed in Ref. 43. To align the topological structure of neural population activity, we time-averaged the activity for each neuron during the execution epoch of each trial in the monkey center-out reaching task. We then analysed the time-averaged data with the previous methodology by performing PCA to find a neural manifold and using CCA to align each pair of sessions (Figure S2.7A). This procedure led to well-aligned 'topological representations' (example in Figure S2.7B). To directly test whether aligning the topological structure of neural population activity is sufficient to uncover preserved latent dynamics, we projected the latent dynamics on the CC axes found through this (static) topological alignment, and calculated the pairwise correlations of the resultant projected latent dynamics. These correlations were remarkably lower than those obtained through alignment of the time-varying latent dynamics (Figure S2.7C,D).

Control analyses on the numbers of conditions and neurons

To establish that the preservation of latent dynamics holds across different degrees of task complexity, we calculated the correlations for increasing numbers of subsampled target conditions for each pair of sessions in the monkey random target task (Figure 2.3F, Figure S2.8B). We randomly subsampled different combinations of target conditions and calculated the degree of preservation of the latent dynamics for up to 10,000 combinations for each number of conditions.

To establish that preserved latent dynamics can be uncovered regardless of the specific measured neurons, we also calculated the correlations for decreasing numbers of neurons in the random target task (Figure 2.3G, Figure S2.8C). For each pair of sessions, we randomly dropped neurons in increments of 10 until we ran out of measured neurons for either session, and repeated this process 50 times, calculating the degree of preservation at each step. For both analyses, we calculated the mean correlations for the top 4 CCs across all subsamples for each pair of sessions.

Comparison of different but related tasks

The central hypothesis of this study is that preserved latent dynamics are the basis for the generation of similar behaviour across individuals from the same species. Here, we sought to further support this hypothesis by showing that the latent dynamics produced by two individuals engaged in the same task are less similar than the latent dynamics produced by the same individual performing two different but related tasks. To this end, we compared our results to our previous study on the relationship of neural population activity underlying different but related wrist manipulation (Figure S2.9) or reach-to-grasp tasks in monkeys[44].

Recurrent neural network models

Model Architecture

To show that the preservation of latent dynamics across animals engaged in the same task is not a trivial byproduct of similar behaviour, we trained recurrent neural networks (RNNs) to perform the same centre-out reaching task as the monkeys. These models were implemented using Pytorch[123]. Similar to previous studies simulating motor cortical dynamics during reaching[49,126–128], we implemented the dynamical system $\dot{\mathbf{x}} = F(\mathbf{x}, \mathbf{s})$ to describe the RNN dynamics:

$$\tau \dot{x}_i(t) = -x_i + \sum_{j=1}^{N} J_{ij} r_j(t) + \sum_{k=1}^{I} B_{ik} s_k(t) + b_i + \eta_i(t)$$

where x_i is the hidden state of the ith unit and r_i is the corresponding firing rate following *tanh* activation of x_i. All networks had $N = 300$ units and $I = 3$ inputs, a time constant $\tau = 0.05s$, and an integration time step $dt = 0.01s$. The noise η was randomly sampled from the Gaussian distribution $\mathcal{N}(0, 0.2)$ for each time step. Each unit had an offset bias, b_i, which was initially set to zero. The initial states $x_{t=0}$ were sampled from the uniform random distribution $\mathcal{U}(-0.2, 0.2)$. All networks were fully recurrently connected, with the recurrent weights J_{ij} initially sampled from the Gaussian distribution $\mathcal{N}(0, \frac{g}{\sqrt{N}})$, where $g = 1.2$. The time-dependent inputs s fed into the network had input weights B initially sampled from the uniform distribution $\mathcal{U}(-0.1, 0.1)$. These inputs consisted of a one-dimensional fixation signal which started at 2 and went to 0 at the go cue, and a target signal that remained at 0 until the visual cue was presented. The two-dimensional target signal $(2\,cos\,\theta^{target}, 2\,sin\,\theta^{target})$ specified the reaching direction θ^{target} of the target.

The networks were trained to produce two-dimensional outputs p corresponding to x and y positions of the experimentally recorded reach trajectories, which were read-out via the linear mapping:

$$p_i(t) = \sum_{k=1}^{N} W_{ik} r_k(t)$$

where the output weights W were sampled from the uniform distribution $\mathcal{U}(-0.1, 0.1)$.

Model training

Networks were trained to generate positions of reach trajectories using the Adam optimiser[129] with a learning rate $l = 0.0005$, first moment estimates decay rate $\beta_1 = 0.9$, second moment estimates decay rate $\beta_2 = 0.999$, and epsilon $\epsilon = 1e - 8$. The loss function L was defined as the MSE between the two-dimensional output and the target positions over each time step t, with the total number of time steps $T = 400$. The first 50 time steps were

not included to allow network dynamics to relax:

$$L = \frac{1}{2B(T-50)} \sum_{b=1}^{B} \sum_{t=50}^{T} \sum_{d=1,2} (p_d^{target}(b,t) - p_d^{output}(b,t))^2.$$

To examine whether two networks could have different latent dynamics while producing the same motor output, we devised a network with additional constraints to perform the behavioural task with distinct latent dynamics (Figure 2.5A). We added a loss term that penalised the CC between the latent dynamics of the 'constrained' network being trained and those of another previously trained 'standard' network during movement execution:

$$L_{constrained} = L + \alpha \sum_{i=1}^{4} c_i^2$$

where c_i is the i-th CC. To examine different degrees of preserved latent dynamics, we trained the networks at varying values of $\alpha = 0$, 0.05, 0.25, or 0.50.

Networks were trained until the average loss of the last ten training trials fell below a threshold of 0.2cm^2, for at least 50 and up to 500 training trials, with a batch size $B = 64$. Each batch had equal numbers of trials for each reach direction. We clipped the gradient norm at 0.2 before the optimization step. Both standard and constrained training were performed on ten different networks initialised from different random seeds. The same set of random seeds were used for constrained networks at different values of α.

Connectivity analyses

By increasing the value of α, we were able to decrease the preservation of the latent dynamics while keeping behavioural performance constant. To examine how this changed the underlying connectivity, we calculated the variance and dimensionality of the weight changes in the recurrent weights J following training (Figure 2.5F,G).

Statistics and reproducibility

We compared the performance of various within-animal and across-animal movement decoders and classifiers using two-sided Wilcoxon's rank sum tests. We replicated the core findings across two species (mice and non-human primates), four behaviors (a centre-out reaching task, a sequential reaching task, and a reach, grasp and pull task, along with during covert movement planning), and two brain regions (motor cortex and dorsolateral striatum). Experiments on each species were performed independently in two different laboratories, and by different scientists. The mice experiments were done in a single cohort, whereas the monkey data was collected in two sets of experiments (one for the centre-out task, another for the random reaching task), each spanning two years. Overall, our neural recordings and behavioural data are in good agreement with related published studies. All attempts at replication were successful.

Code availability

All analyses were implemented using custom python code and using open-source software. All the result panels are reproducible by running Jupyter notebooks. Code to reproduce all the results is openly available in `https://github.com/BeNeuroLab/2022-preserved-dynamics`.

Data availability

The majority of the monkey datasets are publicly available on Dryad (`https://datadryad.org/stash/dataset/doi:10.5061/dryad.xd2547dkt`) and CRCNS (`http://dx.doi.org/10.6080/K0FT8J72`). The remaining monkey datasets and the mouse datasets will be made available upon reasonable request.

2.6 Supplementary Figures

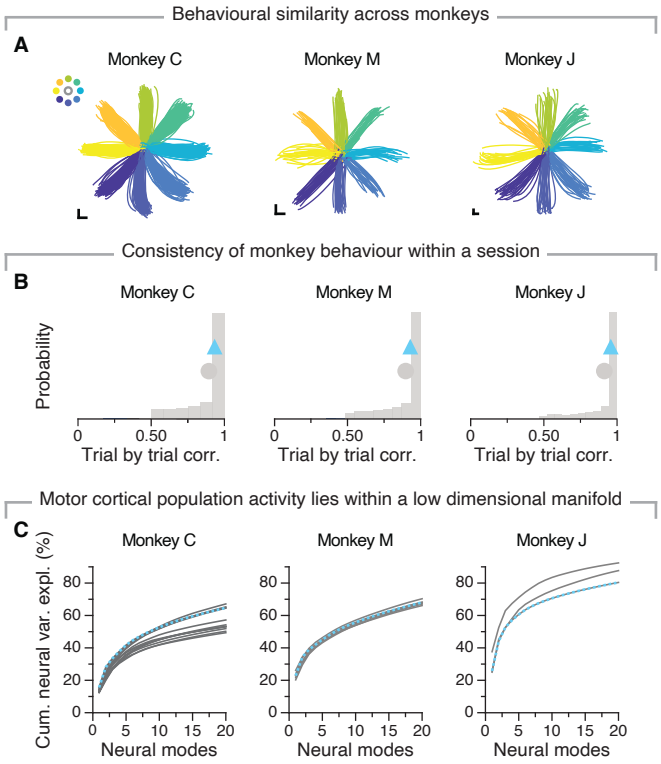


Figure S2.1: Behavioural and neural data from monkeys performing a centre-out reaching task. **A**) Hand trajectories of a single session of every monkey included in this work. Each line represents one trial, colour coded by target (inset). Scale bar, 1 cm. **B**) Distribution of the behavioural correlations for all pairs of trials from every included session for each of the 3 monkeys. Grey circle, mean. Blue triangle, mean for the representative session shown in Panel A. **C**) Cumulative neural variance explained as a function of the number of neural modes included. Data includes 12 sessions for monkey C, 6 sessions for monkey M, and 3 sessions for monkey J. Each line, one session. Dashed blue line, the same session shown in Panel A.

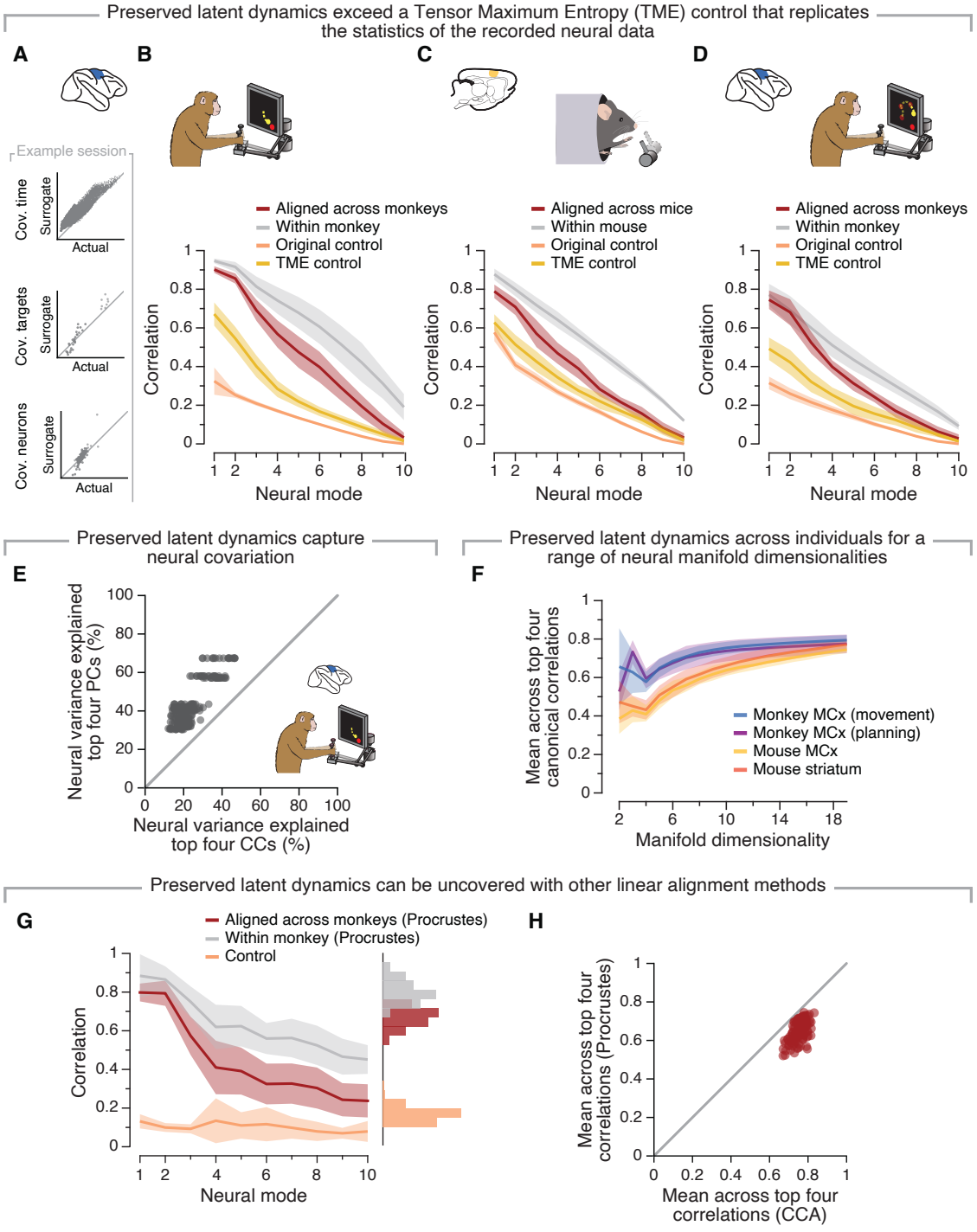


Figure S2.2: A broad range of techniques uncover preserved latent dynamics that capture dominant features of neural population activity. A) We used Tensor Maximum Entropy (TME) to generate surrogate data that preserved all the second order statistics of the actual neural data. Element-wise conservation of the covariance across time, targets and neurons between actual and surrogate data for one example monkey centre-out dataset. **B)** Similarity in latent dynamics across all pairs of sessions from three different monkeys performing the centre-out reaching task. Data include all 21 sessions and 126 pairs of comparisons shown in Figure 2.2G. Note that the correlation between aligned latent dynamics across monkeys (red) was quite close to the within-monkey

correlations (grey), and largely exceeded lower-bound controls from both randomly sampled behavioral epochs (orange) and TME (mustard). Line and shaded area, mean±s.d. **C**) Similar to Panel B but for mouse motor cortex during the reaching and pulling task. Data include all six sessions across four different mice shown in Figure 2.2J. Line and shaded area, mean±s.d. **D**) Similar to Panel B but for the monkey motor cortex during the sequential reaching task. Data include all 11 sessions and ten comparisons across three different monkeys shown in Figure 2.3E. Line and shaded area, mean±s.d. **E**) Preserved latent dynamics capture dominant features of neural population activity. Comparison of the total neural variance explained by the top four PCs and the top four CCs. Single dots, individual comparisons for all monkey centre-out datasets. **F**) Preservation of latent dynamics does not depend on the number of manifold dimensions. Average canonical correlation as function of the number of manifold dimensions (principal components) for each of the four datasets considered (legend). Data pooled across all comparisons separately for each dataset. Line and shaded area, mean±s.d. across the top four canonical correlations. MCx, motor cortex. Line and shaded area, mean±s.d. **G**) Preserved latent dynamics can be uncovered with a different lineal alignment method, Procrustes analysis (details in Methods). Data presented as in Figure 2.2G but using Procrustes analysis rather than CCA to align the latent dynamics. Line and shaded area, mean±s.d. **H**) Pairwise comparison between CCA and Procrustes analysis. Single dots, individual comparisons. Data includes 21 sessions and 126 comparisons across three monkeys, and six sessions and 13 comparisons across four mice.

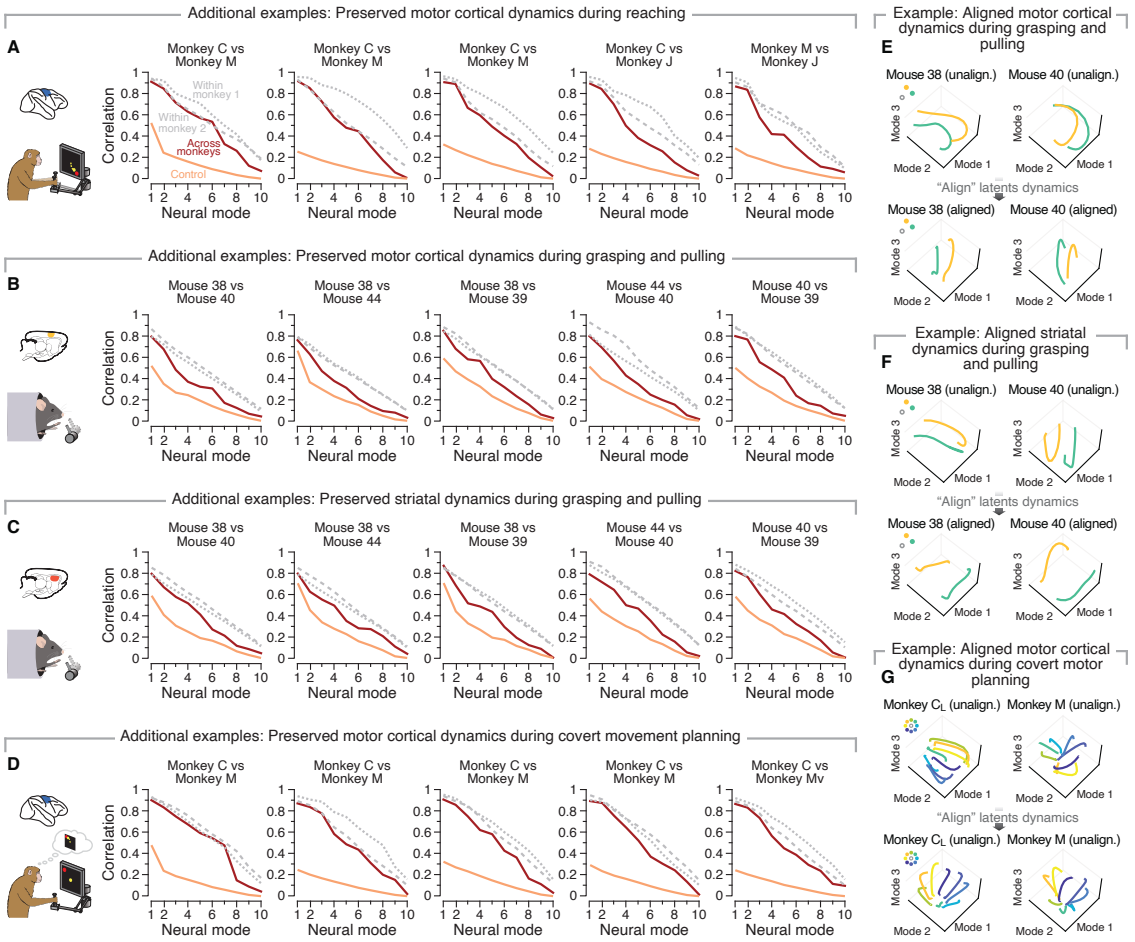


Figure S2.3: Additional examples of canonical correlations, and preserved latent dynamics between pairs of sessions from two different monkeys and mice. A) Example canonical correlations between motor cortical latent dynamics for monkeys during movement execution. Shown are five of the individual comparisons included in the pooled results in Figure 2.2D. **B**) Example canonical correlations between motor cortical latent dynamics for mice during movement execution. Shown are five of the individual comparisons included in the pooled results in Figure 2.3B. **C**) Same as Panel B for mouse striatum. **D**) Same as Panel A but during movement planning. **E**) Trajectories described by the motor cortical latent dynamics before (left) and after (right) alignment for a representative pair of sessions from two different mice. Individual lines, mean latent trajectory to each target (colour coded as in the inset). **F**) Similar to Panel E, for striatal latent dynamics in mice. **G**) Similar to Panel E, for motor cortical latent dynamics during covert movement preparation in monkeys.

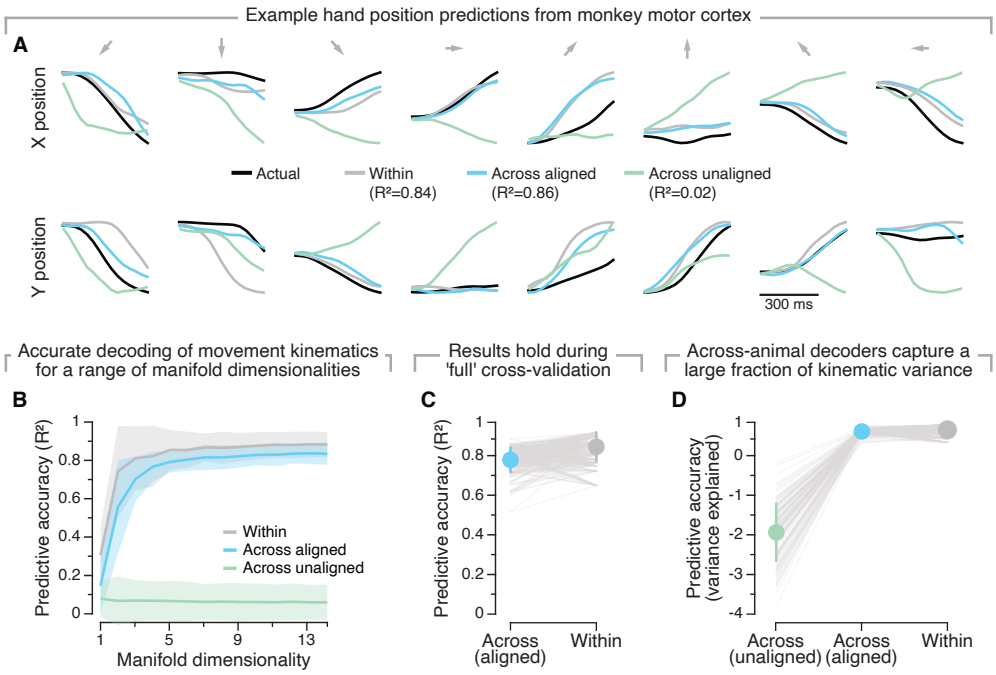


Figure S2.4: Preserved latent dynamics across monkeys are behaviourally relevant. A) Example kinematic predictions across animals. Actual and decoded hand trajectory during one reach to each of the eight targets along the X (top) and Y directions (bottom); arrows on top of each column indicate target direction. The figure shows predictions using three models along with their respective accuracy (legend). Data from the same representative session shown in Figure 2.2B. Note that a decoder trained on a different monkey after alignment of their latent dynamics ('Across aligned') performed virtually as well as a decoder trained and tested on the same session from the monkey being tested ('Within'). **B)** Decoding of hand position based on the latent dynamics for manifolds with increasing dimensionality. Line and shaded area, mean±s.d. **C)** Across-animal decoding accuracy is not significantly impacted by a 'full' cross-validation procedure in which we used three sets of trials for training and testing (one for alignment, one for decoder training, and one for testing). Errorbars, mean±s.d. **D)** Aligned latent dynamics also allow for accurate across-animal decoding of movement kinematics according to a variance explained metric. Data points, individual comparisons between two sessions from different monkeys ($n = 126$). Errorbars, mean±s.d.

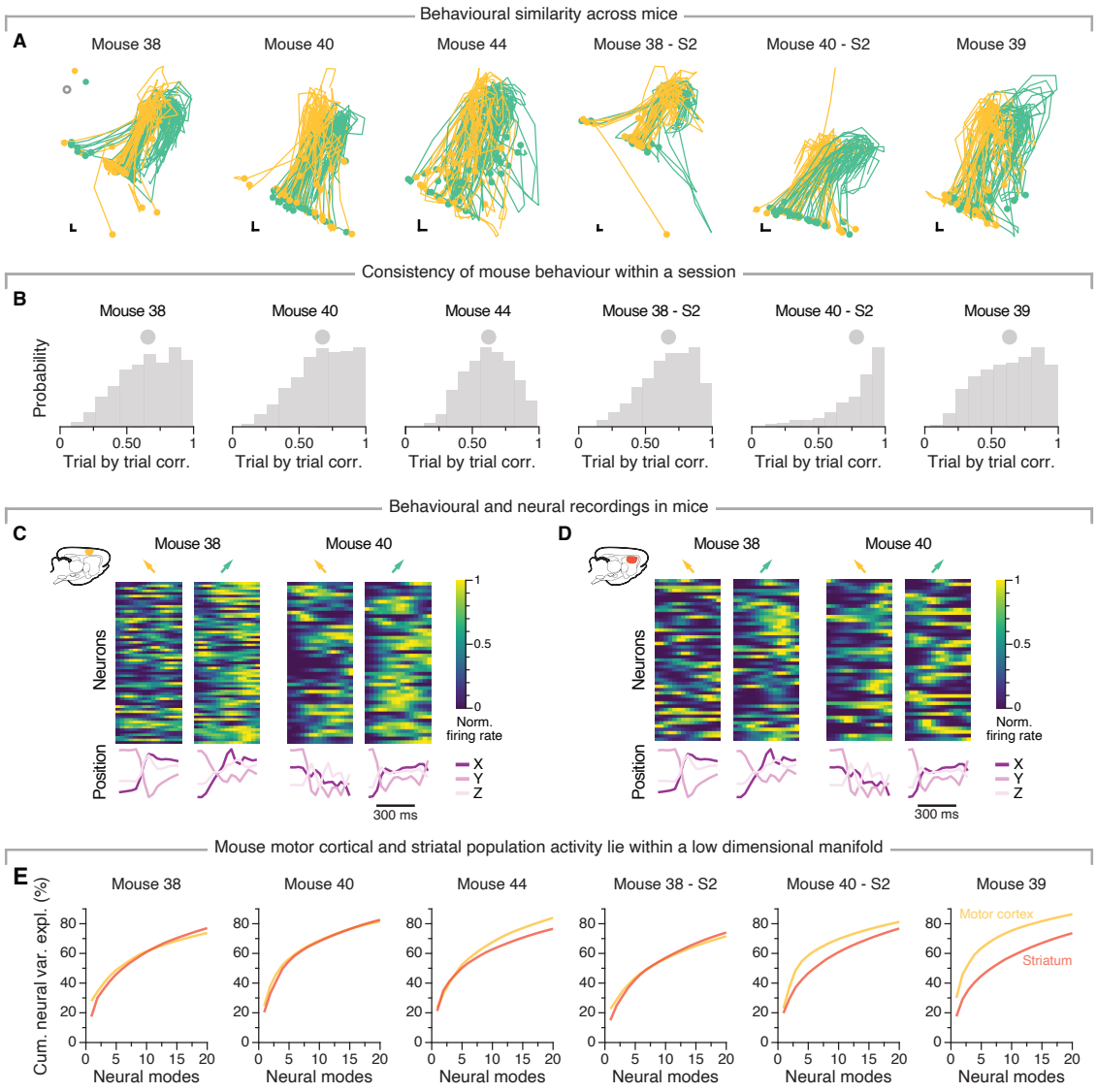


Figure S2.5: Behaviour and neural activity for mice performing a grasping and pulling task. A) Hand trajectories of every session included in this work (notice that two mice have two sessions each). Each line represents one trial, colour coded by target (left, yellow; green, right; see inset). Circle, initial hand position. Scale bar, 1 mm. **B**) Distribution of the behavioural correlations for all pairs of trials from every session. Grey circle, mean. **C**) Example normalised motor cortical firing rates aligned to movement onset for two different mice (top) and corresponding hand trajectories (bottom). Each column, one reach to each of the two directions. Arrows on top of each column, reach direction, colour-coded as in Panel A. **D**) Similar to C but showing the striatal activity during the same example movements. **E**) Cumulative neural variance explained as a function of the number of neural modes included, colour coded by brain region.

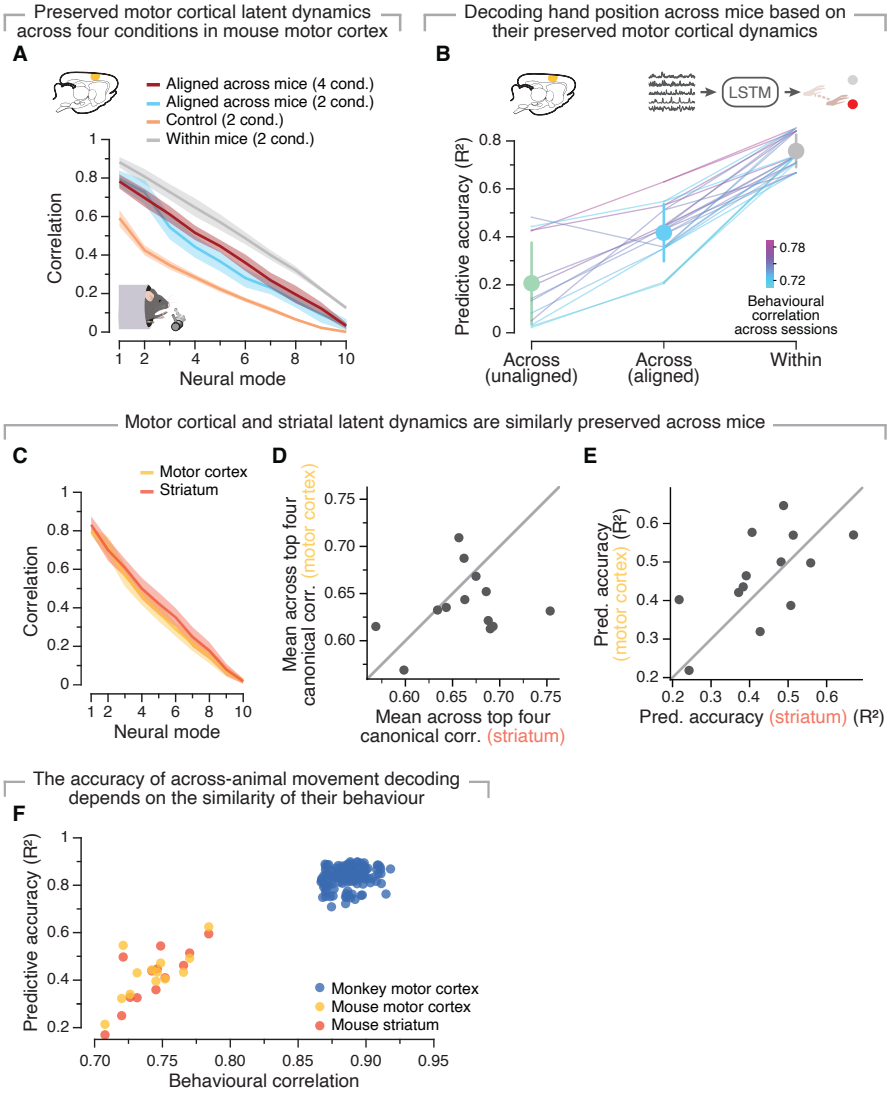


Figure S2.6: Additional data: Preservation of motor cortical and striatal latent dynamics in mice engaged in a grasping and pulling task, and relationship between the accuracy of across-animal decoders and behavioural similarity. A) Preservation of motor cortical latent dynamics during the four conditions (two reach directions $\times$ two loads) of the peri-grasp interval of the mouse grasping and pulling task (red) extends the results during the two reaching conditions (blue) from Figure 2.2J. The lower bound control (orange) and the within-animal upper bound (grey) from Figure 2.2J are shown here for reference. Data include six sessions across four different mice. Line and shaded area, mean$\pm$s.d. **B)** Decoders trained on the aligned latent dynamics from one mouse allowed predicting hand position from a different mouse with reasonable accuracy (blue), considerably better than decoders based on the unaligned latent dynamics (green). Lines, individual comparisons between two sessions from different mice ($n = 6$ sessions and $n = 13$ pairs), colour coded based on the correlation in hand kinematics across them (legend). Errorbars, mean$\pm$s.d. **C-E)** Motor cortical and striatal latent dynamics are similarly preserved during the mouse pulling and grasping task, and allow for comparable decoding of movement kinematics. The across-animal motor cortical (yellow) and striatal (orange) correlations averaged across all comparisons (reproduced from Figure 2.2J and Figure 2.4B) are largely overlapping (Panel D). Line and shaded area, mean$\pm$s.d. Pairwise comparison of the degree of preservation of the latent dynamics (Panel D) and across-animal decoder accuracy (Panel E) for motor cortex and striatum. Single dots, pairs of sessions from two

different mice ($n = 6$ sessions and $n = 13$ pairs). **F)** Decoding performance based on the aligned latent dynamics for each pair of sessions for the three stereotypical movement execution datasets (legend) as function of their mean behavioural correlation. Similar to Figure 2.2K. Dots, individual comparisons across pairs of sessions that includes 21 sessions and 126 comparisons across three monkeys, and six sessions and 13 comparisons across four mice.

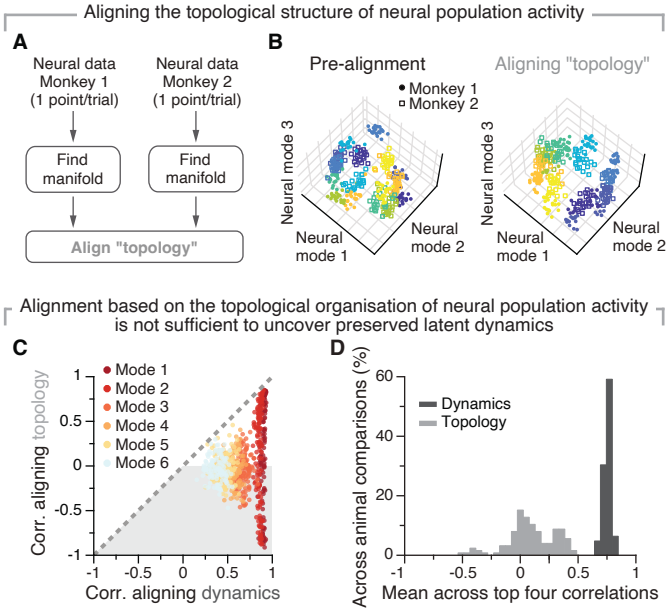


Figure S2.7: Aligning based on the topological organisation of the neural data is not sufficient to uncover preserved latent dynamics across individuals. A) To align the latent dynamics across sessions based on the topological organisation of the population activity, we performed CCA on the population activity averaged across all time points on each trial. This process created a distribution of points with clear topological structure corresponding to the different reaching directions. **B)** This method provided excellent alignment of the topology on a single trial basis. Before alignment (left), the projection of each trial onto the neural manifold from each monkey (shown here overlaid on top of each other) looked different. After alignment, similar target-specific structure is present across both animals. Individual markers, a reach to one of the eight targets (colour code in inset) for Monkey C_L (closed circles) and Monkey M (open squares). **C)** Pairwise comparisons of the correlations after projecting the latent dynamics onto the CC axes found by aligning the topology (vertical axis) and onto the CC axes found by aligning the latent dynamics (horizontal axis). Data shown for the top six neural modes (see legend for colour code). Each dot represents one comparison; data include 21 sessions and 126 pairs of comparisons from three different monkeys during the centre-out reaching task. All dots lie well below the diagonal (dashed grey), indicating that aligning the latent dynamics based on the topology does not reach the correlation values obtained by aligning the latent dynamics. **D)** Correlation values were significantly lower when the alignment was based on preserving the topological structure of the neural population activity rather than its latent dynamics, illustrating the importance of the precise temporal dynamics for uncovering preserved latent dynamics. Same data as Panel C.

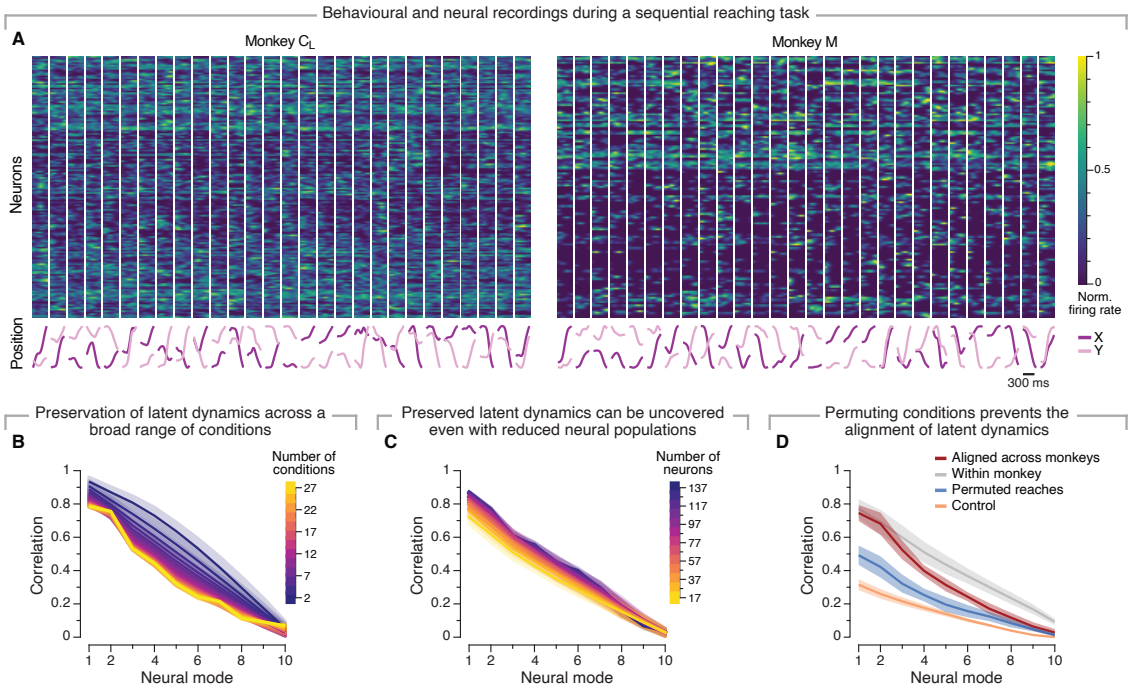


Figure S2.8: Behavioural and neural data from monkeys performing a sequential reaching task, and additional data on the preservation of latent dynamics. **A**) Example normalised neural firing rates aligned to movement onset for two different monkeys (top) and corresponding hand trajectories (bottom). Each column, one reach to each of the 28 conditions, which we defined based on the starting location, duration, and direction of the reaches. **B**) Canonical correlations as function of the number of conditions considered (legend). Note that the traces line up well after ∼9 conditions are considered. Data from the same example dataset as in Panel A. **C**) Canonical correlations as function of the number of neurons (legend). Note that the traces line up very well after ∼57 neurons are considered. Data from the same example dataset as in Panel A. **D**) Permuting the condition labels (blue) considerably decreases the alignment of latent dynamics across animals, bringing the canonical correlations closer to the lower bound control. Data from all 11 sessions and ten comparisons across three different monkeys. Line and shaded area, mean±s.d.

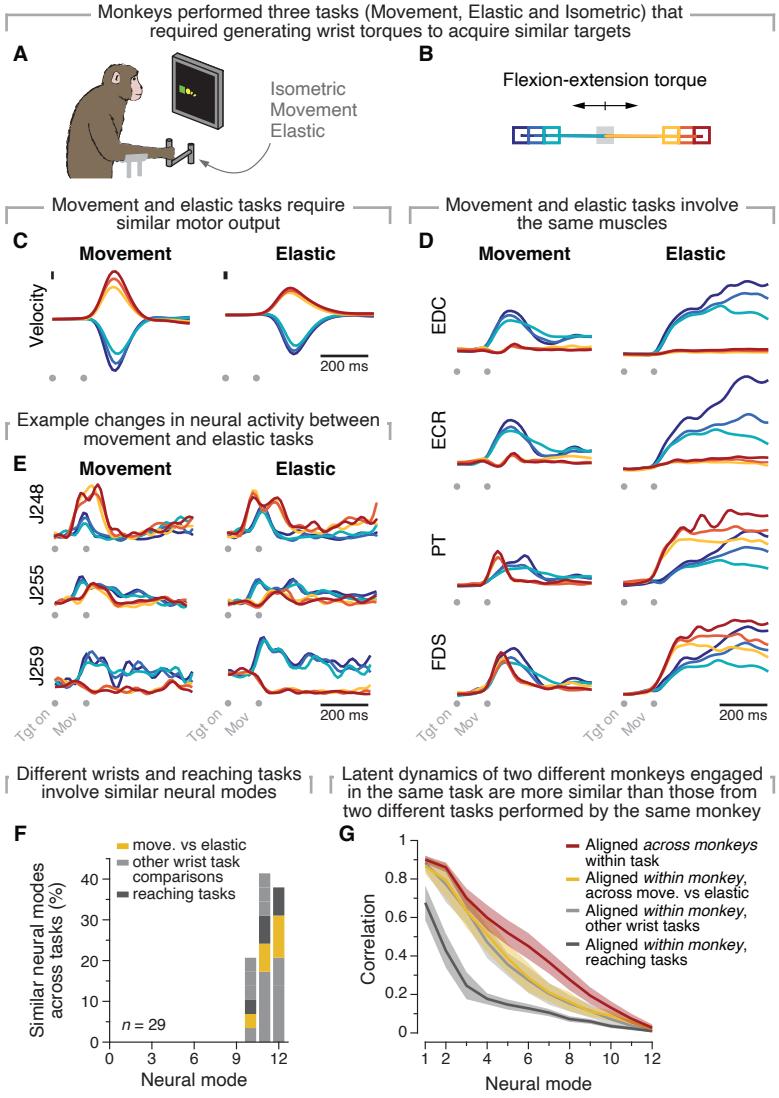


Figure S2.9: The latent dynamics of two different monkeys performing the same task are more correlated than those of the same monkey performing two related tasks. A) We compared the latent dynamics produced by monkeys performing three related visually-guided one-dimensional wrist manipulation tasks using the same manipulandum. We also compared two reaching and grasping tasks (details in the Methods). **B)** The organisation of the targets and the produced cursor trajectories were the same across all three wrist manipulation tasks (note the overlapping cursor trajectories). **C-E)** Here, we show behavioural and neural data from the two wrist manipulation tasks that had the most similar kinematics (Panel C; lines color-couded by targets according to Panel B): the one-dimensional movement task, and the one-dimensional spring-loaded ("elastic") movement task. These two tasks engaged the same muscles (Panel D), and the activity of single units was also relatively similar across them (Panel F shows three example units). **F)** The covariance patterns elicited by these tasks as well as the rest of wrist and reaching tasks was also quite similar across behaviours (this figure shows the number of neural modes out of twelve that are well-aligned across tasks according to a principal angle analysis). **G)** Despite the similarities in task characteristics (Panel B), motor output (Panel C), engaged muscles (Panel D), and even single unit tuning (Panel E), the latent dynamics generated by the same monkey as it performed different wrist tasks (mustard and light grey) or different reaching task (dark grey) within the same day were less similar than those of two different monkeys engaged in the same task (red). The across

monkey results (red) illustrate preservation of the latent dynamics as monkeys performed the same centre-out reaching task; the results are the same as in Figure 2.2E only recalculated with 12 components to match the across-task analysis in Gallego *et al* .[44]. Lines and shaded area, mean±s.d. across the comparisons indicated in the legend. Panels A-G are adapted from Ref. 44; the across-task alignment data in Panel G partly reproduces Supplementary Figure 7 therein.

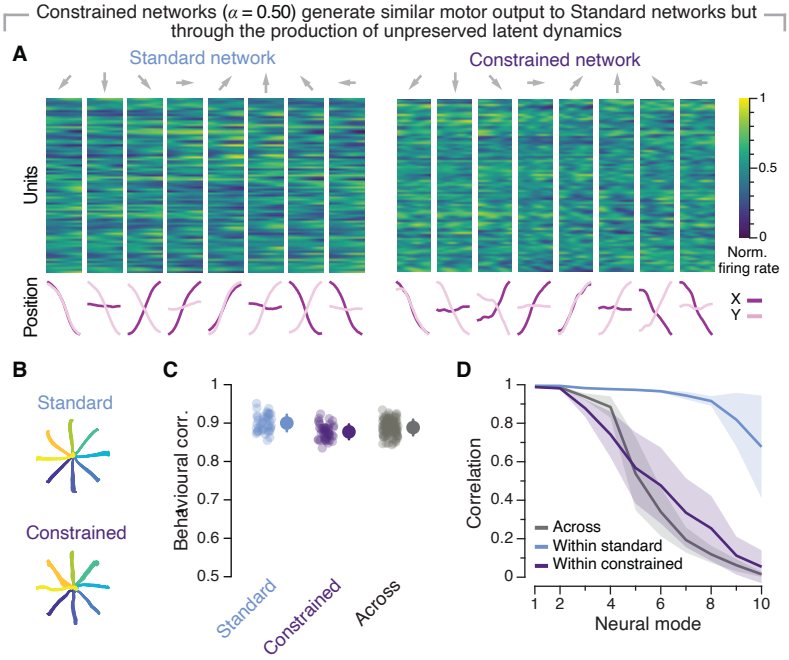


Figure S2.10: Additional data: Behavioural similarity is necessary but not sufficient for the preservation of latent dynamics. A) Behavioral output and network dynamics for each reach direction for example Standard (left) and Constrained (right) networks that were penalised for canonical correlation similarity to the latent dynamics of Standard networks. Results presented as in Figure 2.2B. Data for constrained networks represents results for $\alpha = 0.50$. **B)** Behavioural output produced by the two types of RNNs. The Standard (top) and Constrained (bottom) networks produce similar output. **C)** Behavioural correlation within and across pairs of Standard and Constrained networks. Data show 10 networks initialized from different random seeds for each type of network, 45 comparisons within network types, 100 comparisons across network types. Markers, individual comparisons. Line and shaded surface, mean $\pm$ s.d. **D)** Similarity in latent dynamics between pairs of Standard networks (blue) and Constrained networks (purple), and across network types (grey). Note that Constrained networks have correlations that are much more dissimilar than Standard networks. Likewise, the correlation between the latent dynamics of pairs Standard and Constrained networks are much lower than those between pairs of Standard networks. Thus, behavioural similarity is not sufficient to obtain highly correlated latent dynamics. Line and shaded area, mean$\pm$s.d.

Chapter 3

Structured latent dynamics in motor learning

Having shown the stability of latent dynamics across animals during motor control, we next examine the flexiblity of population dynamics during motor learning. How do animals quickly adjust known movements to ongoing perturbations in the environment? In this chapter, we explore how *de novo* learning of different motor repertoire affects the flexibility of subsequent motor adaptation. Using neural network models, we show that different learning experiences shape the structure of the latent dynamics in neural state space, ultimately determining the population patterns that are easily accessible for adaptation.

The work in this chapter is presented in an article format as published in *Nature Communications* in 2024[128]. I would like to thank my supervisors Claudia Clopath and Juan A. Gallego for their guidance and inspiration. Furthermore, I would like to thank our experimental collaborators Matthew G. Perich and Lee E. Miller for providing their data and insight.

De novo motor learning creates structure in neural activity that shapes adaptation

3.1 Abstract

Animals can quickly adapt learned movements to external perturbations, and their existing motor repertoire likely influences their ease of adaptation. Long-term learning causes lasting changes in neural connectivity, which shapes the activity patterns that can be produced during adaptation. Here, we examined how a neural population's existing activity patterns, acquired through *de novo* learning, affect subsequent adaptation by modeling motor cortical neural population dynamics with recurrent neural networks. We trained networks on different motor repertoires comprising varying numbers of movements, which they acquired following various learning experiences. Networks with multiple movements had more constrained and robust dynamics, which were associated with more defined neural 'structure'—organization in the available population activity patterns. This structure facilitated adaptation, but only when the changes imposed by the perturbation were congruent with the organization of the network inputs and the structure in neural activity space acquired during *de novo* learning. These results highlight trade-offs in skill acquisition and demonstrate how different learning experiences can shape the geometrical properties of neural population activity and subsequent adaptation.

3.2 Introduction

From walking to grasping objects, movement enables us to interact with the world. Mastering a skill requires many hours of practice, be it during development or in adulthood. In contrast to long-term skill learning, adapting existing skills to environmental perturbations is a much faster process: after learning to ride a bike, adapting to foggy weather or uneven roads is much easier. Existing motor repertoires acquired through long-term learning likely form the foundation for short-term motor adaptation, but it is unclear how different repertoires can affect adaptation, even for common experimental perturbations like visuomotor rotations (VR)[130] or force fields[131]. This lack of understanding stems from the experimental challenge of characterising an animal's entire behavioural repertoire learned throughout their lifetime, and implies that the interplay between available neural activity patterns and rapid adaptation remains largely unknown[132].

Recent work has focused on the coordinated activity of neural populations to begin to shed light on the neural basis of motor adaptation[14,37,94]. In this neural population view, brain function is not build upon the independent activity of single neurons, but rather on specific patterns of neural co-variation (from now on, simply 'activity patterns')[13,54,133,134]. In practice, these activity patterns can be examined by building a neural state space (from here, referred to as 'neural space') where each point denotes the state of the neural population. Numerous studies (e.g. Refs. 36,44–46,64,135) have found that the activity of even hundreds of simultaneously recorded motor cortical neurons is well captured by relatively few population-wide activity patterns, an observation consistent with neural population activity being constrained to a low-dimensional surface—a 'neural manifold'—that can be estimated by applying a dimensionality reduction method[33,136–139]. Studies comparing neural population activity patterns evoked by sensory stimuli and optogenetic stimulation provide direct evidence of neural manifolds reflecting circuit connectivity constraints[60,61]. For example, optogenetic stimulation of a small subset of stimulus selective neurons evokes the same activity patterns in the manifold of a population of simultaneously recorded neurons as those observed during 'natural' stimulus presentation[60], indicating

that neural manifolds are at least partly shaped by the underlying connectivity of the network. This presumed relationship between circuit constraints and neural manifold geometry is further supported by the preservation of neural manifolds in the anterior thalamus[138,140] and medial enthorinal cortex[137] between active behavior and sleep.

Previous work suggests that long-term learning causes changes in circuit connectivity[10–12,141], which may in turn change the geometrical properties of an existing neural manifold as well as the activity within it—the so-called 'latent dynamics'. In contrast, experimental and modeling studies of motor adaptation suggest that this form of rapid learning may be achieved without any changes in motor cortical connectivity or, if present, these changes would be very small and low-dimensional[14,15]. In good agreement with these findings, using a brain-computer interface that mapped motor cortical activity onto computer cursor movements, Sadtler et al. showed that it is easier to adapt to perturbations that require only neural states that lie within the existing neural manifold: while animals can learn to produce activity patterns within the existing manifold in a matter of minutes or hours[46], producing activity patterns outside the existing manifold takes several days[16]. To learn outside-manifold perturbations, new activity patterns need to be used, which may necessitate changing the synaptic connectivity of the circuit[16,142]. Combined, these results suggest that long-term motor learning changes the circuit connectivity to create new neural population latent dynamics whereas short-term motor adaptation may reuse existing ones[17].

Yet, studying the interplay between an animal's existing motor repertoire and its ability to adapt a known behaviour remains challenging due to the complex organization of existing motor skills within neural space. Similar behaviors (e.g., various wrist manipulations or grasping tasks) may share similarly oriented task-specific neural manifolds[44], while dissimilar behaviors (e.g., reaching and walking in mice) may have almost orthogonal manifolds even if they share similar movements[45]. Furthermore, experimentally, it is impossible to quantify the entire repertoire of motor skills that an animal has, which poses a challenge to investigate how its behavioral repertoire influences motor adaptation. Similarly, experimenters also often lack access to how skills were acquired in the first place,

which poses an additional challenge to understanding the neural basis of learning because the *de novo* learning process influences future learning[143,144], e.g., based on whether the training examples are meaningfully organized with respect to each other[145]. Finally, in virtually all motor adaptation studies, animals must adjust to perturbations on a specific laboratory task they have already learned, and adaptation is only examined with respect to a 'baseline period' (e.g. Refs. 14,37,46,94,130,131,146,147); this approach ignores, for practical reasons, the relationship between the adapted behavior and all the other motor skills the animal has previously acquired, as well as how these skills were initially learned.

To overcome the experimental challenges of assessing an animal's lifelong experience, here we examined how the existing motor repertoire can affect adaptation differently using recurrent neural networks (RNNs). Similar networks have been able to reproduce motor output and key features of latent dynamics from experimental recordings[49–51,148], including during adaptation[15,142]. We modeled the latent dynamics of the motor cortex during *de novo* learning and subsequent adaptation. We trained our networks on different repertoires with varying numbers of movements through a variety of learning experiences using movement trajectories modified from experimental recordings of monkey reaches[14,112]. We hypothesized that networks with larger motor repertoires would adapt to perturbations more easily since they are already able to produce a broader set of activity patterns, but only if perturbations require changes that are 'meaningful' with respect to their learning experience.

By systematically training networks in two stages comprising of *de novo* learning and subsequent adaptation, we found that larger repertoire networks could adapt to perturbations more quickly, but only under certain circumstances. The way the latent dynamics of multiple movements were organized in neural space shaped subsequent adaptation: adaptation was facilitated when the latent dynamics were organized in a way that was congruent with the changes required by the perturbation, and when only small changes in motor output were needed. This suggests that ease of adaptation is affected not only by its relation to the existing manifold[16,46], but also by the organization of the latent dynamics within it. Crucially, this organization is affected by past learning experiences. These obser-

vations highlight an inherent trade-off in skill-acquisition: mastery of more movements better defines the structure of the neural manifold. This, in turn, facilitates adaptation that requires small changes in behavior, but harms adaptation that requires large changes or when the learning experience is not directly relevant for the new behavior.

3.3 Results

Probing the impact of *de novo* learning on subsequent adaptation with RNNs

To understand how a neural population's existing activity patterns affect its ability to change its activity, we used RNNs to model motor cortical neural population dynamics following *de novo* learning and subsequent adaptation (Figure 3.1a).

To model *de novo* learning, we trained the networks to perform different repertoires with different numbers of movement directions. The movements were modified from experimentally recorded center-out reaches from monkeys (data from Perich et al.[14], Figure 3.1b, Methods). For each reach, monkeys were first presented with a visual target; after a variable delay period, a go cue indicated they could execute the movement. To address our first hypothesis that having a larger motor repertoire would facilitate adaptation, we used repertoires of different sizes, ranging from one to four movement directions (Figure 3.1c). Single movement repertoires comprised a single reach to -10° while multi-movement repertoires comprised movements in directions equally spaced between -10° and -50°, unless otherwise specified. Note that we intentionally limited our networks to four movements for simplicity to make our simulation results intelligible. Importantly, all networks learned to produce the -10° direction, allowing for comparisons across networks trained on different repertoires. We further kept all training parameters constant to simulate a common *de novo* learning experience. Networks were given the angular direction of these movements as inputs along with a 'go' cue (occurring 0.5–1.5 s after the direction cue) to mimic the instructed delay reaching task performed by the monkeys (Methods).

To model motor adaptation, we subsequently trained these networks to counteract a visuomotor rotation (VR) on the shared movement direction they had all learned (Figure 3.1d). VR is a commonly used experimental paradigm to examine motor adaptation[9,14,37,130,149,150] in which a rotational transformation is applied to the motor output. In this case, we applied a 10° rotation (counterclockwise), which the networks had to counteract by producing output in the opposite direction. By probing adaptation on only one common movement, we could assess how entire motor repertoires contribute to adaptation of a given movement, and compare the adaptation performance across repertoires.

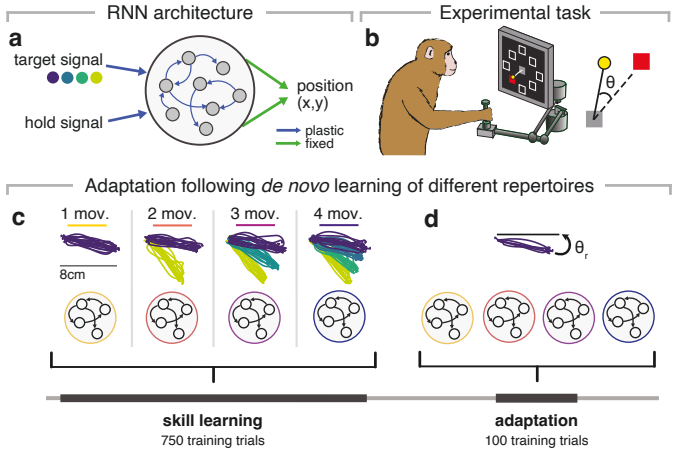


Figure 3.1: Probing *de novo* learning and adaptation with RNNs. a. RNNs were trained to produce 'hand positions' during a standard center-out reaching task as output. They were given a target signal that specifies the reach direction, and a hold signal that indicated movement initiation. The target signal indicated the angular direction, unless otherwise specified. The input and recurrent weights were learned (plastic) while the output weights were fixed, unless otherwise specified. **b**. RNNs were trained on hand trajectories modified from experimental recordings from monkeys performing a similar center-out reaching task (left). Subsequent adaptation was studied using a classic visuomotor rotation paradigm (VR, right), in which visual feedback is rotated by a fixed angle around the center of the workspace. **c**. Networks were trained on repertoires with different numbers of movements (from one to four) to model *de novo* learning. All multi-movement networks (2 mov., 3 mov. and 4 mov.) covered the same angular range. **d**. Networks were later trained to counteract VR perturbations for only the one common movement to understand the influence of existing motor repertoires on adaptation.

Multiple movements produce more constrained and robust dynamics that are structured in neural space

We first trained networks to produce repertoires comprising one to four different movements to understand the resulting differences in underlying network activity. Following the initial *de novo* learning phase, all the networks were able to learn each of the repertoires (Figure 3.2a) with comparable performance (Figure 3.2b), as quantified by the mean squared error between target movement trajectories and produced movement trajectories. This indicates that any differences in their ability to adapt will not be due to how accurately they can generate motor output, but rather to differences in the network dynamics that can be produced based on their acquired motor repertoires.

To compare across different repertoires, we examined the neural dynamics as the networks were producing the common -10°movement. We focused on network activity during both preparation and execution of the same target motor output. We calculated the population latent dynamics by projecting the activity of all 300 units in our networks onto a lower-dimensional neural manifold identified by performing Principal Component Analysis[13,33] (PCA) (Methods). The main differences in the activity produced by networks with different repertoires was between single-movement and multi-movement (two, three, or four movements) networks: multi-movement networks had less variance across trials to the shared target in both single unit activity (Figure 3.2c left, $P = 9.8 \cdot 10^{-4}$, Wilcoxon signed rank test; Figure S3.1a-c; Methods) and latent dynamics (Figure 3.2d, $P = 0.02$, Wilcoxon signed rank test; Figure S3.1d-f) than single-movement networks, suggesting that the network activity becomes more constrained when multiple movements need to be embedded in the neural space. Intriguingly, despite their lower variance in activity, multi-movement networks had greater variance in motor output (Figure 3.2e, $P = 9.8 \cdot 10^{-4}$, Wilcoxon signed rank test; Figure S3.1g-i; Figure S3.2)—which was more comparable to experimental reaches (Figure 3.2e)—even though all the networks were performing the same movement. How can these apparently contradictory observations be reconciled? We hypothesized that the more variable latent dynamics of single-movement networks may

lie on directions of activity space that do not affect motor output, i.e., those defining the 'output-null' subspace[35]. Separately computing the variance of the latent dynamics within the output-null subspace and the 'output-potent' subspace (i.e., the directions that affect the network output), confirmed this prediction: the greater variability of the latent dynamics of single-movement networks was largely confined to the output-null subspace, and thus did not lead to more variable motor output (Figure 3.2c right; Methods). These results also held when networks were trained on synthetic reaches that were more stereotyped than actual monkey reaches (Figure S3.3) and across different ways of calculating inter-trial variability (Figure S3.3; Methods). Thus, greater constraints in the network dynamics were characteristic of multi-movement networks, but these more constrained dynamics did not necessarily translate into more consistent behavioural output.

How may these constraints affect the ability of multi-movement networks to generate robust latent dynamics driving motor output? We predicted that they may lead to greater robustness against noise[151]. We examined this by changing the amount of simulated noise applied to the unit activity, and saw that multi-movement networks were indeed more robust against higher noise levels than single-movement networks (Figure 3.2f, Figure S3.4g). To probe this further, we increased the noise five-fold compared to pre-training and measured how the output was affected (examples in Figure 3.2g and Figure S3.4a,b,d,e, with the two-movement network representative of multi-movement networks). Intriguingly, while motor output for single-movement networks circled back and became twisted, that for two-movement networks shifted toward previously learned outputs at this high noise level (Figure 3.2g, Figure S3.4e). This systematic shift indicates that learning multiple movements creates structure in the neural activity space that maps to structure in the motor output; that is, the trajectories described by the latent dynamics driving each movement are organized in neural space in a way that is congruent with that of the movements. With this underlying 'structure', noise in the activity space caused two-movement networks to explore other activity states that led to movements intermediate to those that had been previously learned (Figure 3.2g, Figure S3.4b,e). This led to mostly linear output trajectories with greater forward movement, as quantified by the velocity in the x direction

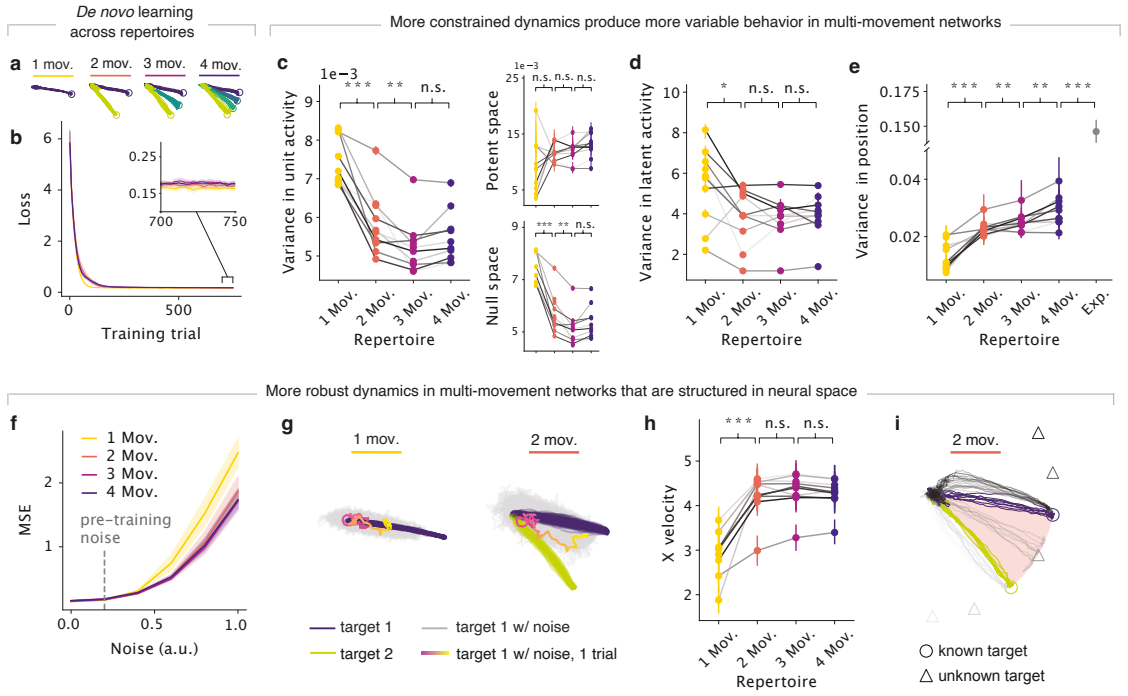


Figure 3.2: Networks that have acquired multiple movements produce more constrained and robust dynamics that are structured in neural space. **a**. Motor output of networks trained on different repertoires following *de novo* learning. **b**. Loss during initial *de novo* learning. Loss was calculated as the mean-squared error between the network output and target positions. Line and shaded surfaces, smoothed mean and 95% confidence interval across networks of different seeds ($n = 10$ random seeds). **c-f, h**. Following *de novo* learning, networks were tested on the one shared movement. **c**. Variance in unit activity for networks trained on different repertories. Variance is calculated for each unit for each timestep across trials. Inset: Variance in the output-null subspace and output-potent subspace of the unit activity with respect to the produced output. Individual lines, different random seeds ($n = 10$); circles and error bars, median and 95% confidence intervals with bootstrapping. *** denotes 0.001, ** 0.01, * 0.05 for one-sided Wilcoxon signed rank tests. **d**. Same as Panel c, but for the variance in latent dynamics. **e**. Same as Panel c but for variance in motor output. Variance for single reaches for monkeys trained on the center-out reach task (data from Ref. 14) is included for comparison. Note that the networks, which only know a few movements, generally have less variance than monkeys that know many more movements. **f**. Mean-squared error of output when noise of increasing magnitude is added to the neural activity. Line and shaded area, median and 95% confidence interval. **g**. Motor output with (grey, pink gradient) and without (purple, green) increased noise added ($\eta = 1$) for example networks with one- and two-movement repertoires. Color gradient, time-course of the movement execution (legend). **h**. Same as Panels c–e but for velocity in the x direction for motor output when noise ($\eta = 1$) is added as per Panel g. **i**. Motor output for different target cues for a sample network with two-movement repertoire. Colors, different target cues; circles and triangles, target endpoints for each movement; pink background, range of known movements.

during execution (Figure 3.2h, $P = 9.8 \cdot 10^{-4}$, Wilcoxon signed rank test), in contrast to the twisted and tangled movements produced by single-movement networks —with these differences in motor output being appropriately reflected in neural space (Figure S3.4c,f).

Moreover, with their underlying structure, two-movement networks were able to generate intermediate movements that they had not previously learned when probed with the appropriate input signal (Figure 3.2i, Figure S3.5b,c). Combined, these results suggest that learning multiple movements creates structure in the neural space that effectively makes more intermediate activity patterns available. The availability of these additional activity patterns may facilitate adaptation.

Networks with larger repertoires can adapt to a perturbation more easily

To directly assess our first prediction that the additional structure in neural space of multi-movement networks may facilitate adaptation, we applied the same small VR perturbation of 10° to networks with different movement repertoires (Figure 3.3a,b). In general, networks that had learned larger repertoires were able to adapt more quickly (Figure 3.3b,c, Figure S3.5d,e), and without catastrophic forgetting (Figure S3.6). Comparison of adaptation performance across networks with different motor repertoires revealed two trends. First, all three types of multi-movement networks adapted much more rapidly than single-movement networks, as predicted. Within 100 adaptation training trials, multi-movement networks converged to a performance comparable to baseline. Single-movement networks, in contrast, could not fully adapt and had average errors 20% larger than that of the multi-movement networks (Figure 3.3b). Second, within multi-movement networks, those with larger repertoires also adapted more quickly than those with smaller repertoires (Figure 3.3c). Thus, having larger motor repertoires, which was associated with greater structure in activity space following the initial *de novo* learning process (Figure 3.2), facilitated adaptation.

The initial *de novo* learning experience shapes the structure of neural activity and influences subsequent adaptation

The previous simulations support our first hypothesis that having a larger behavioral repertoire facilitates adaptation (Figure 3.3), but is more always necessarily better? Task

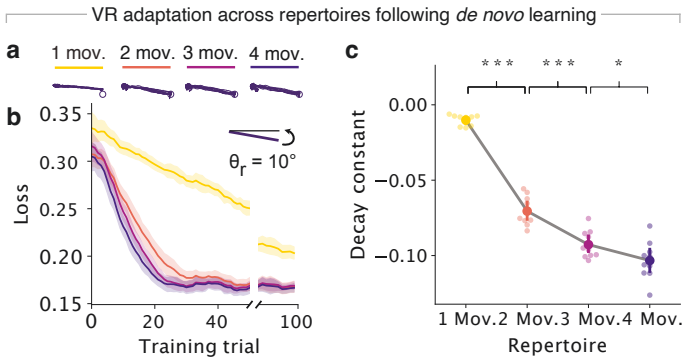


Figure 3.3: Networks with larger repertoires can adapt to perturbations more easily. a. Motor output of networks trained on different repertories (legend) following adaptation to a counterclockwise VR perturbation. **b**. Loss during adaptation, calculated as the mean-squared error between the network output and target positions. Line and shaded surfaces, smoothed mean and 95% confidence interval across networks of different seeds. **c**. Decay constants for exponential curves fitted to the loss curves in Panel b. Circles and error bars, means and 95% confidence intervals with bootstrapping. *** denotes 0.001, ** 0.01, * 0.05 for two-sided paired t-tests. Data includes $n = 10$ networks from 10 random seeds for each repertoire.

performance is shaped by how behavioral components are acquired[143,144], including by whether the examples presented during the initial learning process are meaningfully organized with respect to each other[145]. This led us to hypothesise that the *de novo* learning experience in and of itself would shape adaptation.

To test this hypothesis directly, we trained a second series of networks to perform the exact same reaching task following a different learning experience. We modeled this different learning experience by altering the encoding of the input cues, predicting that this would alter the structure of the network activity. Thus far, we have used continuous angular inputs that specified the direction to the targets (Figure 3.4a). This mimics the process of learning to reach to an endpoint location based on a cue that indicates the take off angle of the movement. An alternative way to perform the same task would be to learn to produce each reach in response to an arbitrary cue, in which case there is no meaningful relationship between target-specific cues. We implemented this alternative learning experience by training networks to produce the same motor output using categorical, one-hot encoded binary vectors that had no angular information (Figure 3.4d, Methods). Even after matching performance following *de novo* learning (Figure 3.4h,i top), the latent dynamics of these networks with categorical inputs no longer had the same organization

during preparation as the motor output (Figure 3.4e), as was the case for the networks with continuous angular inputs Figure 3.4b). This confirms that we could indeed enforce a different structure in neural space by changing the initial learning experience.

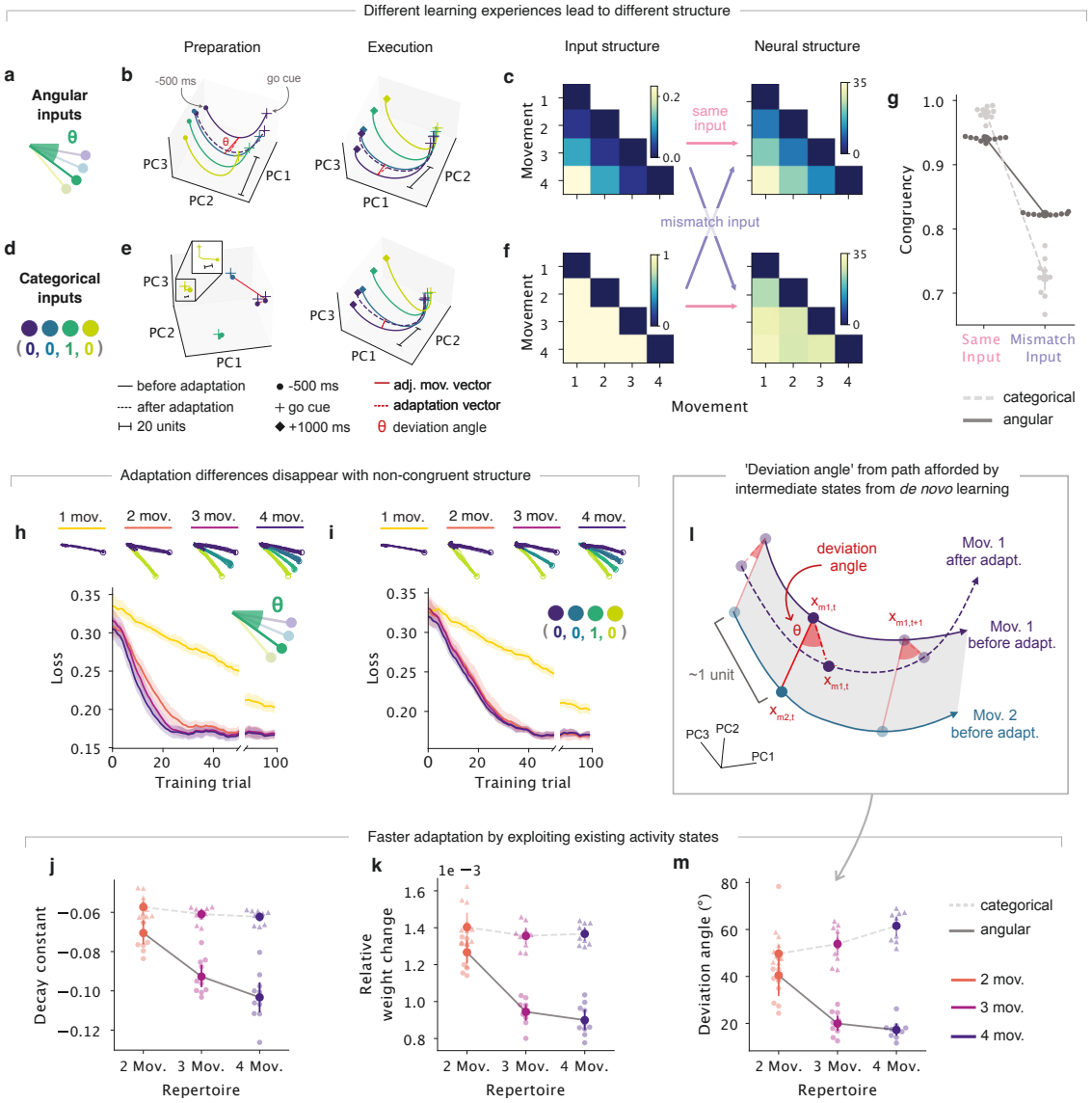


Figure 3.4: The structure in neural space of multi-movement networks resulting from *de novo* learning is responsible for their patterns in adaptation. Networks were given either angular (**a**) or categorical (**d**) inputs to simulate different learning experiences. After training on different repertoires, they had to adapt to a 10° VR perturbation. **b**. Latent activity for an example network with angular inputs trained on four movements during preparation (500 ms before go cue) and execution (1000 ms after go cue). Each trace corresponds to the trial-averaged activity for each movement projected on neural manifold computed before adaptation. Solid lines, activity before adaptation; dotted lines, activity after adaptation. **c**. Left: Input structure, measured as the cosine dissimilaritiy between input vectors for pairs of movements for the network in Panel b. Right: Neural structure, measured as the normalized median Euclidean distances between latent trajectories

during preparation and execution for different movements for the same network. **e-f**. Same as Panels b-c but for a categorical input network. Inset in Panel e: zoomed view. **g**. Congruency between the input and neural structure, quantified as the Pearson's correlation between their dissimilarity matrices. Congruency for mismatched input-activity pairings shown as control. Circle and error bars, mean of congruency values for each seed ($n = 10$) and 95% confidence intervals (CIs) with bootstrapping. **h**. Motor output following skill-learning for angular input networks. Bottom: Loss during adaptation training. Traces and shaded surfaces, smoothed mean and 95% CIs across networks of different seeds ($n = 10$). **i**. Same as Panel h but for categorical input networks. **j**. Decay constants for exponential curves fitted to the loss curves in Panels h and i. Circles and error bars, means and 95% CIs with bootstrapping. **k**. Relative weight changes during adaptation. Circles and error bars, means of the median changes across all weights for each seed and 95% CIs with bootstrapping. **l**. Schematic for a 'deviation angle' between the 'adjacent movement vector' (red solid line in Panel b, e) and the 'adaptation vector' (red dotted line in Panel b, e). **m**. Circles and error bars, means of the median deviation angles across all timesteps for each seed and 95% CIs with bootstrapping.

To further characterize the structure in activity space of each network and compare it across networks with different learning experiences (angular vs. categorical inputs), we measured the Euclidean distances in the neural manifold between the latent trajectories at corresponding time points for different movements (Methods, Figure 3.4c,f right). To allow for comparison between different networks that have different neural spaces, we normalized these distances between movements by the distances between adjacent time points along the same latent trajectory, which should be comparable across networks. We first focused on the networks with angular inputs. During both the preparation and execution epochs, the latent trajectories of these networks were organized similarly to the movements themselves, with movements reaching to adjacent targets also adjacent in neural space (Figure 3.4a,b, Figure S3.7). In contrast, for the networks with categorical inputs, the latent dynamics during preparation for different movements no longer had the same organization as the motor output (Figure 3.4d,e). Rather, the organization of the latent dynamics in both classes of networks reflected the structure of the inputs—quantified by calculating the cosine dissimilarity between pairs of input vectors (Methods)—, suggesting that the structure of the inputs imposes a congruent structure in neural space (Figure 3.4g, Methods). This was the case even if the motor output was equally accurate across network classes (Figure 3.4h,i top) after the initial *de novo* learning phase.

Having two classes of networks that perform the initial center out task equally well fol-

lowing different learning experiences allowed us to directly test the influence of the *de novo* learning experience in subsequent adaptation. For the angular input networks, the activity in neural space was topologically structured (Figure 3.4b), with the latent trajectories organized by the angular direction of the movements. Since a VR perturbation requires angular changes on the motor output, we predicted that the greater congruence between the structure in neural space of angular input networks and the angular changes required by the perturbation would facilitate adaptation. Indeed, while the performance for networks with angular and categorical inputs was comparable following adaptation (Figure 3.4h,i bottom), adaptation was in general faster for networks with more congruence (angular input networks) than networks with less congruence (categorical input networks) (Figure 3.4j), confirming that the experience during initial *de novo* learning influences adaptation. Moreover, for the angular input networks, adaptation was also faster for those with larger movement repertoires (Figure 3.4h), but this trend was absent for the categorical input networks (Figure 3.4i). Single-movement networks lacking structure in neural space were unaffected by the difference in input encoding (Figure 3.4h,i). These results support our second hypothesis that it is not only the motor repertoire that shapes learning: how movements are initially acquired determines the organization of neural activity which also shapes further learning. In particular, the congruence between the structure of the latent trajectories in neural space and the structure in the perturbation is important to facilitate adaptation. Interestingly, when we applied this measure of congruence to experimental data of monkeys performing a center-out reaching task, we could predict the rate of adaptation to a VR perturbation based on the degree of congruence (Figure S3.8f). This lends experimental support for a potential role of congruence between the structure of latent trajectories and that of the perturbation in adaptation.

Adaptation is facilitated by exploiting intermediate states within structured activity

Having established that the motor repertoire and the learning experience jointly shape adaptation, we sought to understand what property of the angular input networks allowed models with greater repertoires to adapt faster. We had previously shown that the structure of these multi-movement networks organizes the neural dynamics and allows them to produce untrained, intermediate movements (Figure 3.2g,i; Figure S3.5b,c). Thus, a larger movement repertoire in angular input networks may lead to more defined neural structure that allow the networks to produce more intermediate activity states. These intermediate states may facilitate adaptation by reducing the need to learn new activity patterns. When modeling learning using RNN models, connectivity must be altered in order to change the activity patterns that can be produced[15,132]. This implies that if large repertoire networks do exploit intermediate states provided by additional movements, they would require smaller adaptive weight changes than smaller repertoire networks. This was indeed the case, but, as anticipated from their differences in the time course of learning (Figure 3.4h), it only occurred in the more congruently structured networks with angular inputs, not in the networks with categorical inputs (Figure 3.4k).

How do intermediate states contribute to the faster adaptation of networks with angular inputs and larger motor repertoires? To explore the role of these intermediate states, we examined how the network activity evolved during adaptation. If we assume that intermediate states exist between the latent trajectories for each movement, we would expect the latent trajectories to move along these states during adaptation, in the direction of other existing latent trajectories that are in the direction of the desired motor output. To measure this, we defined a 'deviation angle' that quantifies how changes in the trajectories during adaptation deviate from the path afforded by the existing potential intermediate states created during *de novo* learning (Figure 3.4l—for validation of this metric on motor cortical recordings from monkeys performing the same VR task[14], see Figure S3.8). If the network is using existing intermediate states, we would expect these angles to be

small. Indeed, networks with angular inputs that had more congruent structure than the networks with categorical inputs also had smaller deviation angles (Figure 3.4m), with the differences paralleling those seen in the rate of adaptation (loss curve decay constants) and relative weight changes (Figure 3.4j,k). Thus, how latent activity is structured in neural space has a large impact on motor adaptation. This structure is jointly shaped by the motor repertoire and how the latent dynamics underlying these movements are organized, which is determined by the learning experience, here modeled as the type of cues. Adaptation may only be facilitated if this structure is congruent to the changes on the motor output that are required by the perturbation (Figure 3.4h,i). Without any structure to guide adaptation, single-movement networks with both types of inputs adapted the most slowly to our VR perturbation. Without congruent structure, multi-movement networks with categorical inputs adapted more slowly than those with angular inputs. Finally, with more movements to provide more intermediate states across the structure, networks with the largest repertoires adapted the fastest.

Structure in neural space resulting from the learning experience can facilitate or impede adaptation

Our previous results suggest that the structure across the inputs, the neural space, and the changes in the output required by the perturbation all need to be congruent for adaptation to be facilitated. To test this directly, we examined how networks adapted to a different type of perturbation that was congruent with the organization of categorical inputs but not with that of angular inputs, as was the case for the VR perturbation. In this 'reassociation' perturbation (Figure 3.5a), the cues and targets were rearranged such that the network needed to re-associate learned reaches to different known target cues[15]. Since this perturbation requires adaptation to categorical rather than angular changes, we predicted that multi-movement networks with categorical inputs would adapt more easily than would those with angular inputs.

Indeed, networks with categorical inputs adapted to the reassociation perturbation (Fig-

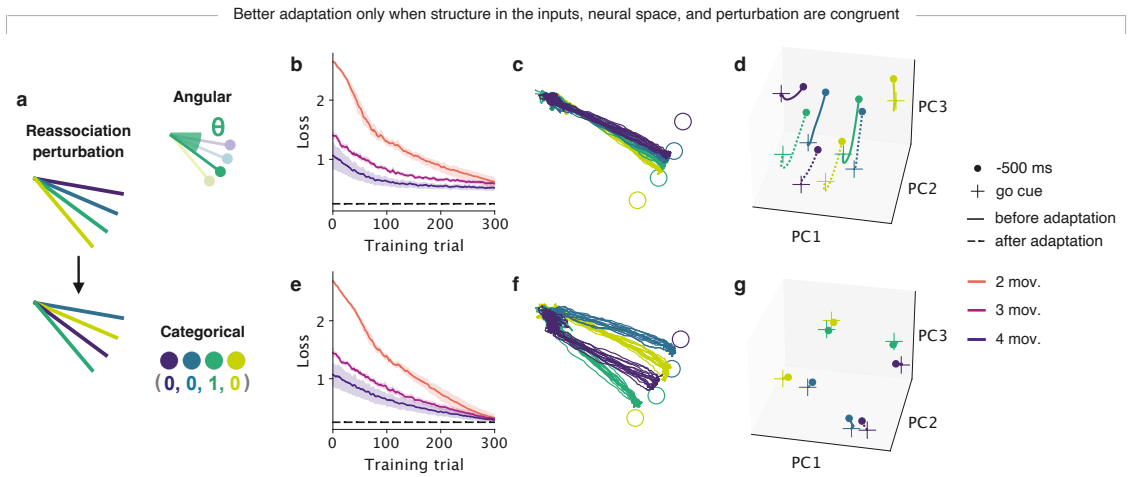


Figure 3.5: The structure in neural space arising from *de novo* learning can facilitate or hinder adaptation. **a**. Networks with angular or categorical inputs adapted to a reassociation rather than a visuomotor rotation perturbation. **b**. Loss during adaptation training for networks with angular inputs with varying motor repertoires. Traces and shaded surfaces, smoothed mean and 95% confidence intervals across networks of different seeds ($n = 10$). The dashed line indicates a loss of 0.25 and is included for easier comparison between the loss in networks with angular or categorical inputs. **c**. Motor output following adaptation training. **d**. Latent trajectories during preparation (see Figure 3.4a) for one example seed. **e-g**. Same as Panels b–d but for networks with categorical inputs. Note that networks with angular encoded inputs could not adapt to the perturbation whereas networks with categorical inputs did since their structure was congruent with the changes required by the perturbation.

ure 3.5 b,c), whereas networks with angular inputs were unable to adapt (Figure 3.5 e,f, Figure S3.9) despite comparable performance following *de novo* learning (Figure 3.4h,i top). For both classes of networks, the latent trajectories maintained their general structure (Figure 3.5d,g), suggesting that the networks adapted by reusing activity patterns in the existing intermediate states rather than by exploring new ones. However, for networks with angular inputs, neighboring latent trajectories interfered with one another during adaptation such that the adapted motor output became overlapped (Figure 3.5c,d). Thus, the existing structure in neural space now harmed adaptation instead of facilitating it. Together, these results further support a fundamental relationship between the structure of neural activity following *de novo* learning and the ability to adapt to a subsequent perturbation: the underlying neural structure can facilitate adaptation if it is congruent to the changes required by the perturbation, but it can also hinder adaptation if the structure of the perturbation is incongruent. Having well-defined structure in activity space also hindered adaptation to other types of perturbations that required greater changes in the

motor output (Figure S3.10). Thus, structure in the neural space can shape adaptation by facilitating adaptation under small changes or interfering adaptation under larger changes. In conclusion, different *de novo* learning experiences can lead to differences in the organization of neural activity, which can in turn lead to dramatic differences in adaptation even across networks with the same motor repertoire.

Generalization to more complex motor repertoires

Our results so far have focused on networks producing reaching movements. To examine the generalizability of our findings to a more complicated task, we trained networks to produce up to four different elliptical movements by generating cosinusoidal and sinusoidal outputs with various amplitudes along the x and y directions, respectively (Figure 3.6a). We chose this task because the oscillatory output requires more complicated timing and output dynamics than a reaching task, and is reminiscent of past studies in motor control[152–154].

We manipulated this new task to test the generalizability of our previous results. First, we previously showed that having a larger motor repertoire can facilitate adaptation by creating intermediate trajectories in activity space (Figure 3.6a). To test this, we trained the networks on two sets of repertoires with different numbers of movements (Figure 3.6b). In one set of repertoires, networks learned one to four movements consisting of cosine waves of different amplitudes, but of sine waves of the same amplitude, creating ellipses of different elongations in the x direction. The other set of repertoires was similar, but with sine waves of different amplitudes and cosine waves of the same amplitude. Second, we previously showed that adaptation is not only influenced by the motor repertoire but also how it was initially learned: to facilitate adaptation, the structure between the inputs, neural space, and the changes imposed by the perturbation all have to be congruent. To test the effects of different learning experiences, we encoded the target signal of the sine wave as a continuous input representing its amplitude, and the target signal of the cosine wave as a categorical one-hot encoded input that represented the same amplitude values

(Figure Figure 3.6a). To test the effects of different perturbations, we asked all networks to adapt to: 1) a 'sine wave perturbation' where they had to produce a movement with a larger sine wave amplitude, 2) a 'cosine wave perturbation' where they had to produce a movement with a larger cosine wave amplitude, or 3) a 'reassociation perturbation' where they had to re-associate target cues to different movements (Figure 3.6b).

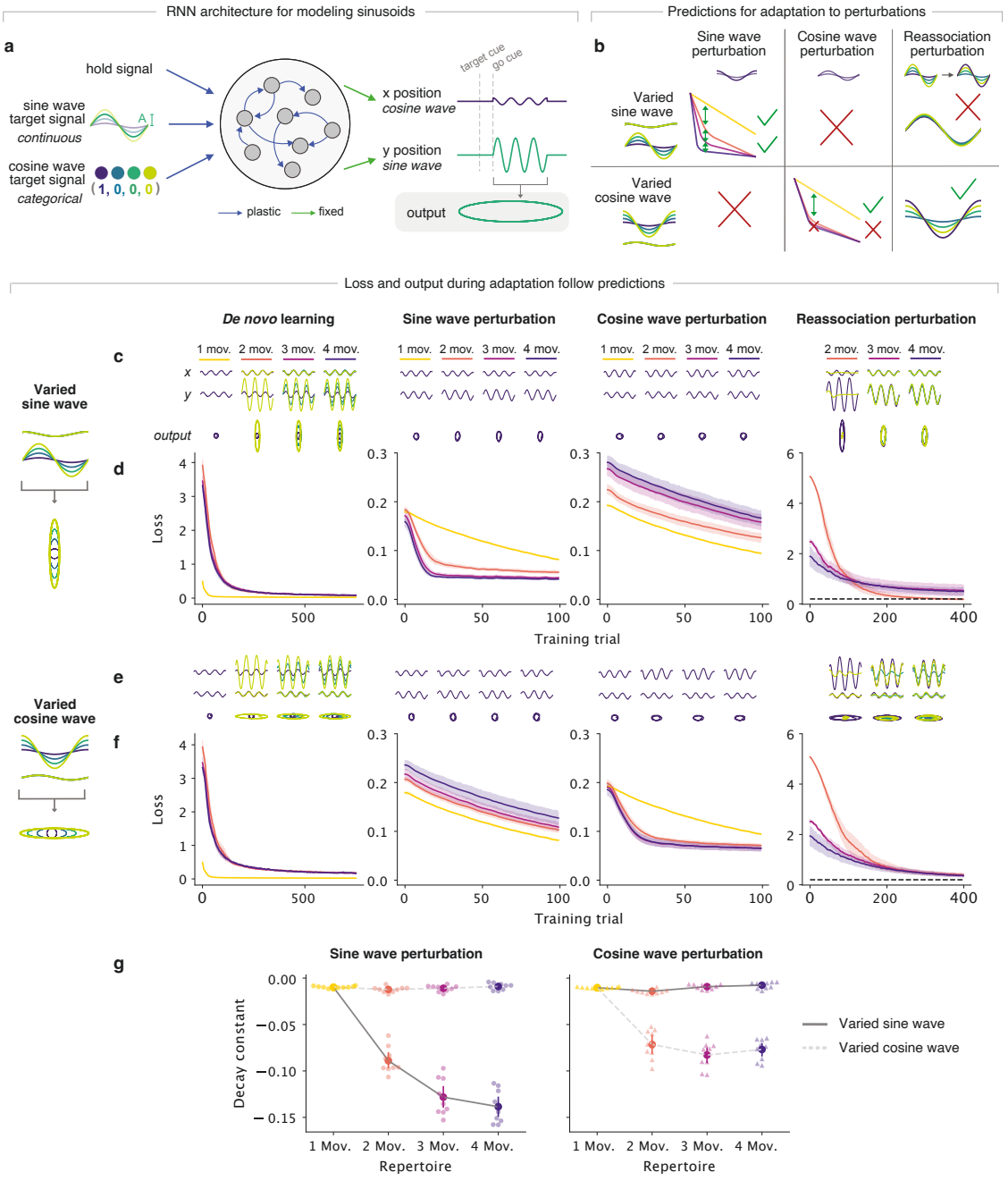


Figure 3.6: The observed joint influence of motor repertoire and *de novo* learning experience

extends to a more complex sinusoidal task. a. Networks were trained to produce ellipses with up to four different geometries by generating cosine and sine waves representing x and y positions, respectively. Networks were given a hold signal that indicated movement initiation, a continuous sine wave target signal that indicated the amplitude of the sine wave, and a categorical 'one-hot-encoded' cosine wave target signal. **b**. Networks were trained on repertoires with different numbers of movements (from one to four) to model *de novo* learning. Within each repertoire, the movements could consist of either different sine waves and a constant cosine wave ('varied sine wave'), or different cosine waves and a constant sine wave ('varied cosine waves'). Following *de novo* learning, networks were separately trained to counteract 1) a sine wave perturbation, 2) a cosine wave perturbation, and 3) a reassociation perturbation where they had to produce different movements given learned target cues. The schematic indicates our predictions of what perturbations would be easier to counteract based on the center-out results. **c**. Cosine waves (x position), sine waves (y position), and resultant elliptical output produced by an example network trained to produce sine waves with various amplitudes after *de novo* learning and each of the perturbations. **d**. Loss during *de novo* learning and adaptation to each of the perturbations. Line and shaded surfaces, smoothed mean and 95% confidence interval across networks of different seeds ($n = 10$ random seeds). Dashed line, loss of 0.2 included for easier comparison between loss in different networks. **e-f**. Same as Panels c–d but for networks trained on repertoires with varied cosine waves. For the reassociation perturbation, 45% of networks trained on either 3- or 4-movement repertoires with varied sine waves were able to adapt (MSE < 0.4), compared to 70% for those with varied cosine waves. **g**. Decay constants for exponential curves fitted to the loss curves in Panels d and f for the sine and cosine wave perturbations.

We could make several predictions based on our previous results (summarised in Figure 3.6b). Regarding the networks that learned varied sine waves, we first predicted that networks with larger repertoires would adapt faster to the sine wave perturbation since the continuous amplitude input encoding of the sine wave is congruent to the continuous amplitude change required by the perturbation. Second, these networks would not adapt faster to the cosine wave perturbation since they have not learned multiple cosine waves and thus do not have the intermediate states necessary to facilitate adaptation. Third, these networks would not adapt well to the reassociation perturbation since their continuous inputs are incongruent to the categorical changes required by the perturbation. Regarding the networks that learned varied cosine waves, we first predicted that they would not adapt faster to the sine wave perturbation since they have not learned multiple sine waves. Second, networks with multi-movement repertoires would adapt faster to the cosine wave perturbation than those with single-movement repertoires since they have some structure. However, larger multi-movement repertoire networks would not adapt faster since the categorical input encoding of the cosine wave is incongruent to the continuous amplitude change required by the cosine wave perturbation. Third, these networks would adapt

well to the reassociation perturbation since their categorical inputs are congruent to the categorical changes required.

The results support all of our predictions (Figure 3.6c-g), showing that our findings on the center-out reaching task generalize to more complicated movement settings. Specifically, different components of a given movement—here, the x and y oscillatory components—can be learned using different input cues, and the way in which these movements are learned can lead to contrasting adaptations to different perturbations. Therefore, our observation that the acquired motor repertoire and the *de novo* learning process jointly shape the structure of the neural activity and influence subsequent adaptation holds for more complex tasks than producing center out reaches.

3.4 Discussion

Adapting to an external perturbation requires the generation of new activity patterns whose availability is likely shaped through long-term motor skill learning. Motor repertoires from different learning experiences thus provide different initial conditions for adaptive activity to evolve. Here, we examined how the learned motor repertoire and how it is acquired jointly shape the underlying structure in neural space and consequently impact motor adaptation using RNNs. We hypothesized that having a larger repertoire of movements could facilitate adaptation since more activity patterns would be readily available, but only if this repertoire was learned in a way that was 'meaningful' for counteracting the perturbation. Indeed, networks with larger repertoires could adapt to perturbations more quickly, but only if the structure of the inputs, the neural space, and the changes imposed by the perturbation were all congruent. These results suggest that adaptation is affected by both the set of existing activity patterns and the organization of these patterns in neural space, which is reflective of the initial *de novo* learning experience.

Previous work showed that, during a carefully designed BCI adaptation experiment, short-term adaptation could be achieved by re-associating existing neural activity patterns, such

that the overall set of neural activity patterns remains the same following adaptation[17]. Other adaptation studies observed that neural activity was shifted following adaptation[14,94]. Notably, however, this shift only occurred under force field perturbation, and not under VR perturbations[14,94], suggesting that our modeling results may fall under the former reassociation regime. Indeed, all our multi-movement networks readily had the ability to produce the activity patterns necessary for adaptation following *de novo* learning: they were able to adapt to perturbations during adaptation training even with frozen recurrent weights (Figure S3.9e,f), showing that recurrent weight changes that produce new activity patterns were not necessary. Within this regime, networks with more structure in activity space still adapted more easily, suggesting that the underlying structure determines how easily activity patterns can be deployed during adaptation. This adds an additional layer of complexity to studies investigating adaptation within the existing neural manifold[16,46,147].

Furthermore, we could define a way by which the *de novo* learning experience shapes subsequent adaptation by establishing a relationship between input structure, neural space structure, and perturbation structure. This relationship lends insight to previous findings. Work on 'structural learning' showed that participants adapted more easily to novel perturbations that have the same structure as perturbations from prior experience[149,155]. They argued based on behavioural data that knowing this structure helps learning by facilitating exploration of a previously acquired low-dimensional 'task-related' space. Here, we showed that the initial *de novo* learning phase creates structure in neural space by organising the acquired latent trajectories in a manner reflective of the learning experience, imposing constraints on the trajectory and even the feasibility of future learning (Figure 3.4, Figure 3.5, Figure 3.6). Contextual information is another way by which the learning experience can be shaped[156] and generalization in motor adaptation is highly dependent on the context[157]. In our model, different types of contexts were created via different types of inputs, and these manipulations led to predictable differences in future learning experiences based on the congruency between the inputs and the changes required by a specific perturbation (Figure 3.4, Figure 3.5, Figure 3.6). Thus, our work suggests a potential common neural basis for various behavioural observations spanning

structural learning, generalization of motor adaptation, and context inference.

We can make several experimental predictions based on our results. First, while it is experimentally difficult to examine long-term learning on the timescales we are interested in, we may be able to test our predictions in long training sessions that employ repeated movements. Verstynen and Sabes[158] showed that participants had more varied reach angles towards a given target when a series of recently performed reaches had more variance. These experience-dependent changes in variance mirror those we saw in our simulations (Figure S3.2), where networks that learned more movements had greater variance and less precision in the motor output (Figure 3.2e). These similarities suggest that we could perhaps use longer training sessions to study some effects of long-term learning on short-term adaptation. Thus, our modeling results (Figure 3.2e,f) predict that there would be a trade-off between motor output precision and robustness to neural noise[151] when more movements are learned through repetition using a paradigm similar to that in Ref. 158.

Second, while there were some differences between multi-movement networks of different sizes, the greatest differences were between single and multi-movement networks (Figure 3.2). Multi-movement networks with the same distribution of learned movements had similar ease in adaptation (Figure 3.3b,c; Figure S3.5d,e), since networks were able to produce activity for intermediate movements within the distribution. Thus, we predict that participants would adapt more quickly if they learn a larger distribution of movements, and learning fine-tuned movements within the distribution would provide smaller benefits (provided that these were acquired through the same learning process).

Third, the *de novo* learning experience, which we modeled by manipulating the structure of the inputs, shaped how the networks adapted under different perturbations by changing the underlying structure in neural space (Figure 3.4, Figure 3.5, Figure 3.6). By comparing networks and monkeys trained on the center-out reaching task, we saw that the deviation angles in the monkey neural activity were more similar to those for networks with angular inputs (Figure S3.8e). Classically, the center-out reaching task has been performed by

showing participants the position of the reach target (e.g. Refs. 14,24,159,160). This is most similar to presenting angular inputs since the cues specify the angular location of the target. Assuming angular inputs in data from monkeys engaged in a VR adaptation task, the speed of within-session adaptation and the degree of congruency between the structure across the inputs, neural activity space, and the perturbation were robustly correlated, giving further experimental support to our modeling result that congruency facilitates learning (Figure S3.8f). To further experimentally examine the potential impact of the structure of the inputs on learning, categorical inputs could be used instead by cuing targets with different shapes or colors rather than the explicit target location. Our results suggest that learning the center-out task based on these two different types of cues would lead to analogously different patterns of adaptation under different kinds of perturbations such as a VR or the reassociation perturbation we have studied.

There are several limitations and ways to expand on our work. To assess how different motor repertoires affect adaptation, we trained the networks in two stages: *de novo* learning of multiple movements followed by adaptation on a single movement. This set-up makes the networks vulnerable to catastrophic forgetting, since the networks may forget the other movements that are not being trained on during adaptation. Training with FORCE learning[161] was especially susceptible to catastrophic forgetting: following adaptation, networks were largely unable to produce output for the other movements (Figure S3.6). In contrast, training with stochastic gradient descent, a technique known to replicate various experimental results in adaptation[15,162], was largely able to overcome catastrophic forgetting (Figure S3.6) without overwriting the initial network (Figure S3.11). Consequently, we decided to use stochastic gradient descent throughout our simulations since it was more behaviorally relevant.

By directly relating the network model activity to motor output, we aimed to model population activity in the motor cortex, which in primates is the main cortical area that projects to the spinal cord to cause movement[163]. While the model has similarities to experimental recordings in monkey motor cortex (Figure S3.8), it has not been explicitly fitted to neural data (e.g. Refs. 164–166). Thus, our model is largely region-agnostic, and it

is still unclear where these neural changes due to motor learning and adaptation may occur. Motor learning seems to be associated with activity changes across cortical regions such as the premotor cortex[14,15,37,167], primary motor cortex[15,168,169], parietal cortex[170,171], as well as other structures such as the cerebellum[4–6] and perhaps even basal ganglia[8,172–174]. How these different regions interact to affect skill learning and adaptation is an ongoing area of study, and future work could use modular and area-specific networks that are constrained by neural data to tease apart their contributions[164–166].

Another potential direction of future research is to extend our current results on the structure of the latent trajectories in neural space to an investigation of the 'dynamical motifs' underpinning these trajectories[133,175]. Since different dynamical motifs arise from learning different tasks[175], different motifs may similarly arise based on different learning processes. Given that the time-varying latent dynamics mediate the planning and execution of movement and exhibit changes during adaptation[14], these dynamical motifs may then predict the result of adaptation to different perturbations.

Future work can also examine more complex motor skills and learning experiences. Here, we have primarily focused on simple movements adapted from the center-out reaching task since it allowed us to compare the motor output and latent dynamics to experimental data—although we have reproduced our key results in a more complex task that simulated tracking ellipses with various geometries (Figure 3.6). We can expand on this work to examine more complex and realistic repertoires that include different behaviors like grasping and manipulation, along with models that consider arm kinematics[176,177] and dynamics[178]. Different actions have been shown to occupy different parts of neural space[44,45], so different combinations of behaviors are likely to require different underlying manifolds, which would affect subsequent adaptation to perturbations on any given behavior. In addition, we can examine how learning is affected by different training processes. Here, we have specified that our networks undergo *de novo* learning since they are trained to perform reaches for the first time, but our results may not be specific to this situation. Rather, our results may be more indicative of how prior training affects further learning in general. To probe this further, future work can examine how networks pretrained on many tasks learn

to perform a new skill, compared to networks that are trained from scratch on the same new skill or networks that are trained on different similar skills in sequence.

In conclusion, we have shown that *de novo* motor learning shapes subsequent adaptation by creating structure in neural space. Most dramatically, two sets of networks that have different neural structure can exhibit opposite trends when adapting to the same perturbation, even if they know the same set of movements equally well. The neural structure is defined by the specific *de novo* learning process, which can be manipulated by providing cues that either have or lack a meaningful relationship among them. Furthermore, knowing a larger repertoire of movements often facilitates adaptation, but only when the changes required by the perturbation are congruent with the structure of the underlying latent dynamics. While we have examined the formation of structure in neural space in the context of motor learning, similar structural constraints may arise in other systems, shaping not only motor but also cognitive processes.

3.5 Methods

Reaching task

We trained recurrent neural networks to perform a standard center-out reach task, which is commonly used in experimental settings to examine motor control. We modified existing experimental data for reaches using the task design of Ref. 14. In the experiment, a monkey controlled a cursor on a computer screen using a 2-D manipulandum. The cursor starts in the middle of a circle with a radius of 8 cm and it must reach to possible targets spaced around the circle. The monkey is shown which target to move to at a target cue, but they must delay movement until a later go cue. To understand how existing skill sets affect motor adaptation, we created four motor repertoires of different sizes, ranging from one to four reach movements to one to four targets, respectively. We adapted experimental reaches of one monkey, Monkey M, from Ref. 14 to create the target reaches for each movement by rotating the experimental reaches to different targets that we defined. In

our case, the targets were equally spaced around an arc of the circle. Unless specified otherwise, the arc spanned from -10° to -50°. Networks were trained on one of these four repertoires. By using different numbers of movements in each skill set, we were able to examine how a network that knows more movements may adapt differently than a network that knows less. Each trial lasted 4.0 s: the target and go cues were randomly selected for each trial, with the target cues between 1.0-2.5 s and the go cues between 2.5-3.0 s.

To examine network performance without trial-to-trial movement variability, which was inherent in experimental reaches, we also created repertoires with synthetic reaches. These synthetic reaches had fixed target and go cues based on the mean target and go cues in the experimental data, and the target position profiles of each reach was the same across trials. Each synthetic reach lasted 1.0 s and the position profile was defined by a sigmoid for each time t for the length of the reach $l = 8$ cm:

$$\frac{l}{1 + e^{-12t+6}}$$

To assess motor adaptation, we examined how networks adapted to a visuomotor rotation (VR), a common perturbation used in experimental settings. To simulate VR, we rotated the output position of the network counterclockwise by θ_r during adaptation trials. We used rotations with $\theta_r = 10°$, 30°, or 60°. We also examined how networks adapted to visuomotor re-associations where the target cues and targets are rearranged, such that the network must reach to a different known target given a known target cue.

Sinusoidal task

To extend our results to a more complex task, we trained recurrent neural networks to produce cosine and sine waves representing x and y positions that trace out ellipses. The task followed a similar trial structure to that of the delayed reaching task: each trial lasted 4.0 s, and the target cues were fixed at 800 ms while the go cues were fixed at 130 ms. We created synthetic movements that lasted 200 ms after the go cue. Each movement consisted

of a triple sine wave and a triple cosine wave, which traced out three cycles of an ellipse. To understand how existing repertoires affect motor adaptation, we created two sets of repertoires of different sizes, ranging from one to four movements tracing out one to four ellipses, respectively. For the first set of repertoires, we varied the amplitude of the sine waves such that different movements in each repertoire had sine wave amplitudes equally spaced within a range of 1 to 7 cm, and a constant cosine wave with an amplitude of 1 cm. The single-movement repertoire had a sine wave amplitude of 1 cm. For the second set of repertoires, we varied the amplitude of the cosine waves instead, while keeping the sine wave amplitude constant.

To assess motor adaptation, we examined how networks adapted separately to three different perturbations. First, we asked networks to adapt to a 'sine wave perturbation' on the one movement shared by all repertoires (the one with sine and cosine wave amplitudes of 1 cm) by producing a sine wave with a larger amplitude of 2 cm . Second, we asked networks to adapt to a 'cosine wave perturbation' on the one shared movement by producing a cosine wave with a larger amplitude of 2 cm. Third, we asked networks to adapt to a reassociation perturbation where the target cues and target movements were rearranged, such that the network must produce a different movement given a known target movement cue.

Neural network model

Network architecture

The model dynamics were given by:

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij} r_j(t) + \sum_{k=1}^{I} B_{ik} s_k(t) + \eta_i(t)$$

where x_i is the hidden state of the ith unit and r_i is the corresponding firing rate following $tanh$ activation of x_i. The network has $N = 300$ units and I inputs. The time constant

$\tau = 0.05s$, the integration time step $dt = 0.01s$, and the noise η is randomly sampled from the Gaussian distribution $\mathcal{N}(0, 0.2)$ for each time step. The initial states $\mathbf{x}_{t=0}$ are sampled from the uniform distribution $\mathcal{U}(-0.1, 0.1)$. The network is fully recurrently connected, with the recurrent weights **J** initially sampled from the Gaussian distribution $\mathcal{N}(0, \frac{g}{\sqrt{N}})$, where $g = 1.2$. The time-dependent stimulus inputs **s** (specified below) are fed into the network, with input weights **B** initially sampled from the uniform distribution $\mathcal{U}(-1.0, 1.0)$.

For the networks trained on the reaching task, two types of sustained inputs **s** were used. For the angular inputs, **s** is 3-D and consists of a 1-D fixation signal and a 2-D target signal $(2\,cos\,\theta^{target}, 2\,sin\,\theta^{target})$ that specifies the reaching direction θ^{target} of the target. For the categorical inputs, **s** is 5-D and consists of a 1-D fixation signal and a 4-D one-hot encoded target signal with the same magnitude (e.g. (0,0,2,0)) that does not provide information about the target's angular direction. For the networks trained on the sinusoidal task, one type of sustained input **s** was used that included both continuous and categorical encoding of target movements. For this task, **s** is 6D and consists of a 1-D fixation signal, a 1-D sine wave target signal ($2A/A_{max}$, where the maximum amplitude A_{max} is 7) that specifies the amplitude A of the sine wave, and a 4-D one-hot encoded cosine wave target signal with the same magnitude (e.g. (0,0,2,0)) that specifies the cosine wave target without providing information about its amplitude. For all types of inputs, the fixation signal starts at 2 and goes to 0 at the go cue and the target signals remains at 0 until the task cue.

The networks were trained to produce 2-D outputs **p** corresponding to x and y positions of motor trajectories, and they are read-out via the linear mapping:

$$p_i(t) = \sum_{k=1}^{N} W_{ik} r_k(t)$$

where the output weights **W** are sampled from the uniform distribution $\mathcal{U}(-1.0, 1.0)$. During VR adaptation trials, **p** is rotated counterclockwise according to a perturbation angle θ_r.

Training the model

For the reaching task, networks were optimized to generate positions of reach trajectories modified from Ref. 14. The training and testing datasets were created by pooling successful trials during baseline epochs across all experimental sessions for Monkey M (2208 trials: 90% training, 10% test). The experimental data was modified for each repertoire (see Reaching task), and equal numbers of trials for each reach direction were included for each repertoire. For the sinusoidal task, networks were optimized to generate synthetic sine and cosine waves representing elliptical movements without trial-to-trial variability for each movement.

Unless otherwise specified, networks had to learn their input weights B and recurrent weights J while the output weights W remained fixed. To model motor skill learning, we initially trained the networks on repertoires with one to four movements, using the Adam optimizer with an initial learning rate $l = 10^{-4}$, first moment estimates decay rate $\beta_1 = 0.9$, second moment estimates decay rate $\beta_2 = 0.999$, and epsilon $\epsilon = 1e - 8$. Then, to model motor adaptation, we trained the pre-trained networks to counteract a perturbation (VR or reassociation for the reaching task; cosine wave, sine wave, or reassociation for the sinuoidal task), using stochastic gradient descent with a fixed learning rate $l = 5^{-3}$, unless otherwise specified. We used a faster learning rate during adaptation to model faster short-term learning compared to long-term skill learning. To assess how existing repertoires may shape adaptation on a given movement, networks were only trained to counteract the VR, sine wave, or cosine wave perturbation on the one target that all repertoires share (i.e. the -10° reach, or the ellipse consisting of cosine and sine waves with an amplitude of 1), such that performance was comparable across networks with different repertoires. Initial training was implemented with 750 training trials and a batch size $B = 64$. For the reaching task, adaptation training was implemented with either 100 training trials for VR perturbations or 300 training trials for reassociation perturbations, and a batch size $B = 64$. For the sinusoidal task, adaptation training was implemented with either 100 training trials for the cosine and sine wave perturbations or 400 training trials for reassociation

perturbations, and a batch size $B = 64$. All training configurations were performed on 10 different networks initialized from different random seeds. To examine adaptation under an alternate learning algorithm, we also trained the networks to counter the VR perturbation during the reaching task using FORCE learning[161] with a learning rate of 100.

The loss L was the mean squared error between the 2-D output and target positions over each time step t, with the total number of time steps T = 400. The first 50 time steps were not included to allow network dynamics to relax:

$$L = \frac{1}{2B(T-50)} \sum_{b=1}^{B} \sum_{t=50}^{T} \sum_{d=1,2} (p_d^{target}(b,t) - p_d^{output}(b,t))^2.$$

To produce dynamics that align more closely to experimentally estimated dynamics[49,50], we added L2 regularization terms for the activity rates and network weights in the overall loss function L_R used for optimization:

$$L_R = L + R_W + R_R$$

where

$$R_R = \frac{\beta}{BTN} \sum_{b=1}^{B} \sum_{t=0}^{T} \sum_{n=1}^{N} r_n(b,t)^2$$

and

$$R_W = \alpha(||J|| + ||B|| + ||W||)$$

where $\beta = 0.5$ and $\alpha = 0.001$. Note that the loss recorded in the main text was the loss L before regularization, and it was smoothed with a backward moving average of five trials. We clipped the gradient norm at 0.2 before applying the optimization step. While we have used the specific parameters described above for training our neural network models, we also trained the models with varied parameters and the trends in our main results remain generally robust to these changes (Figure S3.12).

Data analysis

Analyses on the neural activity were examined for both the preparation and execution epochs of the movement, taken as 500 ms before and 1000 ms after the go cue, respectively.

To assess how latent dynamics change during motor learning and adaptation, we examined the neural activity space of the networks. In the activity space of a population of n neurons, each point denotes the state of the neural population, and each axis corresponds to the firing rate of a specific neuron. To obtain smooth firing rates through time, we applied a Gaussian kernel (std = 50 ms) to the activity rates from the networks. We identified a lower m-dimensional neural manifold in the activity space by applying PCA to the smoothed firing rates of the neurons. PCA finds orthogonal basis vectors (principal components or PCs) that maximally captures the variance in the population activity. PCA finds n PCs for an n-dimensional space, but we sorted the PCs by their corresponding eigenvalues to get the k-leading PCs, or neural modes, that capture the majority of the variance. Here, we used $k = 10$, which captured more than 80% of the variance in our network activity. We projected the original smoothed firing rates onto the neural modes to get the latent dynamics of the networks.

To measure the variability in activity and motor output, we aligned trials by the go cue and measured the variance across trials for the same reach at corresponding time points for each feature (i.e. for each unit, latent dimension, or output dimension). To compare the variance in latent dynamics across different neural spaces, we first normalized the latent dynamics by the median distances between trial-averaged time points within the neural space before calculating the variance. Variance in unit activity and latent dynamics was calculated during both preparation and movement, while variance in output position was calculated during movement. Reach angles were calculated based on the mean angle for the entire reach during movement, and the variance was calculated across trials. To determine how patterns in variance in the motor output can differ from those in the activity, we examined the variance in unit activity in the output-potent and output-null subspaces[35]. Unit activity was directly related to the motor output through the read-out weights W

of the networks. The output-potent dimensions were then the row space of W while the output-null dimensions were the null space of W. We projected the unit activity onto these dimensions to get the activity in the respective subspaces. For each measure, we calculated either the median variance across all features and timepoints (Figure 3.2) or the median total variance summed over all features, across all timepoints (Figure S3.3). Notably, the trends remained the same for the different variance calculations.

To measure the precision of the latent and motor trajectories, we calculated the 'tangling' in both the latent space (Q_x) and neural space (Q_p), adapted from the measure specified in[179]:

$$Q_x(t) = \max_{t'} \frac{||\dot{\mathbf{x}}_t - \dot{\mathbf{x}}_{t'}||^2}{||\mathbf{x}_t - \mathbf{x}_{t'}||^2 + \epsilon}$$

$$Q_p(t) = \max_{t'} \frac{||\dot{\mathbf{p}}_t - \dot{\mathbf{p}}_{t'}||^2}{||\mathbf{p}_t - \mathbf{p}_{t'}||^2 + \epsilon}$$

where $\mathbf{x}_t$ is the latent activity at time t, $\dot{\mathbf{x}}_t$ is its temporal derivative, $\mathbf{p}_t$ is the position at time t, $\dot{\mathbf{p}}_t$ is its temporal derivative, $||\cdot||$ is the Euclidean norm, and ϵ is a small value to prevent division by zero. 'Tangling' measures how dissimilar future states can arise from similar current states, so it can be used to examine the precision of different trajectories.

To assess how the underlying neural manifold changes during motor adaptation, we visualized these changes by projecting both the unit activity before and after adaptation onto the neural modes of the manifold before adaptation (Figure 3.4b,e). We quantified the manifold overlap before and after adaptation training as specified in Ref. 142, adapted from Ref. 180. To find the manifold overlap between the manifolds of two network activities $\mathbf{A}_1$ and $\mathbf{A}_2$, we first find the covariance matrix $\mathbf{C}_1$ of $\mathbf{A}_1$ and project it onto its neural manifold identified through PCA. We then find the covariance matrix $\mathbf{C}_2$ of $\mathbf{A}_2$ and project it onto the neural manifold of $\mathbf{A}_1$. To quantify the variance explained by these projections, we

divide the trace of these projections by the trace of the corresponding covariance matrices:

$$\beta_1 = \frac{Tr(\mathbf{V_1}\,\mathbf{C_1}\,\mathbf{V_1^T})}{Tr(\mathbf{C_1})} \quad \beta_2 = \frac{Tr(\mathbf{V_1}\,\mathbf{C_2}\,\mathbf{V_1^T})}{Tr(\mathbf{C_2})}$$

where $\mathbf{V}_1$ are the first 10 principal components resulting from PCA on $\mathbf{A}_1$. Here, β_1 is the variance in $\mathbf{A}_1$ that can be explained by the neural manifold for $\mathbf{A}_1$ while β_2 is the variance in $\mathbf{A}_2$ that can be explained by the neural manifold for $\mathbf{A}_1$. We then calculate the manifold overlap as the ratio β_2/β_1.

Changes in the neural manifold are driven by changes in synaptic connectivity, so we also measured the relative weight change dW before and after adaptation training:

$$relative\ dW = \left| \frac{W_{after} - W_{before}}{W_{before}} \right|.$$

To examine the direction of these changes in the manifold relative to the initial shape of the manifold, we defined a metric called the 'deviation angle'. First, we define 'adjacent movement vectors' $\mathbf{v}_{adj}$ between corresponding time points of the trial-averaged latent trajectories of the first movement $\mathbf{x}_{m_1}$ and its adjacent movement $\mathbf{x}_{m_2}$ (the second movement) before adaptation:

$$\mathbf{v}_{adj}(t) = \mathbf{x}_{m_2}(t) - \mathbf{x}_{m_1}(t)$$

These vectors quantify the general shift from one movement to the next in neural space and approximate the shape of the manifold between the adjacent movements. Then, we perform a similar computation for the first movement before ($\mathbf{x}_{m_1}$) and after ($\tilde{\mathbf{x}}_{m_1}$) adaptation to get the 'adaptation vector' $\mathbf{v}_{adp}$ that quantifies the general shift during adaptation:

$$\mathbf{v}_{adp}(t) = \tilde{\mathbf{x}}_{m_1}(t) - \mathbf{x}_{m_1}(t).$$

We define the 'deviation angle' as the angle between these two vectors, which measures how changes in adaptation deviate from the path afforded by the existing scaffolding before adaptation.

To examine congruency between the structure of the inputs, neural space, and the changes required by the perturbation, we defined congruence as the representational similarity between any two structures. For the input structure, we constructed a representational dissimilarity matrix (RDM) [181] by calculating the cosine dissimilarity between pairs of target signal input vectors associated with each movement. For the neural space structure, we constructed a RDM by calculating the distances between the latent trajectories associated with each movement. Specifically, we quantified the median distance D between the trial-averaged latent dynamics $\mathbf{x}_1$ and $\mathbf{x}_2$ for each pair of movements over all corresponding time points t:

$$D = med(||\mathbf{x}_1(t) - \mathbf{x}_2(t)||).$$

To compare these distances between movements across different neural spaces, we normalized by the median distances between time points within each movement, pooled across all movements m:

$$\hat{D} = \frac{D}{med(||\mathbf{x}_m(t) - \mathbf{x}_m(t-1)||)}.$$

We can then measure the congruence between the input and neural space structures by calculating the Pearson's correlation coefficient between the upper triangular regions of their respective RDMs. The structure of the changes required by the perturbation is harder to define mathematically since the nature of the perturbation can vary widely. Instead, we described the congruence qualitatively: an angular perturbation like a VR perturbation is congruent to angular inputs and a categorical perturbation like a reassociation perturbation is congruent to categorical inputs.

Experimental comparison

To verify that the networks produced realistic latent dynamics, we trained the networks on the original 8-target center-out reach task defined in Ref. 14 (see Reaching task). The target trajectories for training and testing were based on reach trajectories during successful trials pooled from all sessions for Monkey C (see Training the model). Following training,

we compared the simulated activity to experimental activity recorded from motor cortex during one baseline session for Monkey C[14]. We pre-processed the experimental recordings by removing units with trial-averaged firing rates less than 5 Hz and applying a Gaussian kernel (std = 50 ms) for the binned square-root transformed firings of each unit (bin size= 30 ms). We then substracted the cross-condition mean. To compare the latent dynamics (see Data analysis), we used Canonical Correlation Analysis (CCA), which finds new directions (canonical correlations or CCs) in the neural manifold that maximize the pairwise correlations between two data sets when they are projected on these directions[122]. Canonical correlation values that are associated with these CCs range from 0 to 1, with 1 being complete correlation.

To verify the 'deviation angle' metric, we trained the networks to counteract a 30° VR perturbation on all 8 targets and compared the deviation angles to those found in monkeys performing the same adaptation (3 sessions for Monkey C, 3 sessions for Monkey M, 10 sessions for Monkey M2; Monkey M and M2 are the same monkey, but they are treated separately since the recordings are from different regions in motor cortex). Deviation angles were found for the latent dynamics corresponding to reaches for all 8 targets. We also calculated deviation angles for shuffled targets and time points as a control for the networks.

To examine congruency between the structure of the inputs and neural space, we assumed t¡hat monkeys were given angular inputs during the center-out reach task since they were shown the location of the target centered around a circle. We calculated the congruency between the structure of angular inputs and the structure of their neural spaces. To compare congruency with the extent of adaptation, we calculated error curves for each session based on the angular error of monkey reaches in the first 150 ms and calculated decay constants for exponential curves fitted to these error curves, as we did for the loss curves during training for the networks.

Data availability

We did not acquire new monkey datasets for experimental comparison, and instead relied on existing datasets from Ref. 14. Most of the monkey datasets used have been previously analyzed[14,43,112,182] and made publicly available on Dryad (`https://datadryad.org/stash/dataset/doi:10.5061/dryad.xd2547dkt`). The remaining datasets will be made available on request.

Code availability

All analyses were implemented using custom python code (Python 3.6) and open-source software. All the figures are reproducible by running Jupyter notebooks. Code to reproduce all the results is openly available at `https://github.com/JoannaChang/de_novo_learning_structure`.

3.6 Supplementary Figures

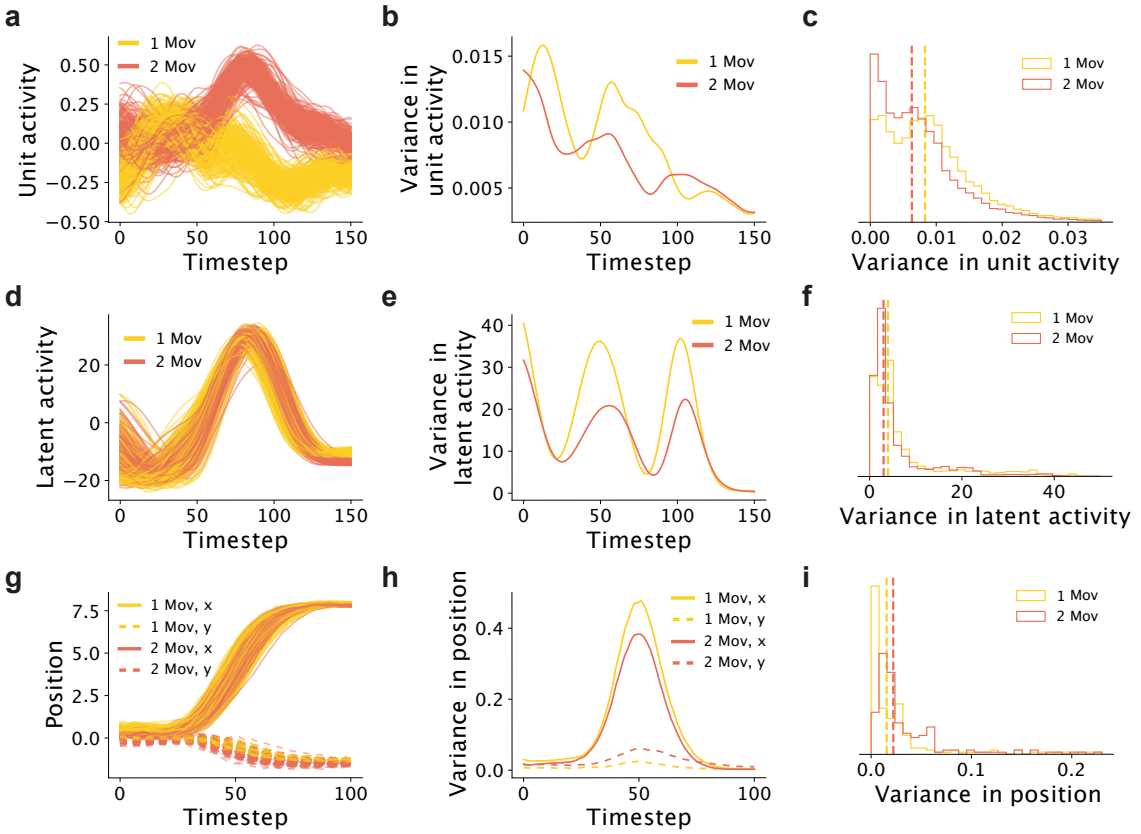


Figure S3.1: Networks that can generate multiple movements produce more constrained neural dynamics. This figure presents additional data for Figure 3.2b–d, and compares different variables across networks with different repertoires producing the same common movement. **a**. Unit activity for one example unit for networks trained on one (yellow) or two (orange) movements. Networks had the same random seed. Traces, different trials. **b**. Variance in unit activity, calculated per timestep across all trials for example unit in Panel a. **c**. Distributions for variance in unit activity, pooled across all units for one example seed. Dashed lines, median. **d-f**. Same as Panels a-c but for the second dimension of the latent dynamics in Panels d-e and for all dimensions in Panel f. **g-i**. Same as Panels a-c but for the motor output. In Panels g-h: Solid line, position along the x axis; dotted line, position along the y axis.

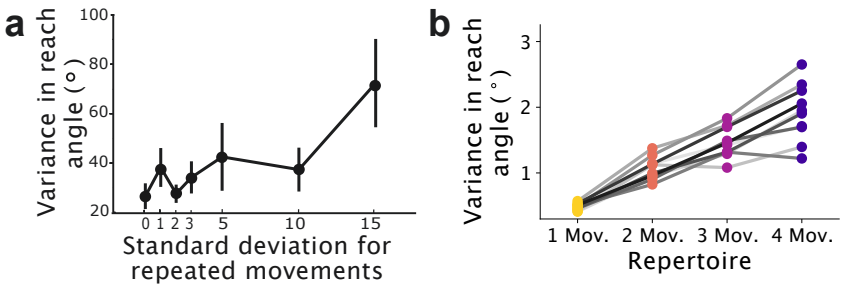


Figure S3.2: Networks reproduce experimental results from Ref. 158. a. This study examined the variance in reach angles to a given target following repeated movements sampled from a normal distribution around the target with different standard deviations. Larger variance of known movements was correlated with larger variance in reach angle. Figure modified from Fig. 2a from Ref. 158. **b**. Networks initially trained on different repertoires were tested on one shared movement. Networks trained on larger repertoires had a larger variance of known movements, and this was also correlated with larger variance in reach angle.

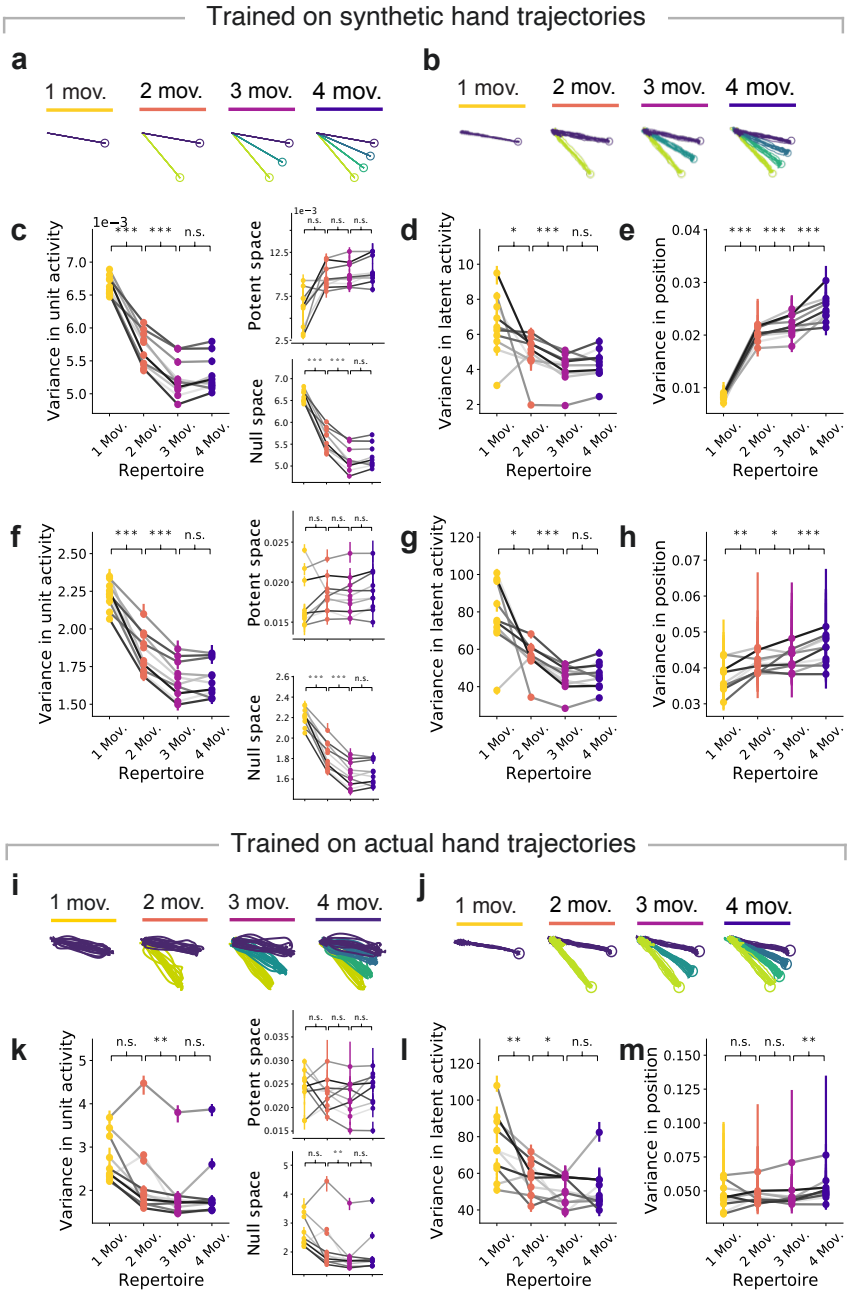


Figure S3.3: Trends in variance remain robust for synthetic movements and different measures of variance. a-h. Networks were trained to produce motor output (**b**) based on simulated (**a**) rather than actual "hand trajectories" based on monkey movements. Simulated trajectories did not have trial-to-trial variability for a given movement. **c-e**. Same as Fig.3.2c–d but for networks trained on simulated hand trajectories. **f–g**. Same as Panels c–e but for total variance summed across all features (across all units for unit activity, latent dimensions for latent activity, and output dimensions for position) rather than the variance for each individual feature. Variance is still calculated for each timestep across trials. Note that the patterns remained the same for synthetic trajectories as when trained on actual hand trajectories, showing that constrained dynamics are not a byproduct of variability in the monkey movements they were trained on. **i–m**. Networks were trained to produce motor output (**i**) based on actual "hand trajectories" based on monkey movements (**j**). These are the same simulations as in the main text. **k–m**. Same as Panels f–g but for

networks trained on experimental trajectories. Note that the trends remain the same. While some trends are not as strong, they are likely due to masking from the variability inherent in experimental trajectories, since the trends remain strong for the networks trained on synthetic trajectories.

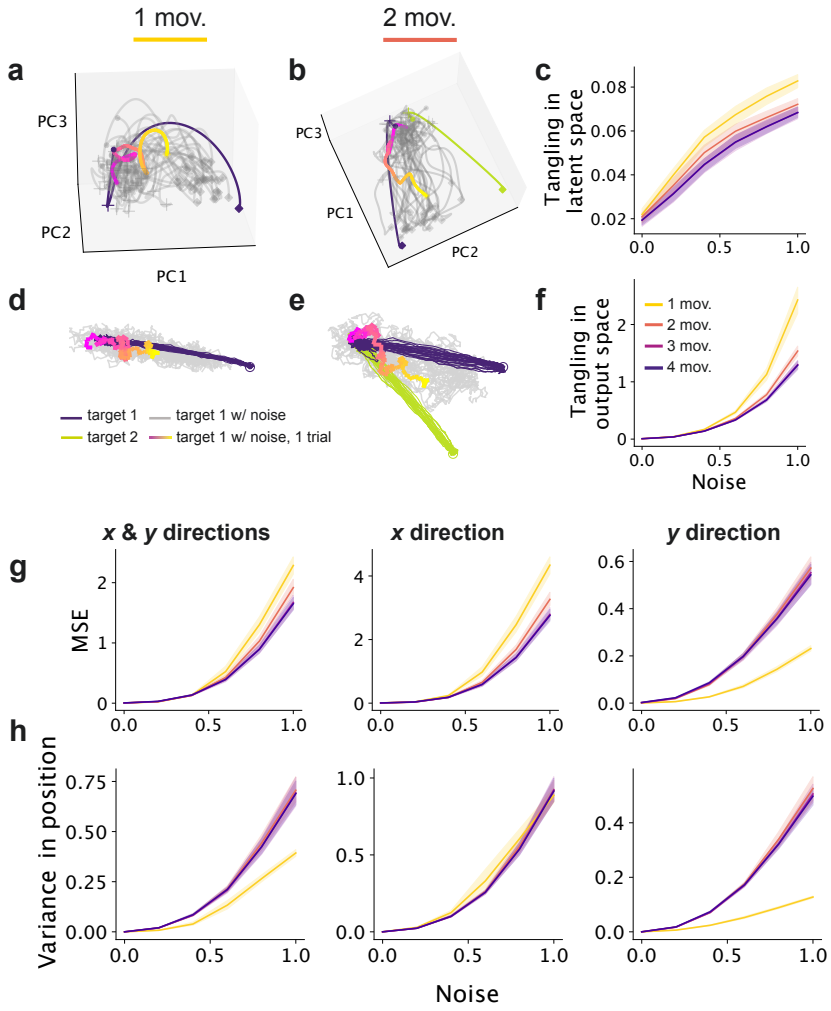


Figure S3.4: Addition of noise shows underlying structure in multi-movement networks. Noise was applied to networks trained on synthetic trajectories without trial-to-trial variability to guarantee that differences would be due to different repertoires rather than different variability in the training data. **a**. Latent trajectories for an example single-movement network. Purple line, trial-averaged trajectory without increased noise added. Pink gradient and grey lines, trajectories for example trials when increased noise ($\eta = 1$) is added. Color gradient, time-course of the trial. **b**. Same as Panel a but for an example two-movement network. Noise is only added to the first movement. **c**. 90th-percentile tangling in the neural space [183], which quantifies how deterministic future states are from current states. **d-f**. Same as Panels a–c but for the motor output and output space. **g**. Mean-squared error (MSE) of output when noise of increasing magnitude is added to the neural activity. Line and shaded area, median and 95% confidence interval. Left: MSE for both x and y directions. Middle: MSE only for the x direction. Right: MSE only for the y direction. **h**. Same as Panel g but for the variance in position.

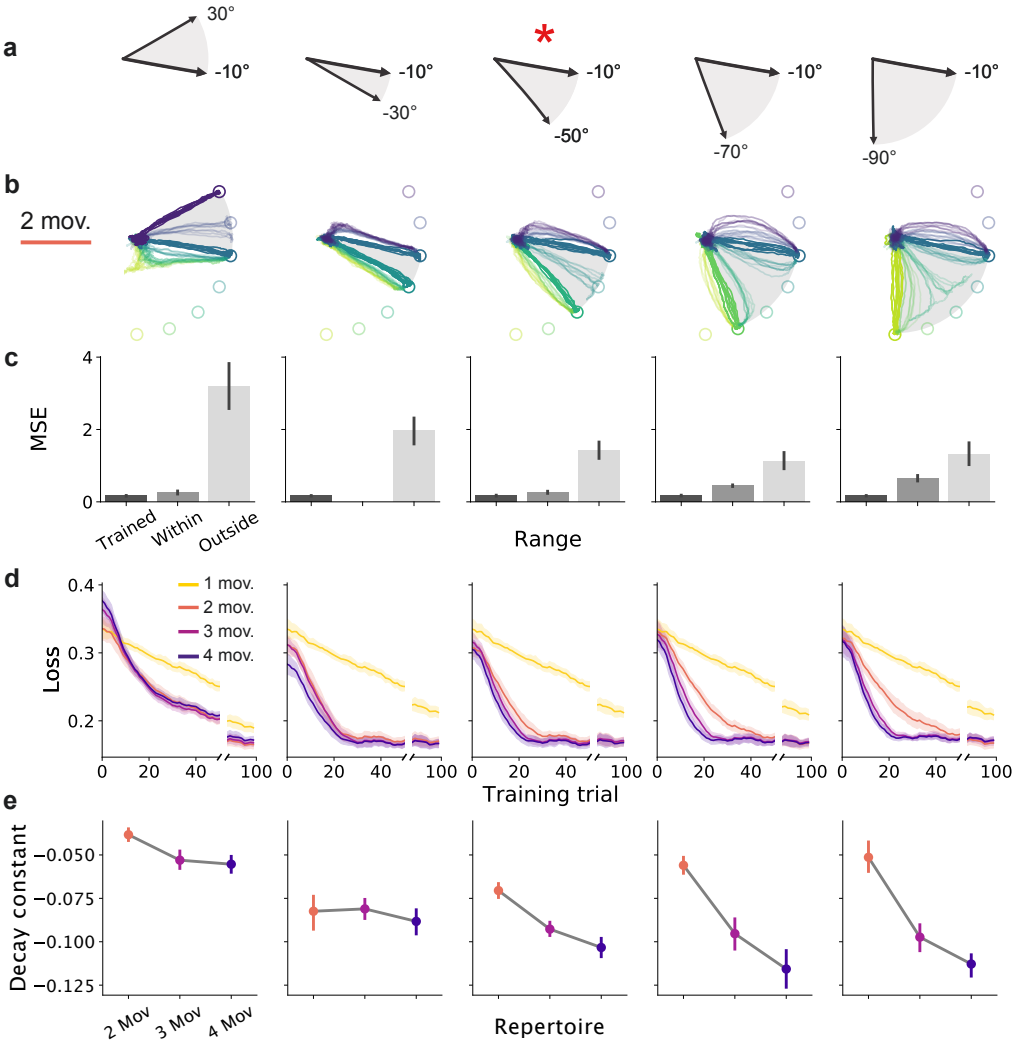


Figure S3.5: Networks can generalize and adapt to perturbations that require movements within a learned range. a. Networks were trained on repertoires with movements that spanned different ranges (top). The range -10° to -50°, denoted by an red asterisk, was used for all simulations in the main text. **b-c**. Networks trained on two-movement repertoires in their respective ranges (i.e. a movement to -10° and a movement to -50° for the -10° to -50° range) were tested on target cues for movements equally spaced between 30° and -90° to assess whether they could generalize to movements they were not trained on. The target cues were chosen such that networks were tested on movements they knew ('Trained'), movements that were within the range of known movements ('Within'), and movements that were outside the range ('Outside'). **b**. Motor output for the target cues for each movement for a sample network. Colors, target cues; circles, targets for each movement. Colors for target cues that the networks have not been trained on have lower opacity. Grey backgrounds denote the range of known movements. **c**. Mean-squared error between the network output and target positions for previously 'Trained' movements, and movements 'Within' or 'Outside' the range of known movements. Note that MSE was lower for 'Within' movements, showing that the networks can generalize. **d**. Loss during adaptation training with counterclockwise VR perturbations of 10°. Traces and shaded areas, mean and 95% confidence interval across networks of different seeds. **e**. Decay constants for exponential curves fitted to the loss curves in Panel d. Circles and error bars, mean and 95% confidence interval with bootstrapping.

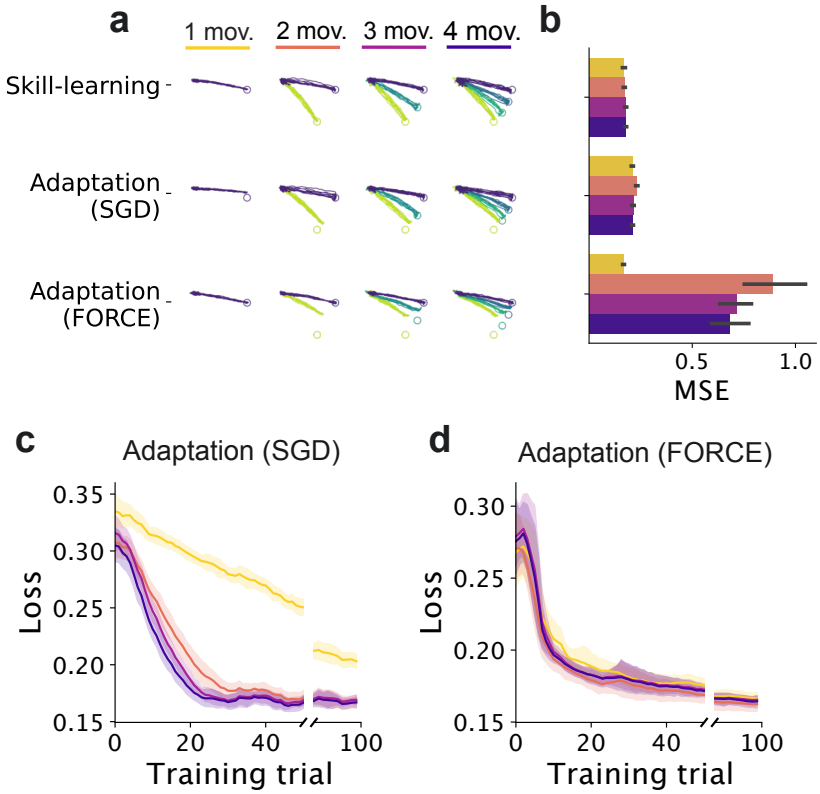


Figure S3.6: Adaptation training with stochastic gradient descent (SGD) is better at overcoming catastrophic forgetting than FORCE. Following skill learning, networks were trained to adapt to counterclockwise VR perturbations of 10° on one movement, using either SGD or FORCE. Motor output (**a**) and MSE (**b**) for all movements in known repertoires following adaptation training. Note that networks trained with FORCE had worse 'catastrophic forgetting' (that is, forgetting of previously trained tasks when learning new tasks) of other movements that were not perturbed during adaptation training. Loss during adaptation training with SGD (**c**) and FORCE (**d**). Traces and shaded areas, mean and 95% confidence intervals across networks of different seeds.

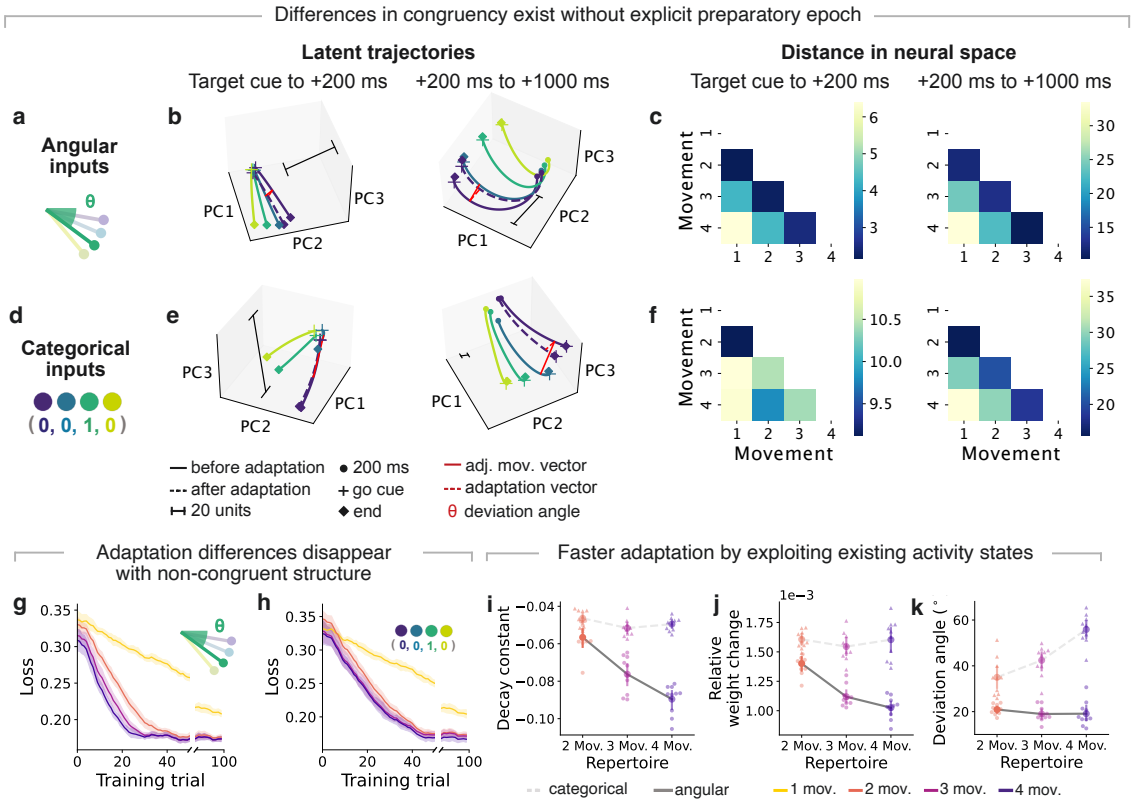


Figure S3.7: Explicit preparatory epochs are not necessary for differences in adaptation and structure. Networks were trained to produce movements without an explicit preparatory epoch. That is, networks started movement at the target cue rather than waiting for a go cue. Without an explicit preparatory epoch, networks still had preparatory-like activity in the first 200 ms after the target cue and execution-like activity for the next 800 ms after the target cue. **a–f**. Same measures as Figure 3.4a–f, but with the distances in neural space separated into the first 200 ms after the target cue and 200 to 1000 ms after the target cue. **g–k**. Same measures as Figure 3.4g-l.

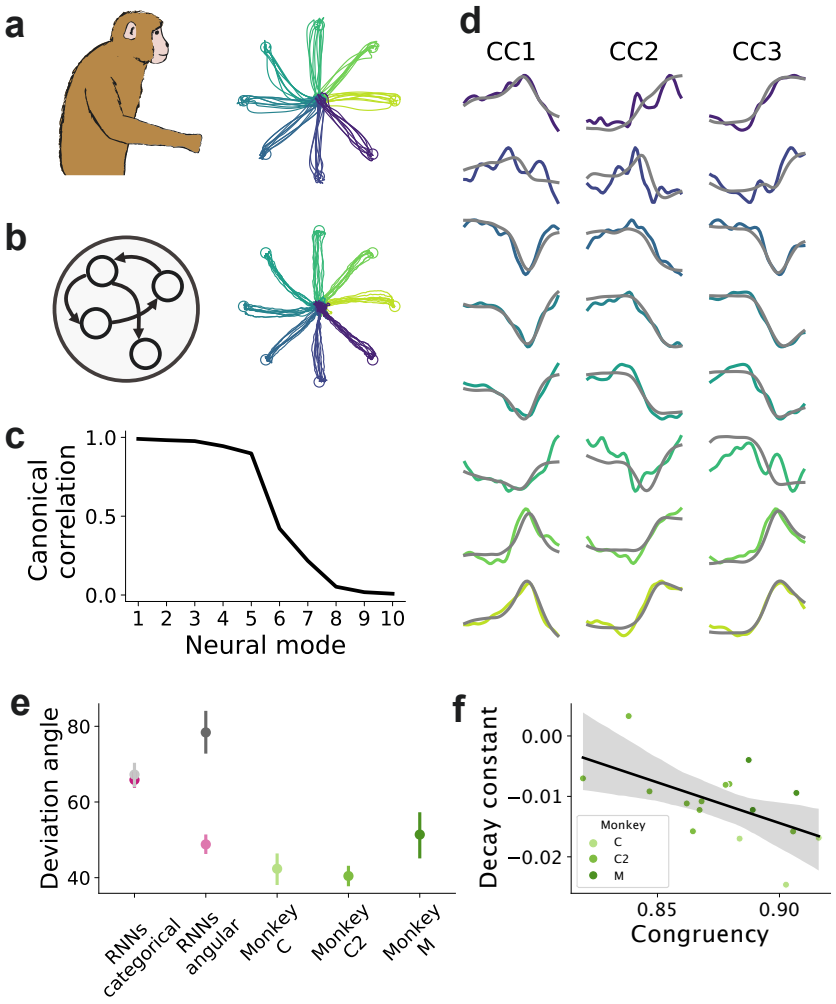


Figure S3.8: Simulation results reflect experimental monkey data. a. Hand trajectories produced by a monkey performing an eight-target center-out reaching task. Data from Ref. 14 (pooled from all sessions for Monkey M, see Methods). **b**. Simulated motor output when RNNs were trained on the same center-out reaching task with angular inputs. **c** Canonical correlation values between experimental and simulated latent dynamics. **d**. Projections of the experimental (colored) and simulated (grey) latent dynamics onto the first three axes identified through canonical correlation analysis. **e-f**. RNNs were trained on the center-out reaching task with either one-hot encoded ('categorical') or angular inputs, and compared to monkeys trained on the same task (Monkey C, Monkey C2, and Monkey M). Data from Ref. 14. **e**. The 'deviation angle' (Figure 3.4k) was calculated and pooled for the latent activity of all targets before and after adaptation to a visuomotor rotation. Circles and error bars, median and 95% confidence intervals with bootstrapping. Grey, control with shuffled targets and time points. **f**. Congruency between angular input structure and neural space structure (see Methods) compared to the decay constants fit to learning curves based on the angular reach errors.

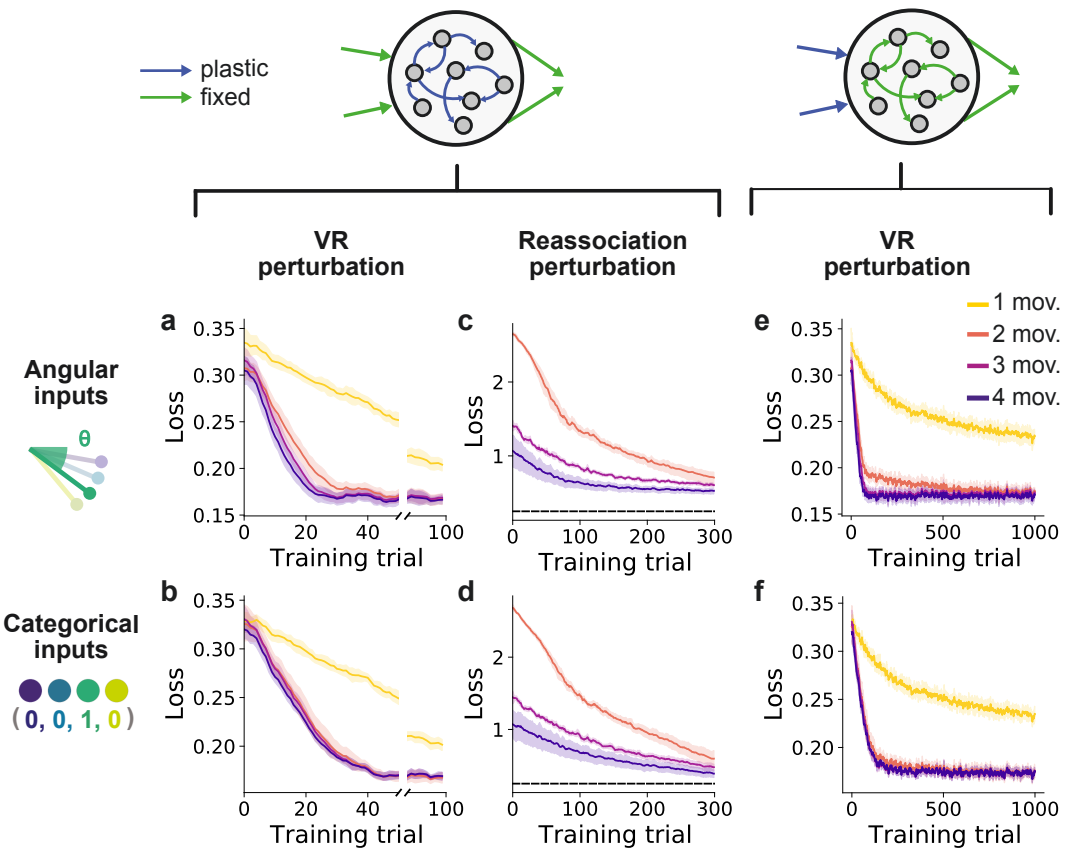


Figure S3.9: Adaptation with fixed input or recurrent weights. Loss during adaptation training with 10° counterclockwise VR perturbations with fixed input weights for networks with angular (**a**) and categorical (**b**) inputs (for 100 training trials). Traces and shaded areas, smoothed mean and 95% confidence intervals across networks of different seeds. **c, d**. Same as Panel a, b but for reassociation perturbations (for 300 training trials). Note that the results from Figure 3.5a,d held when the input weights were frozen such that the networks could not simply rely on changing the input weights to counteract reassociation. **e, f**. Same as Panels a, b but with fixed recurrent weights (for 1000 training trials). Note that the patterns remain the same as in the main text, where networks were trained with plastic input and recurrent weights, although at longer timescales.

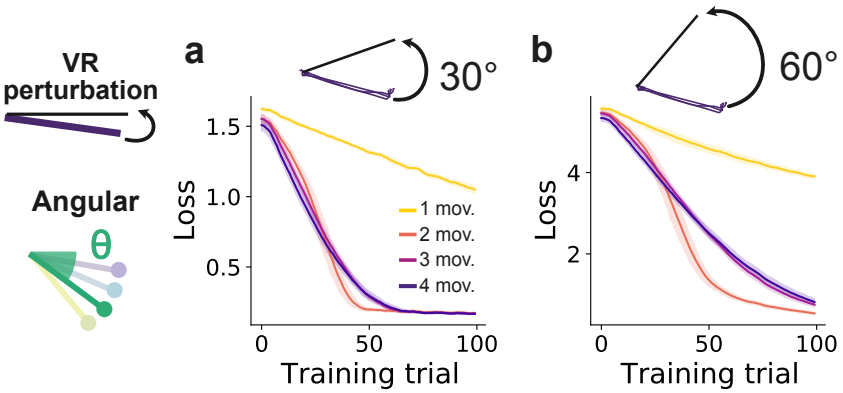


Figure S3.10: Structure in neural space hinders adaptation when larger changes are required. Networks were given angular inputs and adapted to 'more challenging' VR perturbations than the 10°rotations examined thus far. **a**. Loss during adaptation training for a VR of 30°. **b**. Loss during adaptation training for a VR of 60°. For both perturbations, smaller multi-movement networks adapted more quickly than larger multi-movement networks, contrary to previous results under smaller perturbations of 10° (Figure 3.3d). Note that this effect occurred when the movement needed to counter the adaptation was both within (for the 30° VR) and outside the range (for the 60° VR) of the known movements.

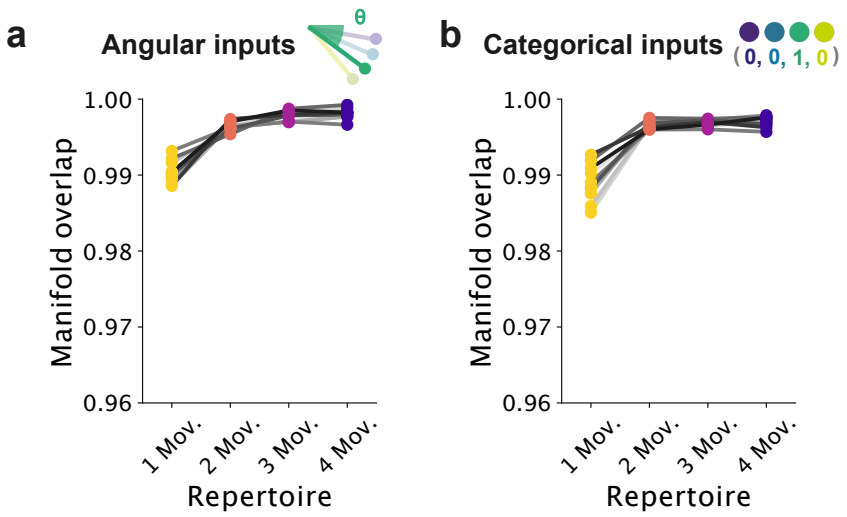


Figure S3.11: Adaptation to visuomotor rotation perturbation is within-manifold. Manifold overlap (see Methods) for network activity between skill learning and adaptation to a VR perturbation of 10° for networks with angular (**a**) and categorical (**b**) inputs. Line colors denote different random seeds.

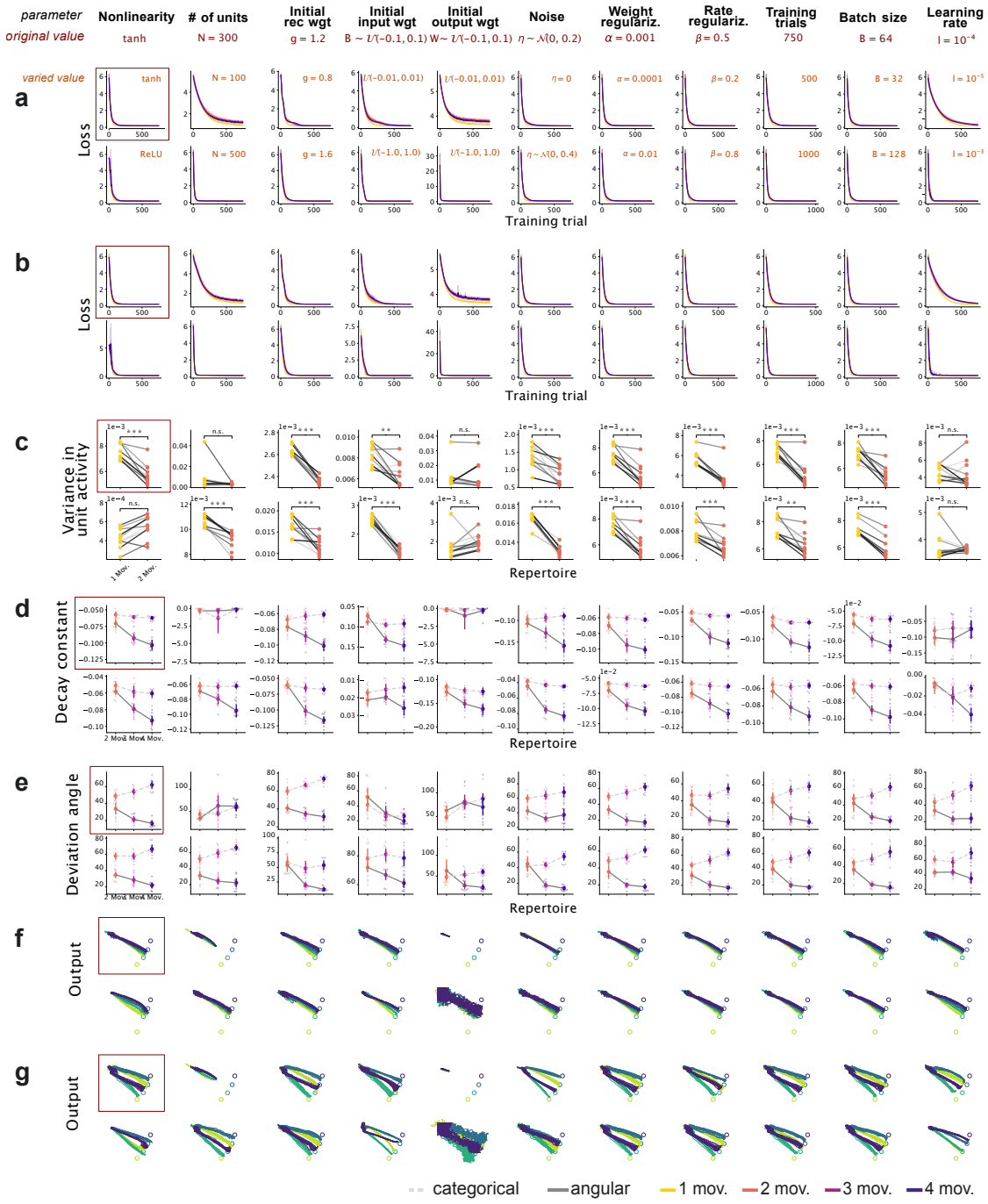


Figure S3.12: Main results remain robust to changes in modeling parameters. Parameters in the neural network model were varied one at a time and simulations were run with these changes. The parameter values in red denote the original value used in the simulations in the main text. The red boxed plots denote the results for the simulations in the main text. The values in orange in (**a**) denote the changed values. The trends in the results were generally robust to changes in parameters. **a**. Loss curves during initial training for networks with angular inputs. Presented in Figure 3.3b. **b**. Loss curves during initial training for networks with categorical inputs. **c**. Variance in unit activity for angular input networks. Presented in Figure 3.3c. **d**. Decay constants for angular and categorical input networks. Presented in Figure 3.4i. **e**. Decay constants for angular and categorical

input networks. Presented in Figure 3.4l. **f**. Output for reassociation perturbation for angular input networks. Decay constants for angular and categorical input networks. Presented in Figure 3.5c. **g**. Output for reassociation perturbation for angular input networks. Decay constants for angular and categorical input networks. Presented in Figure 3.5f.

Chapter 4

Motor memories in sequential learning

While latent dynamics need to be stable for animals to generate reliable behavior, they also need to flexible for animals to learn and adapt to changing environments. How do brains flexibly change their activity when learning new tasks without interfering with the stable control of familiar tasks? In this chapter, we examine how learning a new task changes the neural activity for a familar task, focusing on two neural signatures that have been recently identified during sequential learning experiments. Using neural network models, we explored how these signatures arise and how they relate to continual learning to promote flexible yet stable behavior.

The work in this chapter is presented in an article format as preprinted in *bioRxiv* in 2025[184]. The work has also been submitted for publication and is under review as of the submission of this thesis. I would like to thank my supervisors Claudia Clopath and Juan A. Gallego for their guidance and inspiration.

Neural signatures of motor memories emerge in neural network models

4.1 Abstract

Animals can learn and seamlessly perform a great number of behaviors. However, it is unclear how neural activity can accommodate new behaviors without interfering with those an animal has already acquired. Recent studies in monkeys performing motor and brain-computer interface (BCI) learning tasks have identified neural signatures—so-called "memory traces" and "uniform shifts"—that appear in the neural activity of a familiar task after learning a new task. Here we asked when these signatures arise and how they are related to continual learning. By modeling a BCI learning paradigm, we show that both signatures emerge naturally as a consequence of learning, without requiring a specific mechanism. In general, memory traces and uniform shifts reflected savings by capturing how information from different tasks coexisted in the same neural activity patterns. Yet, although the properties of these two different signatures were both indicative of savings, they were uncorrelated with each other. When we added contextual inputs that separated the activity for the different tasks, these signatures decreased even when savings were maintained, demonstrating the challenges of defining a clear relationship between neural activity changes and continual learning.

4.2 Introduction

Animals can acquire, retain, and execute many different behaviors in a lifelong cycle in which new behaviors are learned without interfering with familiar ones. Learning leads to changes in synaptic connectivity[10–12,141] and neural activity[16,94,112,185–189]. These changes can be retained on different timescales, from hours[190] to decades[191], and correlate with activity changes in different brain regions, from hippocampus[185] to the basal ganglia[192], cerebellum[187,188], and the motor cortices[16,37,94,112,189]. While it is evident that learning new behaviors leads to new patterns in neural activity, it is unclear how brains can accommodate these new neural activity patterns with old ones. Specifically, how do brains integrate activity patterns that improve future performance of newly learned behaviors—a phenomenon called savings[9] (Figure 4.1A)—without decreasing performance for older, familiar behaviors?

Several recent studies have begun to elucidate how brains incorporate motor memories of new behaviors by adopting a neural population view[94,147]. In this view, brain function is thought to be based on collective patterns of co-variation across a neural population rather than on the independent activity of its constituent neurons[13,134,193]. Since relatively few activity patterns are needed to explain the majority of variance in a neural population during a given task, neural population activity can be described mathematically based on a low-dimensional surface, a neural manifold[13]. Neural manifolds are thought to reflect biophysical (e.g., connectivity-related) constraints on neural activity[16,46,60,61] and have been leveraged to gain insights into the neural basis for learning[16,37,46,94,147].

Adopting this view, Sadtler et al[46] probed sequential learning using a brain-computer interface (BCI) paradigm where a linear map transformed neural population activity in monkey motor cortex to cursor movement, (Figure 4.1B). In each session, monkeys switched from a familiar Map A within their existing neural manifold (Task A1) to a new Map B (Task B1) and then back to the familiar Map A (Task A2). In different sessions, two different kinds of Map Bs were used to support the hypothesis that neural manifolds capture

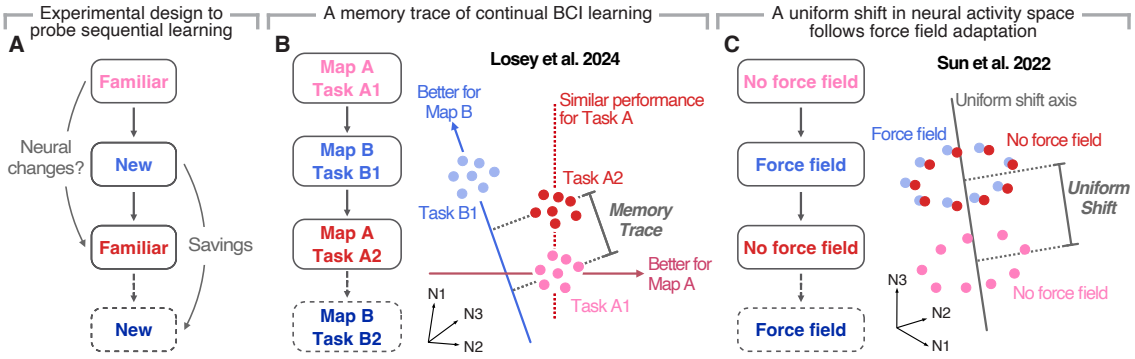


Figure 4.1: Potential neural signatures of learning. A) Learning a new task may cause changes in the neural activity used for a familiar task, and these signatures may help maintain knowledge about the new task to lead to retention savings if the new task is encountered again. **B)** Losey et al[147] studied a BCI learning task where monkeys performed a center-out reaching task with a familiar Map A (Task A1, pink), learned a new Map B (Task B1, light blue), then returned to the familiar Map A (Task A2, red). They identified a memory trace that made neural activity (colored dots) during movement execution for the familiar Map A more useful for a new Map B after learning this new Map B. Importantly, the memory trace existed without compromising the performance for Map A. **C)** Sun et al[94] studied a motor learning task where monkeys performed a baseline center-out reaching task (pink), learned to counteract a force field (light blue), then returned to the baseline task (red). They identified a uniform shift (colored dots) in the neural population space during movement preparation when monkeys learned to adapt to the force field that was sustained after returning to the baseline task.

constraints on neural population activity: Map Bs that required activity patterns within an existing manifold ("within-manifold perturbations") could be learned in a matter of minutes, whereas Map Bs that required activity outside the manifold ("outside-manifold perturbations") could not be learned within a session, requiring progressive learning over several days[16]. Focusing on within-manifold perturbations in this BCI paradigm, Losey et al[147] demonstrated that learning a new Map B altered the neural activity of the familiar Map A when monkeys were re-exposed to it (Task A2, Figure 4.1B). Crucially, this change made the activity during the familiar Map A more beneficial for the new Map B without negatively impacting performance of the familiar task (Figure 4.1B). The authors could show this directly by taking advantage of their BCI paradigm which, in contrast to traditional motor tasks where the relationship between motor cortex activity and behavior is not known, gave them full access to the relationship between neural activity and behavior since it is fully defined by the corresponding map: when they "projected" Task A2 activity onto Map B offline, it led to better performance for Map B than Task A1 activity. Thus, they proposed that Task A2 activity retained a "memory trace" of the new Map B that did

not compromise the performance for the familiar Map A[147].

Using a classic motor learning task in which participants—in this case, monkeys—performing reaching movements need to learn to counteract a force field[131], Sun et al[94] identified a "uniform shift" in the motor cortical population activity during movement preparation that emerged during learning. Similar to the "memory trace" in the previous BCI study, this shift was sustained after monkeys returned to the baseline task without a force field (Figure 4.1C). Since this shift was in an axis orthogonal to the directions in the neural manifold that were predictive of motor output, the authors argued that it could potentially index a motor memory of the new force field task, again in a way that would not compromise performance for the familiar baseline task[94]. Together, these BCI learning and motor learning studies suggest that learning new tasks may produce changes in neural activity that incorporate information about the new task in the activity of familiar tasks, potentially leading to savings if the new task is encountered again (Figure 4.1A). As a result, these signatures may be important for continual learning, that is, for learning tasks sequentially without forgetting previously acquired knowledge.

Here, we aimed to examine when these two potential signatures—the sustained uniform shift during preparation identified by Sun et al[94] and the memory traces during execution identified by Losey et al[147]—arise and how they may relate to continual learning. To this end, we studied both signatures during the same learning paradigm using recurrent neural networks (RNNs). We used RNNs because similar models have been shown to replicate key features of neural activity and motor output from experimental recordings[49–51,142,194], enabling investigation of scenarios that are difficult to probe experimentally[128,194,195].

We modeled an adapted version of the sequential learning BCI task described above, which allowed us to compare both the behavior and the changes in network activity—including the two potential signatures of motor memories we are interested in—to those observed in monkey motor cortex in the original studies. We found that both sustained uniform shifts and memory traces emerged naturally in our RNNs, without modeling any explicit additional mechanisms to force them to arise. To relate these signatures

to continual learning, we examined a common way to measure the extent of continual learning: savings, or improved relearning of a task upon re-exposure, compared to the initial exposure[9]. Savings can be observed as either an increase in performance due to better retention immediately after re-exposure ("retention savings"), or a faster rate of learning during prolonged exposure ("learning rate savings"). Here, we focused on retention savings since learning rate savings might be confounded by different initial errors for our models trained with gradient descent. While the magnitude of the memory traces were closely correlated with retention savings, that of the uniform shift was not correlated with neither the magnitude of the memory traces nor the retention savings. Extending Losey et al's[147] analyses to outside-manifold perturbations, we saw that memory traces were larger for outside-manifold perturbations when compared to within-manifold perturbations. This was due to greater forgetting in within-manifold perturbations when reverting back to the familiar map in Task A2. Inspired by behavioral studies on context-dependent learning[156,196], we added a context cue to better separate activity patterns and reduce interference to prevent forgetting. With context cues, a uniform shift was observed during learning of the new map in Task B1, but it did not persist during Task A2. Similarly, the magnitude of memory traces decreased and was no longer indicative of savings, even though there was still increased savings of the new map. Therefore, uniform shifts and memory traces arise naturally when learning a new task that shares activity patterns with a familiar task, but these signatures of learning for a novel task are no longer observed in the activity of the familiar task if they are separated by features like strong context cues.

4.3 Results

An RNN model of sequential learning naturally recapitulates experimental results

To understand how memory traces[147] and uniform shifts[94] arise and to explore their features in the context of continual learning, we used RNNs to model the same BCI

experiments as in Losey et al[147] (Figure 4.2A). In these experiments, monkeys used their motor cortical activity to control a computer cursor through a BCI map that translated neural activity onto cursor velocity. To study how neural activity changed across different tasks, monkeys had to control the BCI to produce center-out reaches using different BCI maps that were presented in a sequential block design.

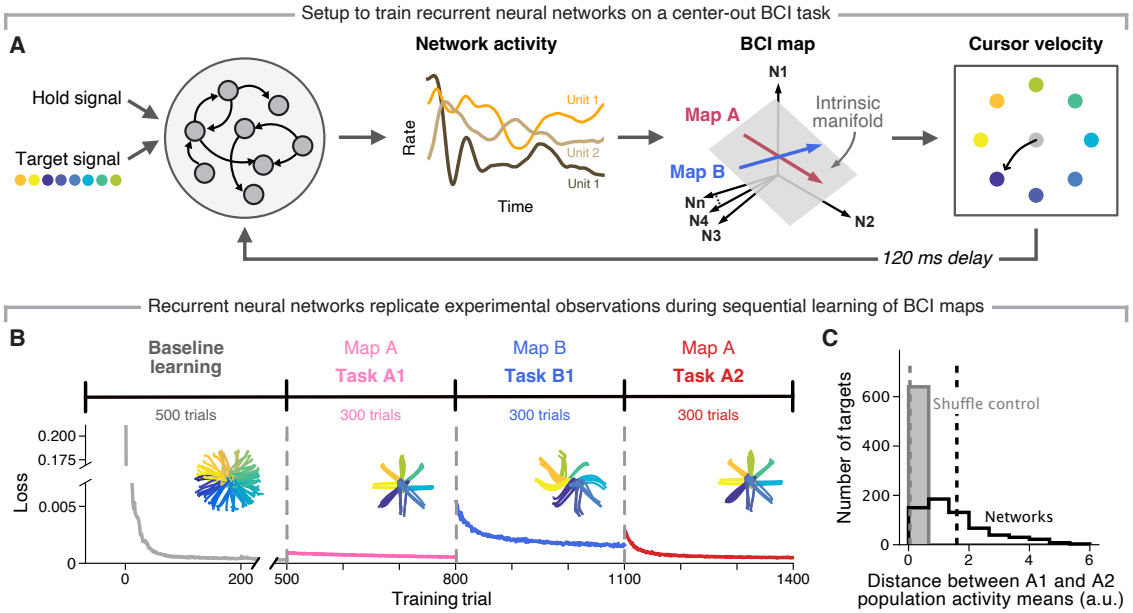


Figure 4.2: RNNs recapitulate experimental results in sequential learning. **A**) Networks were trained to perform center-out reaches by using different BCI maps which linearly transformed network activity to output velocities. Map A was the intuitive map based on the "intrinsic manifold" –a low-dimensional flat surface that captured the dominant covariation patterns of the network activity after baseline learning–, while Map B was a permutation of Map A. Produced velocities were fed back into the network after a 120 ms delay. **B**) Sequential training procedure: we first trained the recurrent, input, and output network weights during an initial training phase where networks had to produce random center-out reaches. Then, networks were trained to produce center-out reaches to eight fixed targets using different BCI maps (A or B) as a fixed output layer. Networks learned to produce output through these maps in blocks: from Map A (Task A1), to Map B (Task B1), then back to Map A (Task A2). Loss curves during each task block for an example network. Inset: position trajectories at the end of each block. **C**) Mahalanobis distance between population activity means for Task A1 and Task A2 for each target (black), and for controls where the task labels were shuffled (gray). Dotted lines, means across 10 different maps for each random seed (n=8 random seeds).

Here, we first used gradient descent to pretrain RNNs to produce velocities for varied delayed center-out reaches to model the monkeys' ability to produce the necessary neural activity for controlling their arms (Figure 4.2B). We incorporated a delay period to examine possible uniform shifts during preparation, even if the original BCI experiments did

not have one. RNNs were also given delayed feedback of their produced velocities[194] (Figure 4.2A), and their performance was quantified as the mean-squared error (MSE) between the produced and synthetic target velocities (Methods). Following this initial training, RNNs were trained to perform the eight-target center-out BCI task using different BCI maps, which were modeled as fixed output weights. Each BCI map consisted of two transformations. First, activity from all 400 units in our RNNs was projected onto an eight-dimensional flat neural manifold, which we identified using Principal Component Analysis (PCA) and thus captured the majority (89-91%) of the variance in the network activity. Second, activity within this manifold was projected to the two-dimensional subspace defined by the corresponding BCI map to obtain the output cursor velocity along the horizontal and vertical axes (Figure 4.2A).

To probe whether and how learning a new task affected the neural activity of a familiar task, the RNNs were trained on the two different maps in sequential blocks: first with Map A (Task A1), then with a new Map B (Task B1), then back to the familiar Map A (Task A2, Figure 4.2B). Replicating the experiments[46], Map A was determined based on the "intuitive manifold" map that captured the majority of the variance in the activity following baseline training, and Map B was a within-manifold perturbation of Map A that required activity in the same low-dimensional neural manifold (Methods).

The RNNs were able to learn Map A and Map B in sequence, with similar rates of learning and performance as the experimental results[46,147] (Figure 4.2B). Recapitulating experimental results, neural activity patterns for a given target changed from Task A1 to Task A2 in our models, even if the mapping during Tasks A1 and A2 was the same[147] (Figure 4.2C, Figure S4.1A). These changes were significant across different targets and significantly greater than a shuffled control (Figure 4.2C, $P = 8.9 \cdot 10^{-107}$, one-sided Wilcoxon signed-rank test). Thus, learning a new task (even a new version of the same task, in this case defined by Map B) altered the RNN activity for a familiar task without the need for an additional learning mechanism.

Uniform shifts and memory traces arise from sequential learning without adding explicit mechanisms

Examining the properties of these activity changes revealed that they could be partially attributed to a uniform shift in the preparatory population activity from Task A1 to Task B1, which was orthogonal to the directions in neural space predictive of motor output (Figure 4.3A, Methods), as experimentally observed in monkey motor cortex[94]. These activity shifts generally persisted in Task A2 (Figure 4.3A-C; Figure S4.2A,B), further replicating experimental results[94]. Due to these changes, the different blocked tasks (Task A1, Task B1, and Task A2), could be easily distinguished based on their preparatory unit activity alone using linear discriminant analysis classifiers (Figure S4.1A), demonstrating that the tasks were indeed using different activity patterns.

Having established that uniform shifts emerge naturally as a result of continual learning in RNNs using standard training techniques, we next asked whether we could observe similar memory traces as those found in experimental recordings[197]. Specifically, we were interested in whether memory traces spontaneously arise from our simple sequential learning model, or if an additional explicit mechanism was necessary for them to arise.

The advantage of a BCI setup is that we can probe how useful the neural activity for a familiar map is for a new BCI map before and after learning the new BCI map. While Task A1 and Task A2 used the same Map A online, we could project their activity through Map B offline to evaluate how much the activity changed from Task A1 to Task A2 to accommodate new activity patterns that were directly useful for Map B. In other words, we asked how much learning Map B left a memory trace of this new task in the activity of the familiar task defined by Map A (Figure 4.3E). When projected through Map A, Task A1 and Task A2 activity both produced output position trajectories that were similar to the target output, without a clear difference in performance (example in Figure 4.3E). However, when projected through Map B, Task A2 activity produced output trajectories that were closer to the target output than that of Task A1 activity (Figure 4.3E), indicating the existence of a memory trace of Map B. We quantified this memory trace as the increase

in performance, or decrease in MSE, for Map B from Task A1 to Task A2 (Figure 4.3F). From here, we will denote "Task X activity projected through Map Y" as "Task X→Map Y", such that the memory trace is the decrease in MSE from Task A1→Map B to Task A2→Map B.

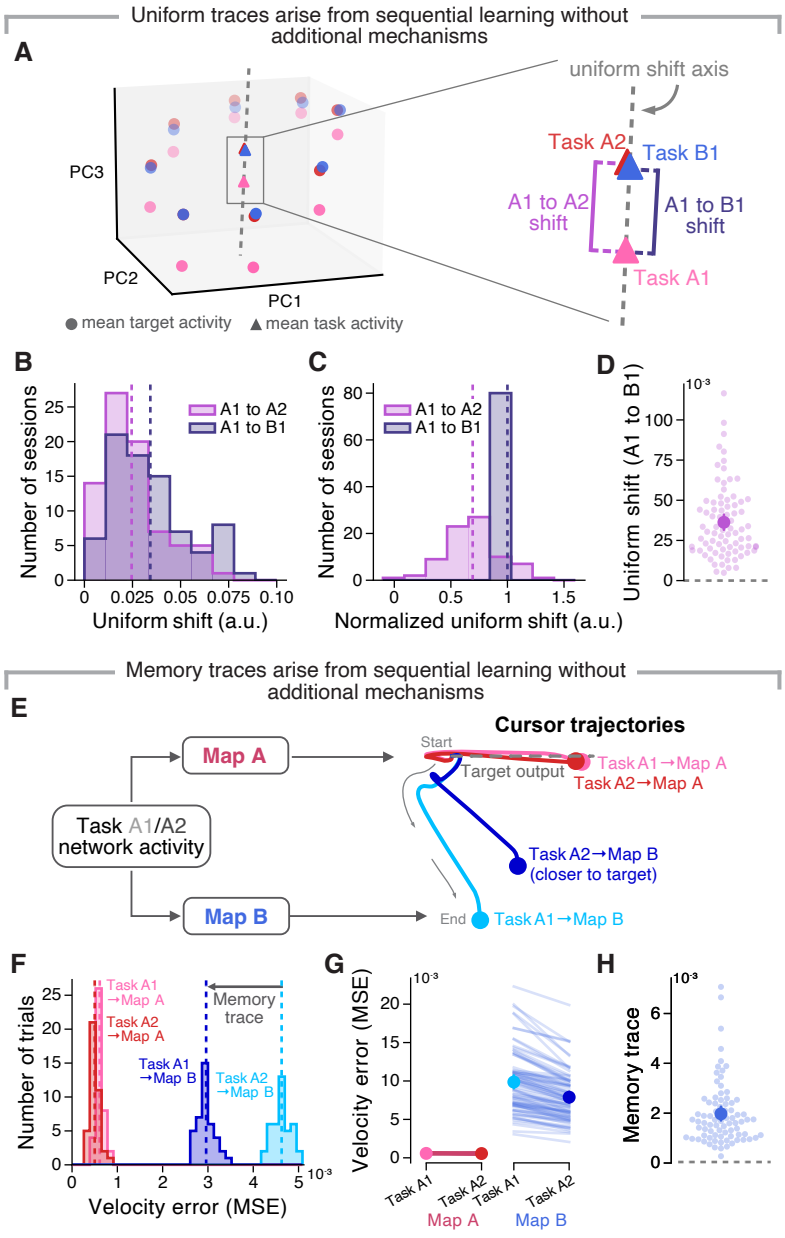


Figure 4.3: Uniform shifts and memory traces arise from sequential learning without explicit mechanisms. A) Population activity during preparation (first 200 ms of each trial) for an example session. Circles, time-averaged, trial-averaged activity for each target (mean target activity) in each task block. Triangles, time-, trial-, and target-averaged activity (i.e., mean task activity) in each block. The uniform shift axis is quantified by the direction of the vector connecting the mean task activity from Task A1 to Task B1. The uniform shift across tasks is calculated by projecting the

mean task activity onto the uniform shift axis and calculating the resulting distance between tasks. **B**) Uniform shift from Task A1 to Task B1 and Task A1 to Task A2. Dotted lines, means across 10 different maps for each random seed. **C**) Normalized uniform shift, calculated by dividing by the uniform shift from Task A1 to Task B1 for each session. **D**) Uniform shift from Task A1 to Task B1 across targets. Big circle and error bars, means and 95% confidence intervals with bootstrapping for 10 different maps for each random seed (n=8 random seeds); small circles, different maps. Note that uniform shifts are consistently greater than zero (dotted line) indicating a shift in population activity. **E**) Task A1 and Task A2 activity were projected into Map A online or into Map B offline. The resulting output was used to assess how useful the activity is for each map, i.e., how close to the target output it brings the cursor. **F**) MSE between the produced and target output for each projection for an example target in an example session. Dashed lines, mean MSE across trials for that target. The memory trace is quantified as the decrease in mean MSE (i.e., the increase in performance) from Task A1→Map B to Task A2→Map B. **G**) MSE across all targets for projections onto the two different maps. Circles and error bars, means and 95% confidence intervals with bootstrapping for 10 different maps for each random seed (n=8 random seeds). Traces, different sessions. **H**) Same as D but for the memory trace. Note that memory traces are consistently greater than zero (dotted line) indicating increased offline performance for Task B with activity used for Task A2 compared to that for Task A1.

We observed that memory traces consistently arose across different targets (Figure 4.3G,H; Figure S4.3A). These increases in performance for Map B from Task A1 to Task A2 were significantly greater than a control based on the performance for Map A (Figure 4.3G, $P = 3.9 \cdot 10^{-15}$, one-sided Wilcoxon signed-rank test), demonstrating that the presence of these memory traces did not result in a trade-off or loss of performance for Map A. This indicates that information from both maps coexisted in the activity of the familiar task as RNNs performed Task A2. Importantly, these memory traces could be specifically attributed to the process of learning Map B: when we reset the input and recurrent network weights after Task B or randomly reallocated the weight changes that occurred during Task B, the memory traces disappeared (Figure S4.3B,C). As a result, our models showed that both sustained uniform shifts and memory traces can arise spontaneously through sequential learning, potentially allowing the networks to fit both tasks in their activity repertoire even without any added mechanisms to reduce interference.

The magnitude of memory traces, but not of uniform shifts, correlates with savings and initial behavioral error

Having established that our RNN models readily reproduce potential neural signatures of motor memories observed during BCI learning[147] and motor learning[94] experiments, we asked whether these signatures have a clear relationship with behavior.

First, we studied whether they captured the formation of a motor memory. Behaviorally, exposure to a perturbation leads to the formation of a motor memory, which is expressed by increased task performance upon re-exposure. One way motor memories manifest is through retention savings, defined as the increase in task performance immediately after (re-)exposure. We quantified these retention savings by comparing performance when networks were re-exposed to Map B during a new Task B2 compared to initial exposure to this same map during Task B1. Interestingly, the magnitude of the memory trace was closely related (Figure S4.4A) and highly correlated with the magnitude of the savings (Figure 4.4A, Figure S4.5D), but the magnitude of the uniform shift was not (Figure 4.4C, Figure 4.4I). Thus, even if both signatures spontaneously arose during continual learning, their relationship with behavior is different, and only the memory trace seems to be related to immediate retention savings.

Having defined these relationships, we asked whether we could predict their magnitude from task properties alone. The BCI learning paradigm, combined with our modeling setup that allows us to run many "experiments", offers a unique opportunity to answer this question. We characterized the Map B perturbations based on the directions of the neural manifold they affected (Figure S4.5A), since recent theoretical work suggests that large variance directions may be easier to control than their lower variance counterparts[162]. While "high variance perturbations" led to larger memory traces than "low variance perturbations" (Figure S4.5B), such difference in variance did not affect the magnitude of the uniform shifts (Figure S4.5C). Other measures such as the principal angles between BCI maps, or their relative amount of neural variance explained also had inexistent or weak associations with uniform shifts and memory traces (Figure S4.5E,F,G,J,K,L). In contrast,

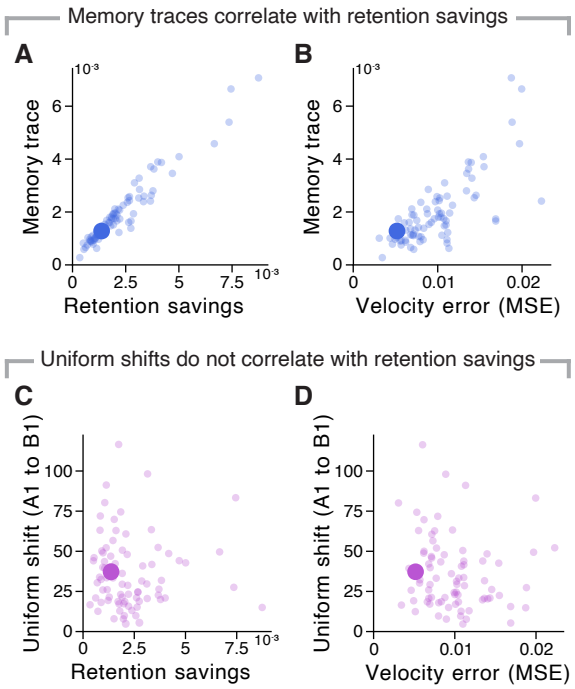


Figure 4.4: The magnitude of memory traces, but not that of uniform shifts, correlates with retention savings. A) Mean memory trace across targets compared to the retention savings. Small circles, mean for each map (n=10 maps) for each random seed (n=8 random seeds); large circle, example shared across A-D. **B)** Mean memory trace across targets compared to the initial velocity error (MSE) when Task A1 activity is projected into the new Map Bs before learning in Task B1 (i.e. Task A1→Map B MSE). **C)** Uniform shift from Task A1 to Task B1 across targets compared to the retention savings. **D)** Uniform shift from Task A1 to Task B1 across targets compared to the initial velocity error (MSE).

the initial behavioral disruption imposed by the perturbation was a strong predictor of the magnitude of the memory trace (Figure 4.4B, Figure S4.5H), although it did not predict that of the uniform shift (Figure 4.4D, Figure S4.5M). This suggests that more information is retained about a given task as measured by the memory trace if the task requires more learning at the behavioral level.

Experimentally hard-to-learn outside-manifold perturbations lead to larger memory traces and smaller uniform shifts

We have studied uniform shifts and memory traces for BCI maps that impose within-manifold perturbations that require new activity patterns within the existing neural manifold. Here, we extended our investigation to outside-manifold perturbations that require

activity patterns that are not available within the existing manifold (Figure 4.5A). Behaviorally, these perturbations are very different: while within-manifold perturbations can be learned within a few minutes[46], outside-manifold perturbations cannot, requiring several days[16].

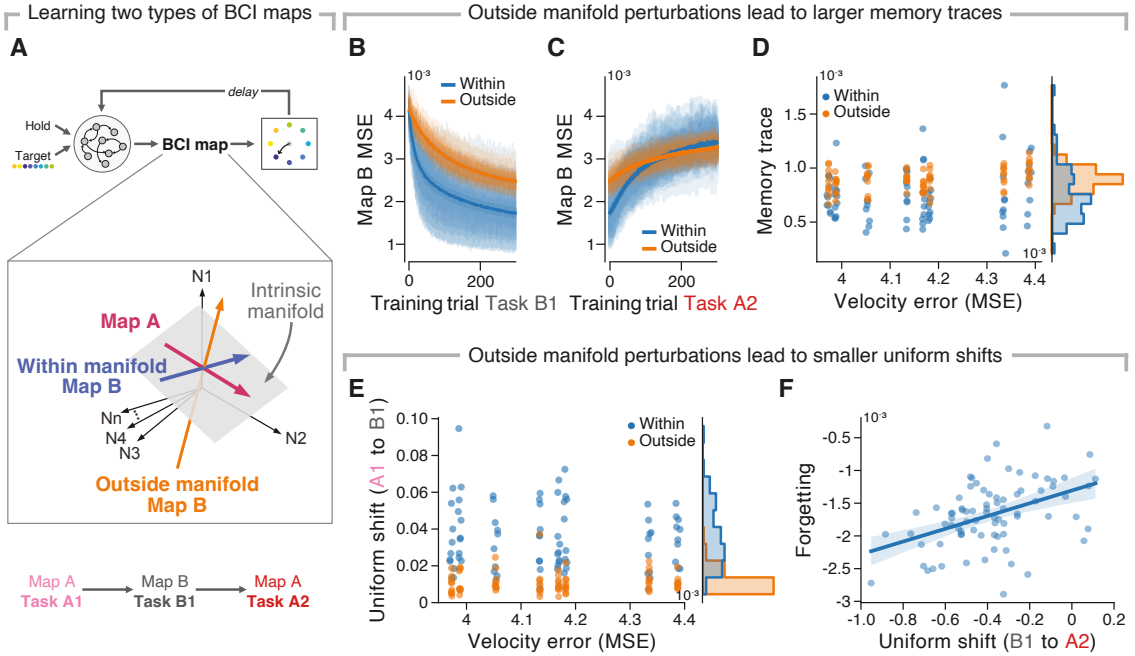


Figure 4.5: Experimentally hard-to-learn outside-manifold perturbations lead to larger memory traces and smaller uniform shifts. A) Networks were trained on different Map Bs representing different within- and outside-manifold perturbations that had the same Task A1→Map B MSE. **B)** Map B MSE during initial learning of the new Map B in Task B1. Lines and shaded surfaces, mean and 95% confidence interval across 10 maps for each perturbation across networks of different seeds (n=8 random seeds). Lines, different maps for different seeds. **C)** Map B MSE during reversion to the familiar Map A in Task A2. **D)** Mean memory trace across targets compared to the Task A1→Map B MSE. Circles, mean for each map for each random seed. Note that points with the same Task A1→Map B MSE belong to the same seed since this was controlled for. **E)** Uniform shift from Task A1 to Task B1 across targets compared to the Task A1→Map B MSE. **F)** Uniform shift from Task B1 to Task A2 compared to forgetting, defined as the decrease in performance from Task B1→Map B to Task A2→Map B, for within-manifold perturbations. Lines and shaded surfaces, mean and 95% confidence interval across different maps for different seeds; circles, different maps for different seeds.

Having established that the initial task error was a strong predictor of memory trace magnitude for within-manifold perturbations, we asked how uniform shifts and memory traces change when we control for this factor for both inside- and outside-manifold perturbations (Figure 4.5A). To do so, we projected Task A1 activity into thousands of within-manifold perturbation and outside-manifold perturbation BCI maps and selected a subset of maps

of each type with the same MSE to use as the new Map Bs ($n = 10$ for each perturbation type, Methods, Figure S4.6A). Recapitulating experimental findings[46], outside-manifold perturbations were harder to learn than within-manifold perturbations, with larger errors at the end of Task B1 when the same number of training trials were used to learn the maps (Figure 4.5B, $P = 8.2 \cdot 10^{-30}$, one-sided Student's t-test).

Learning outside-manifold maps less well would suggest smaller initial retention savings of these maps compared to within-manifold maps. However, when we examined the memory traces at the end of Task A2, outside-manifold perturbation led to surprisingly larger memory traces and savings than within-manifold maps (Figure 4.5D, $P = 2.3 \cdot 10^{-7}$, one-sided Mann-Whitney U test; Figure S4.6C, Figure S4.7A).

To reconcile these findings, we examined performance for the new maps during re-exposure to the original map (i.e., Task A2) and saw that the performance for within-manifold maps dropped more quickly than that for outside-manifold maps (Figure 4.5C). This suggests that during Task A2, information about within-manifold maps was more easily overwritten and forgotten, likely because these maps were too similar to coexist together without interference. To directly examine this, we trained networks initialized from the same random seeds to learn both Map A and Map B simultaneously right after initialization. RNNs had more difficulty learning two within-manifold maps simultaneously than one within-manifold and one outside-manifold map, showing that it was harder for within-manifold maps to coexist (Figure S4.7B). This supports our interpretation that the larger memory trace after learning an outside-manifold perturbation mapping happened because, although these maps were learned less well in the first place, more information about the new maps was retained after switching back to the familiar task (Figure 4.5C), leading to greater retention savings. The greater savings during outside-manifold perturbations were again correlated with greater memory traces (Figure S4.6B; note that this effect did not result from a tradeoff in performance for the familiar map, Figure S4.8). In contrast, our second potential neural signature of motor memory exhibited the opposite behavior: the magnitude of the uniform shift from Task A1 to Task B1 was smaller for outside-manifold perturbations than within-manifold perturbations (Figure 4.5E, $P = 1.6 \cdot 10^{-22}$, one-sided

Mann-Whitney U test; Figure S4.2E,F), and was not directly correlated with retention savings (Figure S4.6H). However, the degree of forgetting of Map B was correlated with retraction along the uniform shift axis during Task A2 for within-manifold perturbations (Figure 4.5F), showing that, the degree of retention of the uniform shift, rather than its absolute length, was indicative of savings. Together, these results show that more similar tasks, like within-manifold perturbation maps, may be initially easier to learn, but may also cause greater interference and less retention during continual learning.

Explicit context cues alter memory traces, uniform shifts, and their relationship to savings

Since forgetting due to interference was a driving factor in reducing savings for within-manifold perturbations, we examined ways to reduce interference to promote continual learning. Thus far, we have not introduced any explicit mechanisms to separate the activity patterns for different tasks. However, experimental work shows that humans and monkeys are better at learning tasks in blocks than when they are interleaved[94,198,199], a phenomenon that is likely due to the incorporation of contextual information[156,196,199]. Based on this, we predicted that explicitly cuing our RNNs on the identity of the task would allow different task-specific activity patterns to be learned sequentially without overwriting. This intuition was built on the notion that an explicit cue signal would promote activity patterns for the different tasks to emerge in different parts of activity space.

To test this intuition, we added explicit context cues as inputs to our networks that indicated whether they were controlling the familiar Map A or the new Map B using one-hot encoded binary vectors (Figure 4.6A). We trained networks on the same sequential learning setup as before, but now with this additional context cue: Context A was used during baseline learning, Task A1, and Task A2, while Context B was used during Task B1. Using this setup, we asked how continual learning was modulated by the degree of discriminability between contexts—since such discriminability will likely indicate the degree that activity is driven to occupy different parts of state space—, and how this would translate into differences in

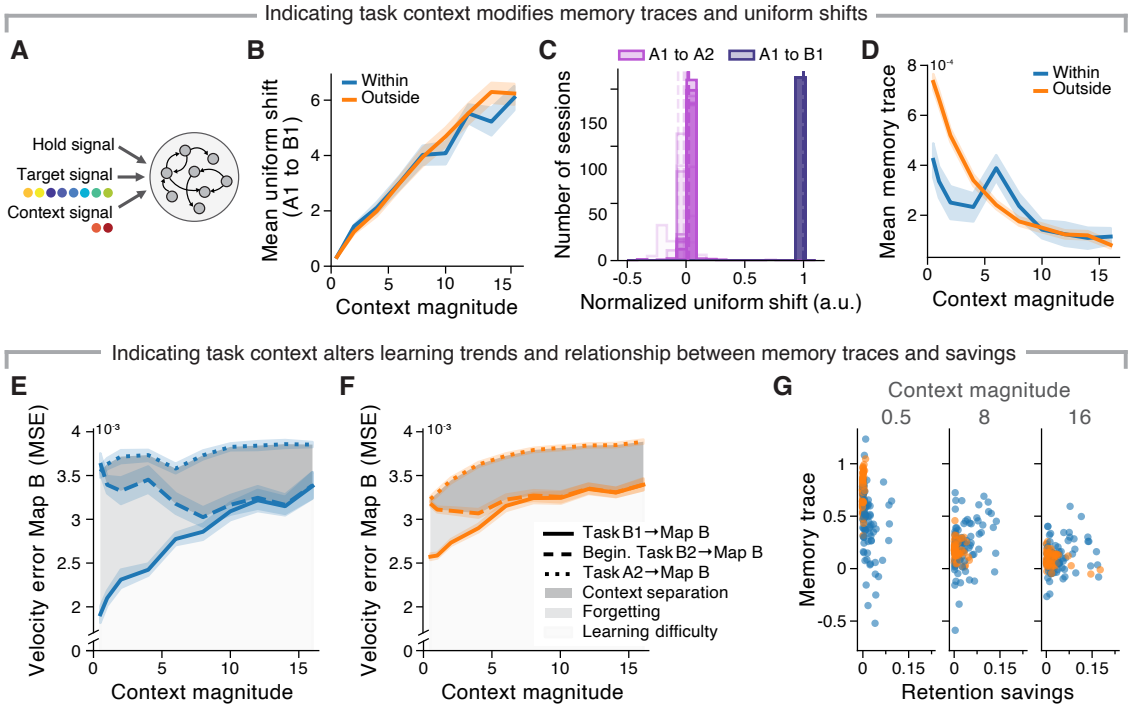


Figure 4.6: Context cues alter uniform shifts and memory traces. A) Networks were given an additional context cue that signaled whether the familiar Map A or the new Map B was being used. The context signal was a one-hot encoded vector, and the vector was multiplied by different factors to create different context magnitudes. **B)** Uniform shifts from Task A1 to Task B1 for different context magnitudes. Line and shaded surfaces, mean and 95% confidence interval across 10 maps for each perturbation across networks of different seeds (n=8 random seeds). **C)** Normalized uniform shifts from Task A1 to Task B1 and from Task A1 to Task A2, averaged across targets for each session. Shifts were normalized by the length of the shift from Task A1 to Task B1. Opacities, different context magnitudes (most transparent for smallest magnitude). Dashed lines, mean for each context magnitude. **D)** Memory trace for different context magnitudes. Line and shaded surfaces, mean and 95% confidence interval across maps for each perturbation across networks of different seeds. **E)** MSE when activity from the end of Task B1, the beginning of Task B2, and the end of Task A2 are projected into Map B for different context magnitudes for within-manifold perturbations. Task B1→Map B MSE indicates learning difficulty, the difference between Task B1→Map B and beginning of Task B2→Map B MSE indicates forgetting, and the difference between the beginning of Task B2→Map B and Task B1→Map B MSE indicates activity changes due to context. **F)** Same as E but for outside-manifold perturbations. **G)** Mean memory trace across targets compared to the retention savings. Small circles, mean for each map for each random seed.

uniform shifts and memory traces. We controlled the degree of discriminability between contexts by multiplying the binary vectors representing the context (Map A or Map B) by different factors to create context cues of varying magnitudes, and separately trained networks with each context magnitude.

As a result of introducing context, the trends in uniform shifts and memory traces that

we previously saw dramatically changed. First, uniform shifts were still evident when learning the new map, and the magnitude of the shifts increased with greater context magnitudes (Figure 4.6B, Figure S4.2G,Figure S4.9A). However, the uniform shifts were not sustained: upon returning to the familiar Map A in Task A2, the uniform shift disappeared (Figure 4.6C, Figure S4.2G). This suggests that context signals pushed activity for the new map to occupy different spaces from the familiar map, but these changes were not incorporated into activity for the familiar map. Thus, context can lead to a dramatic change in this signature of learning.

Second, the inclusion of context cues also changed the trends in memory traces. Initially, with small context magnitudes, the trends were similar to those in networks without any context cues (Figure 4.5): outside-manifold maps led to greater memory traces than within-manifold maps (Figure 4.6D). Moreover, even though within-manifold perturbations were easier to learn during Task B1, there was also greater forgetting during Task A2, leading to greater Task A2→Map B MSE and smaller memory traces overall[1] (Figure 4.6E,F).

As the context magnitude increased, however, the magnitude of the memory traces became indistinguishable for outside-manifold perturbation and within-manifold maps (Figure 4.6D). This trend was caused by larger context magnitudes leading to more disparate activity patterns which occupied a different part of state space ((Figure S4.10), making learning new maps more difficult. With less learning (Figure 4.6E,F), there was less knowledge about the new map that could be retained, contributing to a general decrease in memory traces with larger context magnitudes (Figure 4.6 E,F). With activity patterns in a different part of activity space due to the increased context magnitude, there was also less overwriting and forgetting of Map B during Task A2 (Figure 4.6E,F). Thus, greater context magnitudes made within-manifold perturbations more similar to outside-manifold perturbations: both were equally difficult to learn and equally resistant to forgetting, with essentially perfect retention savings (Figure 4.6E,F). However, unlike our previous net-

[1]Recall that the memory trace is quantified as the increase in performance, or decrease in MSE, from Task A1→Map B to Task A2→Map B, so it is inversely proportional to the Task A2→Map B MSE when we have controlled for the initial output error, the Task A1→Map B MSE. We can thus use the trends in Task A2→Map B MSE to examine the trends in memory traces in Figure 4.6E and F).

works without context cues, here the savings were no longer closely correlated with the memory trace (Figure 4.6G, Figure S4.9, Figure S4.4A). With context however, the context cue shifted Task B2 into a different part of state space from Task A2, such that Task A2 activity became less relevant for Map B (Figure S4.4A). This context separation made larger context magnitudes lead to smaller memory traces of Task B even when there was perfect retention savings. Together, these results show that there can still be motor memories of a new task that lead to savings even when the uniform shift is no longer sustained or if the memory trace is decreased. With little or no context, memory traces were good indicators for initial retention savings and the degree of retention of the uniform shift was a good indicator for forgetting. However, these relationships deteriorated when a large context input was added.

4.4 Discussion

Using RNNs, we showed that learning a new task naturally causes activity changes that are observed when returning to a familiar task, recapitulating results from several experimental learning studies in monkeys[46,94,147]. Learning new tasks both caused uniform shifts in the preparatory activity[94] that were sustained when reverting to a familiar task, and left memory traces that made activity for the familiar task more appropriate for the new task during movement execution[147]. These changes appeared naturally, without the addition of an additional mechanism, and allowed information about both tasks to coexist in the same neural activity patterns, facilitating retention savings. However, when we added contextual inputs that separated the activity for the different tasks, both the retention of the uniform shift and the memory trace were reduced, showing that these signatures of motor memory may not be present even if savings are maintained.

Relationship to continual learning

Our results show that uniform shifts and memory traces indeed are signatures for continual learning, but have certain caveats. To measure the extent of continual learning, we have focused on the retention savings, the increase in performance immediately after re-exposure due to greater retention. While Sun et al[94] specifically correlated the magnitude of the uniform shift during relearning to learning rate savings, Losey et al[147] did not directly relate memory traces to any measures of savings. Here, we focused on retention savings since learning rate savings might be confounded by starting at different motor output errors at the beginning of re-exposure (Task B2) for our models trained with gradient descent. In our models, the magnitude of memory traces was closely correlated with retention savings following both within-manifold and outside-manifold perturbations (Figure 4.4A, Figure S4.4A, Figure S4.5D) for networks without contextual inputs, while the magnitude of uniform shifts was not (Figure 4.4D, Figure S4.5I). Rather, uniform shifts were indirectly related to retention savings: the relative retention of the uniform shifts was correlated with the amount of forgetting (Figure 4.5F), which affected the amount of savings for within-manifold perturbations (Figure 4.5C). Thus, our simulations and analyses show that while memory traces and uniform shifts could both be related to continual learning, they are uncorrelated with each other (Figure S4.6B).

Importantly, while these measures could be useful signatures for continual learning, they are not always indicative of savings, and they are not necessary for continual learning to occur. By including context signals, we pushed activity to occupy different parts of neural activity space, leading to a decrease in memory traces and a lack of retention of uniform shifts (Figure 4.6C,D). This suggests that if context inputs or other "top-down" inputs drive tasks to use very different activity patterns for producing motor output across different tasks—in the extreme case, pushing different tasks to be performed by non-overlapping subpopulations of neurons—, these signatures can disappear even if motor memories of new tasks are retained. Here, our RNN specifically modeled neural activity in motor cortex which is directly used for BCI control, so such inputs may originate from activity

upstream of motor cortex. However, motor learning and adaptation involve contributions from many regions other than motor cortex, including parietal cortex[170,171], cerebellum[4–6], and basal ganglia[8,174]. While learning was constrained to our RNN model of motor cortex for the majority of this work, we also examined results when we added a second upstream RNN module and constrained learning to happen in this module. Even when learning was restricted to this upstream module, memory traces and sustained uniform shifts still arose (Figure S4.11), suggesting that the presence of these phenomena does not necessitate learning in the area that produces the behavioral output, in this case the motor cortex.

Effects of context on within-manifold and outside-manifold perturbation learning

By modeling the BCI paradigm from Sadtler et al[46] with many different within- and outside-manifold perturbations, we were able to further examine the differences between these perturbations. Previous work has largely distinguished the two classes of perturbations based on factors such as learning timescales[46], neural changes required for learning[16,17], reliance on correct feedback[142], and modeled synaptic weight changes[200]. Here, we showed that there was variability in the memory traces and uniform shifts within each class of perturbations, highlighting differences between different perturbations in each class. While outside-manifold perturbations generally led to larger memory traces, some still overlapped with the distribution of within-manifold perturbations, suggesting that the changes that within- and outside-manifold perturbations impose on neural activity may live on a continuum, rather than fall within two discrete categories. Moreover, adding and increasing the magnitude of a context signal led to less forgetting for within-manifold perturbations (Figure 4.6E), causing them to adopt similar results as outside-manifold perturbations (Figure 4.6D,E,F).

These unexpected similarities between within- and outside-manifold perturbations in the presence of large context cues were likely caused by these cues pushing supposedly within-manifold perturbations to be outside the estimated manifold, as the original manifold

estimated during baseline learning did not capture activity patterns imposed by different contexts. This is supported by large decreases in the activity's overlap with the manifold when the context switched at the beginning of Task B, even before learning occurred (Figure S4.10). Previous work has shown that neural manifolds are generally nonlinear[139], so large upstream inputs like context signals may drive activity into a part of activity space that is not captured by the same flat manifold estimated using a linear dimensionality reduction method—as we have done here and in the original experimental work. Thus, depending on the magnitude of these signals, we may explore activity that is more or less within the estimated flat manifold, rather than distinctly fully within or outside the actual nonlinear manifold. Indeed, this continuum between within- and outside-manifold perturbations may explain why even in experimental settings, some within-manifold perturbations required similar amounts of learning as some outside-manifold perturbations[46]. As a result, a wide range of inputs and tasks must be considered to capture better estimates of the "true" underlying manifold, to more fully capture different neural constraints on learning.

Our context manipulations highlighted tradeoffs between interference and generalization within continual learning. Greater context signals were able to prevent overwriting by avoiding interference, particularly when learning a new within-manifold map. However, by separating the activity for different maps, context signals forced RNNs to create new activity patterns for a new within-manifold Map B instead of generalizing existing patterns that were already available for the within-manifold Map A, making learning Map B slower and more difficult (Figure 4.6E,F). While our implementation of context forced activity to be in very different spaces, minimizing interference but abolishing generalization, actual neural activity in different real-world contexts may be organized so as to find a compromise between interference and generalization. For example, a comparison of different but related upper limb tasks indicated that their underlying patterns were shifted with respect to each other but these tasks still shared dynamical and geometric features[44]. How the brain accomplishes such a compromise with different contexts to promote continual learning would be an interesting avenue for future work.

Neural mechanisms for uniform shifts and memory traces

While uniform shifts and memory traces were consistently present in our RNNs without additional context inputs, we did not explicitly model any mechanisms to maintain information about different tasks in the same activity. Simply through sequential learning through gradient descent, the RNNs were pushed towards solutions that accommodated information about both tasks, rather than reverting to the original state before the new task was learned. Without explicit mechanisms, the magnitude of uniform shifts and memory traces observed may serve as a baseline minimum. While we have shown that this baseline is affected by many factors like the initial behavioral error, the brain may implement additional mechanisms that might organize these neural activity changes in a way that further fosters continual learning.

Within the machine learning literature, many neuroscience-inspired methods exist that try to encourage continual learning without catastrophic forgetting of previous tasks, including methods that regulate levels of synaptic plasticity to protect existing knowledge[201,202], push activity changes to be in orthogonal subspaces in the population space[203,204], incorporate context to separate activity[205,206], and use complementary learning systems that promote replay and consolidation[207]. By implementing different methods, contrasting their effects on uniform shifts and memory traces, and comparing them to experimental results, we may be able to pinpoint specific ways by which the brain may accommodate different neural activity patterns underpinning different behaviors. Since these motor memory signatures are directly experimentally measurable, they have the potential to serve as markers, enabling us to disentangle and distinguish between different types of learning.

4.5 Conclusion

Animals build an increasingly rich behavioral repertoire by acquiring different skills during development and a lifetime of learning. Yet, how the brain accommodates novel activity patterns necessary for the generation of these new behaviors remains an open question.

Recent experimental work in monkeys has proposed two potential motor cortical signatures of motor memories underpinning learning. Here, we leveraged the full access afforded by RNN models to simulate this continual learning process and systematically examine how properties of these neural signatures relate to the formation of a motor memory. We found that these signatures emerge naturally as a byproduct of sequential learning, and that their relationship with motor memories is complex: their correlation depends on the details of the perturbation that causes learning and even on whether this perturbation is explicitly cued, as expected from recent behavioral work in humans. Thus, our work identifies fundamental challenges to define a clear relationship between neural activity changes and the formation of motor memories, but also provides insights and metrics to interpret future experimental results.

4.6 Methods

Tasks

We trained RNNs to perform a standard delayed center-out reach task, which is commonly used in experimental settings to examine motor control. In experiments, the subject would control a cursor on a computer screen. The cursor starts in the center of a circle, and the subject has to reach one of eight possible targets evenly spaced around the circle. The subject is shown which target to move to with a target cue, but they have to delay movement until a later go cue –a design typically used to study motor planning. In our case, each trial lasted 2.0 s: the target cue was presented from the beginning of the trial, the go cue was presented at 0.2 s, and each reach lasted 1.5 s. We trained RNNs to produce velocities for different reaches, and the target velocities were defined by Gaussian profiles for each timestep t from the go cue, where each timestep equals 10 ms for a total of $T = 150$ time steps for each reach:

$$exp(\frac{(t - T/2)^2}{-2(T/8)^2}).$$

The target velocities were 0 before the go cue and after the end of the reach.

To model how experimental subjects already know how to perform different reaches prior to experiments, we first pre-trained the RNNs to produce different reaches with targets randomly distributed around a circle of 9.4 cm during a baseline training phase. Then, to model the experimental task, we trained RNNs to perform reaches to eight targets evenly spaced around the circle using a BCI map. The RNNs were used to model motor cortical neural activity, and the BCI map transformed unit firing rates to x and y output velocities. To examine sequential learning, we trained RNNs to use different BCI maps in blocks: first with Map A (Task A1), then with a new Map B (Task B), then back to the familiar Map A (Task A2).

BCI maps

Map A consisted of two transformations. First, a projection matrix $\mathbf{C}$ was used to project the firing rates from all $n = 400$ units onto an 8D space, or manifold, that captured the majority of the variance in the activity. This "intuitive manifold" was determined by performing principal component analysis (PCA) on the firing rates at the end of baseline training concatenated over trials. PCA finds n orthogonal basis vectors (principal components or PCs) that maximally capture the variance in the population activity, sorted by their corresponding amount of variance explained. We kept the eight leading PCs, which captured more than 80% of the variance in our network activity. Second, a decoder matrix $\mathbf{D}$ was used to project activity from the manifold to a two-dimensional space indicating the output velocities. The decoder was determined offline by performing linear regression of the produced velocities on the principal components. The full map was then given by:

$$\mathbf{M_A} = \mathbf{DC}.$$

Map B was either a within-manifold perturbation or an outside-manifold perturbation of Map A. For within-manifold perturbations, the transformation from the manifold to the output velocities was disrupted with a permutation matrix $\eta_{\mathbf{wm}}$ inserted between the first

and second linear transformations:

$$\mathbf{M_B} = \mathbf{D}\eta_{\mathbf{WM}}\mathbf{C}.$$

To examine the effects of perturbing different dimensions of the manifold, we examined within-manifold perturbations that permuted only the "high-variance" (first four) dimensions, only the "low-variance" (last four) dimensions, or all of the dimensions. In Figures 4.2-4.4, S4.1, S4.3, and S4.11, we examined the effects of 10 maps for random permutations of all dimensions. In Figure S4.2B-D and S4.5, we examined 24 maps each from all possible permutations $(4! = 24)$ of the "high-variance" and "low-variance" dimensions, and 10 maps for random permutations of all dimensions.

For outside-manifold perturbations, the transformation from the firing rates to the manifold was disrupted with a permutation matrix $\eta_{\mathbf{OM}}$ inserted before the first linear transformation:

$$\mathbf{M_B} = \mathbf{DC}\eta_{\mathbf{OM}}.$$

Note that η_{WM} is a 8x8 permutation matrix such that there are 8 factorial (40,320) possible η_{WM}s, while η_{OM} is a 400x400 permutation matrix such that there are 400 factorial (>10868) possibilities. To equalize the dimensions of the search space for within- and outside-manifold perturbations, we followed methods similar to Sadtler et al. 2014. In short, we separated the 400 units into 8 groups and only considered outside-manifold perturbations where we permuted these groups, keeping all units in the same group together. We then considered all within-manifold maps and this limited set of outside-manifold maps as candidate Map Bs. Since we saw that the memory trace was highly correlated with the Task A1→Map B output MSE (Figure 4.4B), in Figures 4.5, 4.6, S4.2E-G, and S4.6-S4.10, we controlled for this factor by calculating the Task A1→Map B output MSE for all candidate Map Bs and choosing the 10 within- and 10 outside-manifold maps with MSEs closest to the median of all values.

Neural network model

Network architecture

The model dynamics were given by:

$$\tau \dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} W_{ij}^{rec} r_j(t) + \sum_{k=1}^{I} W_{ik}^{in} s_k(t) + \sum_{l=1}^{F} W_{il}^{fb} f_l(t) + \eta_i(t),$$

where x_i was the hidden state of the ith unit and r_i was the corresponding firing rate following $tanh$ activation of x_i. The network had $N = 400$ units, I inputs, and $F = 2$ feedback signals. The time constant τ was set to 0.05 s, the integration time step dt was set to 0.01 s, and the noise η was randomly sampled from the Gaussian distribution $\mathcal{N}(0, 0.1)$ for each time step. The initial states $\mathbf{x}_{t=0}$ were sampled from the uniform distribution $\mathcal{U}(-0.1, 0.1)$. The network was fully recurrently connected, with the recurrent weights $\mathbf{W^{rec}}$ initially sampled from the Gaussian distribution $\mathcal{N}(0, \frac{g}{\sqrt{N}})$, where $g = 1.2$. The time-dependent stimulus inputs $\mathbf{s}$ (specified below) were fed into the network, with input weights $\mathbf{W^{in}}$. The delayed feedback $\mathbf{f}$ was equivalent to the two-dimensional output velocities $\mathbf{v}$, but delayed by 120 ms, with feedback weights $\mathbf{W^{fb}}$. The output v corresponded to x and y velocities of motor trajectories, and they were read out via the linear map:

$$v_i(t) = \sum_{j=1}^{N} W_{ij}^{out} r_j(t)$$

with the output weights $\mathbf{W^{out}}$. $\mathbf{W^{in}}$, $\mathbf{W^{fb}}$, and $\mathbf{W^{out}}$ all initially sampled from the uniform distribution $\mathcal{U}(-0.1, 0.1)$.

To model networks with and without context, two types of sustained inputs $\mathbf{s}$ were used. For the networks without context, $\mathbf{s}$ was four-dimensional and consisted of a one-dimensional hold signal, a two-dimensional target signal $(2\,cos\,\theta^{target}, 2\,sin\,\theta^{target})$ that specified the reaching direction θ^{target} of the target, and a one-dimensional length signal that specified the normalized length of the reach. All reaches were the same length so they were kept at

2, with the same magnitude as the target and hold signals. For the networks with context, $\mathbf{s}$ was six-dimensional, with the additional dimensions consisting of a two-dimensional one-hot encoded context signal. Context A (denoted by the vector $(2, 0)$) was used for baseline training, Task A1, and Task A2 while Context B (denoted by $(0, 2)$) was used for Task B1 and Task B2. To examine the effects of different context magnitudes, we multiplied these vectors by different factors: 0.5, 1, 2, 4, 6, 8, 10, 12, 14, and 16. For both types of inputs, the hold signal started at 2 and went to 0 at the go cue while the target, length, and context signals were presented throughout the trial. While all weights ($\mathbf{W^{in}}$, $\mathbf{W^{rec}}$, $\mathbf{W^{fb}}$, $\mathbf{W^{out}}$) were initialized at the beginning of baseline training, the input weights specific for Context B were also reinitialized at the beginning of Task A1 since the weights otherwise decreased dramatically towards 0 during baseline training when Context B was not in use.

To model upstream brain regions, we added an additional RNN module similar to Feulner et al[15]. This module received sustained inputs as described above, and was fully connected to the downstream module.

Model training

Networks were optimized to generate velocities of synthetic center-out reaches. To model baseline learning, we pretrained the networks on reaches with targets randomly distributed around a circle of 8 cm. During baseline training, networks had to learn all the weights ($\mathbf{W^{in}}$, $\mathbf{W^{rec}}$, $\mathbf{W^{fb}}$, $\mathbf{W^{out}}$) using the Adam optimizer with an initial learning rate $l = 10^{-4}$, first moment estimates decay rate $\beta_1 = 0.9$, second moment estimates decay rate $\beta_2 = 0.999$, and epsilon $\epsilon = 1e - 8$. Baseline training was implemented with 500 training trials and a batch size $B = 64$. Then, to model sequential learning, we trained the networks to control different BCI maps in sequential blocks: first with Map A (Task A1), then with a new Map B (Task B), then back to the familiar Map A (Task A2). During each block, networks had to learn $\mathbf{W^{in}}$, $\mathbf{W^{rec}}$, and $\mathbf{W^{fb}}$, but the output weights $\mathbf{W^{out}}$ were fixed as either Map A ($\mathbf{M_A}$) or Map B ($\mathbf{M_B}$). The weights were learned using stochastic gradient descent with a fixed learning rate $l = 5^{-3}$. We used a faster learning rate during sequential learning to

model faster short-term learning compared to long-term baseline learning. Each block was implemented with 300 training trials and a batch size $B = 64$. To examine the effects of learning different tasks, we trained networks separately on different Map Bs following Task A1 (see *Tasks* for delineation of different Map Bs used).

To examine the effects of using the intuitive manifold as the familiar map, we also trained networks on the same pairs of Map As and Map Bs, but we used the Map Bs as the familiar map instead; in other words, training proceeded from baseline training, to Map B, to a new Map A, then back to the familiar Map B. To examine the effects of simultaneous rather than sequential learning, we simultaneously trained networks on the same pairs of maps during baseline learning following network initialization. All training configurations were performed on eight different networks initialized from different random seeds.

The loss L was the mean-squared error between the two-dimensional output and target velocities over each time step t, with the total number of time steps $T = 200$:

$$L = \frac{1}{2BT} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{d=1}^{2} \left(v_d^{tar}(b,t) - v_d^{out}(b,t) \right)^2 .$$

To produce network activity that aligned more closely to experimentally measured neural activity[49,50], we added L2 regularization terms for the activity rates and network weights in the overall loss function LR used for optimization:

$$L_R = L + R_W + R_R,$$

where

$$R_R = \frac{\beta}{BTN} \sum_{b=1}^{B} \sum_{t=0}^{T} \sum_{n=1}^{N} r_n(b,t)^2$$

and

$$R_W = \alpha \left(||\mathbf{W^{in}}|| + ||\mathbf{W^{rec}}|| + ||\mathbf{W^{out}}|| \right),$$

where $\beta = 0.2$ and $\alpha = 0.001$. Note that the loss recorded in the main text was the loss L before regularization. We clipped the gradient norm at 0.2 before applying the optimization

step. Following training, we evaluated the networks at each block using testing data for 320 trials (40 per target).

Data analysis

Distance between population activity means

To assess how learning a new task changes activity for a familiar task, we examined how network population activity changed from Task A1 to Task A2. We obtained the population activity by projecting network activity from each of the three task blocks into the eight-dimensional manifold. To compare our results with experimental results, we followed the methods outlined for the experimental analyses in Losey et al. 2024[147]. In brief, we calculated the Mahalanobis distances between the population activity means for Task A1 and Task A2 for each target, using the covariance of Task A1 activity per target. As a control, we recalculated the Mahalanobis distances with shuffled task labels. To further demonstrate that activity is distinctly different between the tasks, we used linear discriminant analysis with five-fold cross-validation to classify the three task blocks using only the preparatory activity (first 200 ms of each trial).

Uniform shifts

To measure uniform shifts in the preparatory space, we followed the methods outlined in Sun et al.[94] In short, we examined population activity only in the preparatory period of each trial, which lasted 200 ms from trial start to the go cue. To estimate the manifold in the preparatory space, we applied PCA to the preparatory activity at the end of baseline training, then projected preparatory unit activity from all three task blocks into this manifold to get the preparatory population activity. We defined the uniform shift axis as the vector connecting the centroids of the time-averaged, trial-averaged population activity for all targets from Task A1 to Task B. We wanted to make sure that the uniform shift was not capturing changes due to behavioral output differences, so we performed targeted

dimensionality reduction (TDR) on the preparatory activity and RNN outputs from Task A1 to identify a neural subspace that was predictive of behavioral output (velocities) in the upcoming reach, and we orthogonalized the uniform shift axis against the TDR axes. For each target, we projected the time-averaged, trial-averaged population activity onto the orthogonalized uniform shift axis to quantify the uniform shift per target. We obtained the normalized shifts by normalizing by the length of the target-averaged uniform shift from Task A1 to Task B. Thus, the average normalized uniform shift for all targets from Task A1 to Task B would be 1, and uniform shifts are sustained if they remain positive and close to 1 from Task A1 to Task A2.

Memory traces

To measure how learning a new task can leave a memory trace in the activity of a familiar task, we examined how activity patterns for a familiar task may become more useful for a new task after learning the new task. Specifically, we examined how activity patterns in Task A2 may be more appropriate for Map B than the patterns of Task A1. We projected activity from Task A1 and Task A2 into Map B offline and quantified the memory trace m as the increase in performance, or decrease in MSE, from Task A1 to Task A2:

$$m = \frac{1}{2BT} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{d=1}^{2} \left(v_d^{tar}(b,t) - v_d^{A1\to B}(b,t)\right)^2 - \left(v_d^{tar}(b,t) - v_d^{A2\to B}(b,t)\right)^2,$$

where

$$\mathbf{v}^{A1\to B} = \mathbf{r}^{A1}\mathbf{M}_B^{\mathsf{T}} \qquad \text{and} \qquad \mathbf{v}^{A2\to B} = \mathbf{r}^{A2}\mathbf{M}_B^{\mathsf{T}}$$

are the output velocities when Task A1 and Task A2 activity are projected into Map B, respectively.

Task properties

To understand how properties regarding the relationship between tasks may affect the magnitude of the memory trace, we examined two measures that quantified the relationship between Map A and Map B: the principal angles and the variance accounted for (VAF) between the maps.

Principal angles measure the relative alignment between two m-dimensional subspaces in an n dimensional space ($m < n$) by quantifying the m angles between pairs of basis vectors in the subspaces that minimize the angles between them. We followed the methods from Bjorck and Golub[208] to calculate the principal angles: given the maps $\mathbf{M_A}$ and $\mathbf{M_B}$, we perform QR decomposition to obtain orthonormal bases $\mathbf{Q_A}$ and $\mathbf{Q_B}$. We then construct the inner product of these bases and perform singular value decomposition to obtain:

$$\mathbf{Q}_A^{\mathsf{T}}\mathbf{Q}_B = \mathbf{P}_A\Sigma\mathbf{P}_B^{\mathsf{T}},$$

where $\mathbf{P}_A$ and $\mathbf{P}_B$ specify the directions that minimize the principal angles and $\mathbf{\Sigma}$ is a diagonal matrix whose entries are the ranked cosines of the principal angles.

Variance accounted for (VAF) measures how much the variance associated with a given task can be accounted for when projected into the map for a different task. We computed the variance when Task A1 activity was projected into Map A and the variance when it was projected into the new bases for Map B found using principal angles:

$$\sigma^2_{A1\to A} = Var(\mathbf{r}^{A1}\mathbf{M}_A^{\mathsf{T}}) \qquad \sigma^2_{A1\to B} = Var(\mathbf{r}^{A1}\mathbf{M}_B^{\mathsf{T}}\mathbf{P}_B).$$

VAF was the ratio of these two variances:

$$\frac{\sigma^2_{A1\to B}}{\sigma^2_{A1\to A}}.$$

To understand how the magnitude of the memory traces relate to the initial behavioral

disruption imposed by a new task, we measured the initial error using the Task A1→Map B output MSE and the Task A1→Map B angular error. The Task A1→Map B output MSE has been described above, whereas the Task A1→Map B angular error was computed based on the difference between the angle to the target and the mean angle for the entire reach during movement.

To understand how learning a new task changed the activity with respect to the manifold, we calculated the manifold overlap[180] of Task B1 activity $\mathbf{r^{B1}}$ with the original intuitive manifold. We first calculated the covariance matrix $\mathbf{S}$ of $\mathbf{r^{B1}}$ and projected it into $\mathbf{C}$, the first eight principal components that describe the intuitive manifold. To quantify the variance explained by the projection, we divided the trace of the projection by the trace of the corresponding covariance matrix:

$$\frac{Tr(\mathbf{CSC}^{\mathsf{T}})}{Tr(\mathbf{S})}$$

Control analysis

To verify that memory traces of Map B arose specifically due to the process of learning Task B, rather than due to changes that may arise from other processes, we ran two controls. First, we reset the weights learned in Task B ($\mathbf{W^{in}}$, $\mathbf{W^{fb}}$, and $\mathbf{W^{rec}}$) to those at the end of Task A1 before Task A2 to check if memory traces may arise due to generally more learning of Map A. Second, we randomly redistributed the weight changes incurred during Task B before Task A2 to check if memory traces may arise due to random weight changes of the same magnitude that are unrelated to learning Map B.

Code availability

All analyses were implemented using custom python code (Python 3.8) and open-source software. All the figures are reproducible by running Jupyter notebooks. Code to reproduce all the results is openly available on GitHub at https://github.com/JoannaChang/motor_memory.

4.7 Supplementary Figures

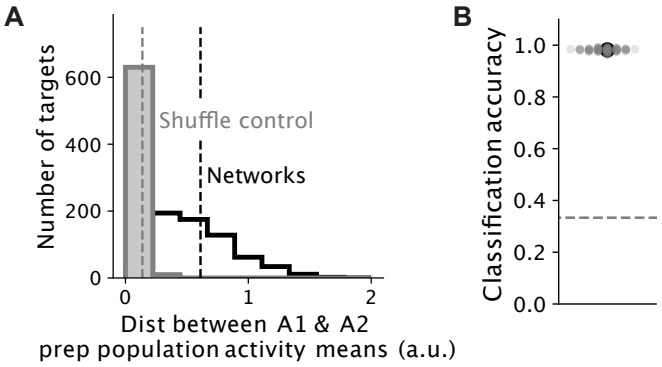


Figure S4.1: Neural activity changes from Task A1 to Task A2. A) Mahalanobis distance between population activity means for Task A1 and Task A2 for each target during preparation (first 200 ms of each trial, black), and for controls where the task labels were shuffled (gray). Dotted lines, means across 10 different maps for each random seed (n=8 random seeds). **B)** Classification accuracy for classifying tasks (Task A1, Task B1, Task B2) from neural activity during preparation using linear discriminant analysis. Large circle, mean across maps and random seeds; small circles, accuracy for each map for each random seed; colors, different random seeds; dotted line, chance accuracy level.

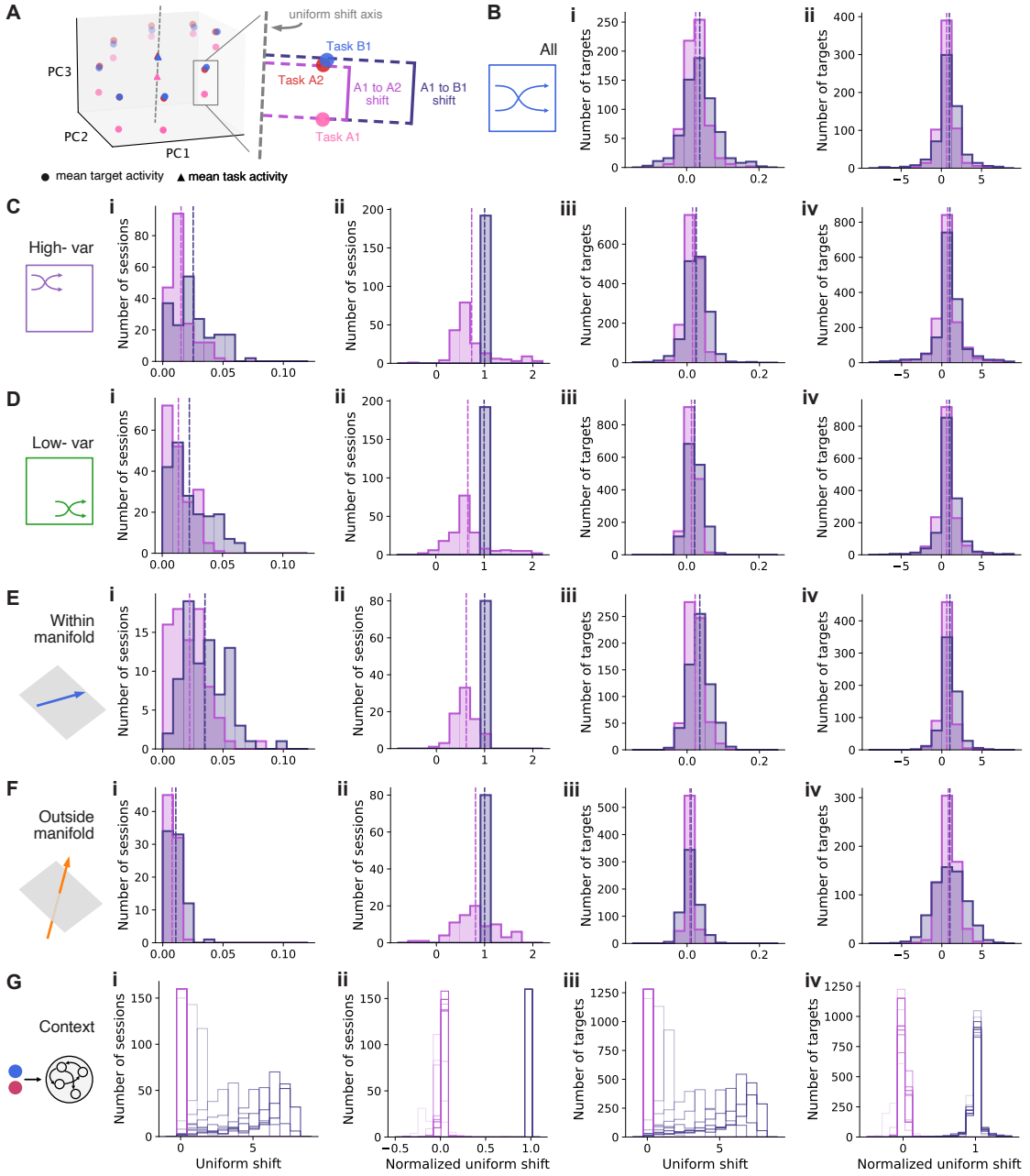


Figure S4.2: Uniform shifts for different simulations. A) Uniform shifts can be quantified per target for each session rather than averaged across all targets for each session as in Fig. 3B. To quantify uniform shifts per target, mean target activity (circles) was projected on the uniform shift axis determined based on mean task activity (triangles) for Task A1 and Task B1. The uniform shifts were then the resulting distances between these projected mean states. **B**) Uniform shifts for 10 different maps based on within-manifold permutations that shuffled all manifold dimensions (used in Figs 2-4). **B.i**) Uniform shifts from Task A1 to Task B1 and Task A1 to Task A2 for each target. Dotted lines, means across 10 different maps for each random seed. **B.ii**) Uniform shift per target, normalized by the uniform shift averaged across targets from Task A1 to Task B1 for each session. **C**) Uniform shifts for 24 different maps based on within-manifold perturbations that shuffled the first four high-variance dimensions (used in Fig 4). **C.i**) Uniform shift averaged

across targets (as in Fig. 3B). **C.ii**) Normalized uniform shift averaged across targets (as in Fig. 3C). **C.iii-iv**) Same as B.i-ii but for high-variance maps. **D**) Same as C but for 24 different maps based on within-manifold perturbations that shuffled the last four low-variance dimensions (used in Fig S6). **E**) Same as C but for 10 different within-manifold perturbations with the same initial output MSE (used in Fig 5). **F**) Same as C but for 10 different outside-manifold perturbations with the same initial output MSE (used in Fig 5). **G**) Same as C but for 10 outside-manifold and 10 within-manifold perturbations with the same initial output MSE for networks with different context magnitudes (used in Fig 6). Opacities, different context magnitudes (most transparent for smallest magnitude).

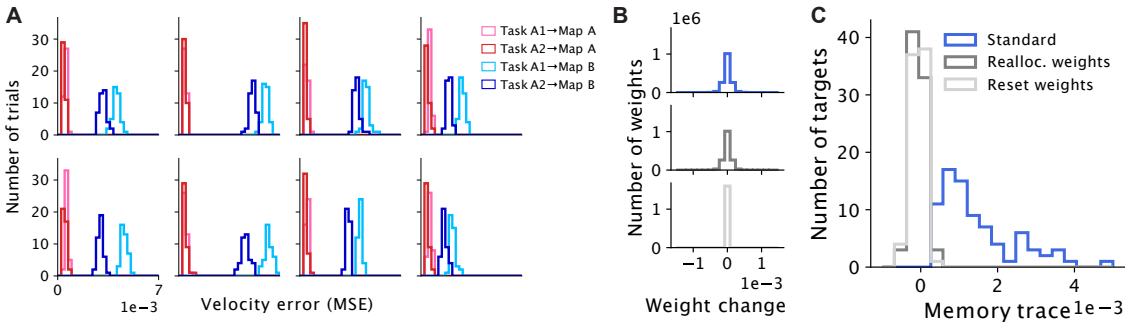


Figure S4.3: Memory traces consistently arise and are due to learning Map B. A) MSE between the produced and target output for Task A1 and Task A2 activity when projected into Map A and Map B for each target in an example session (same session as Fig 3E). **B-C**) To show that memory traces arise specifically due to changes incurred from learning Map B, networks were trained on Task A2 without altered weights (standard, blue), with weight changes from Task B randomly reallocated before Task A2 (dark gray), or with weights reset before Task A2 to those at the end of Task A1 (light gray). **B**) Weight changes from end of Task A1 to beginning of Task A2. **C**) Memory traces for each target for each session for 10 different maps for each random seed (n=8 random seeds) after the weight manipulations outlined in B.

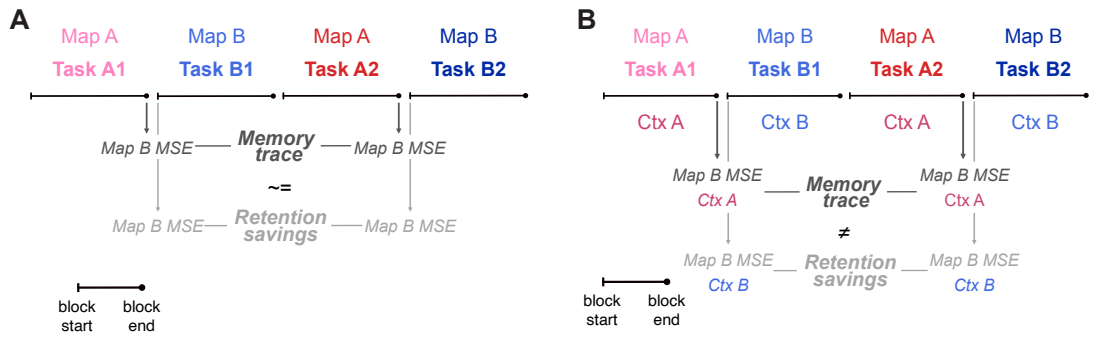


Figure S4.4: Relationship between memory trace and retention savings. A) Schematic for how memory trace and retention savings are measured based on MSEs at different stages of learning without context cues. Note that the memory trace is closely related to retention savings. The memory trace is defined as the increase in performance from Task A1→Map B to Task A2→Map B while retention savings is defined as the increase in performance from the beginning of Task B1→Map B to the beginning of Task B2→Map B. Without context cues, the network dynamics are identical at the end of Task A1 (Task A2) compared to the beginning of Task B1 (Task B2), other than minimal contributions from feedback, such that the memory trace is equivalent to retention savings, other than the contributions from feedback. **B)** Schematic for how memory trace and retention savings are measured based on MSEs at different stages of learning with context cues. Note that memory trace is not equivalent to retention savings when context cues are added because the network dynamics at the end of Task A1 (Task A2) are not identical to those at the beginning of Task B1 (Task B2).

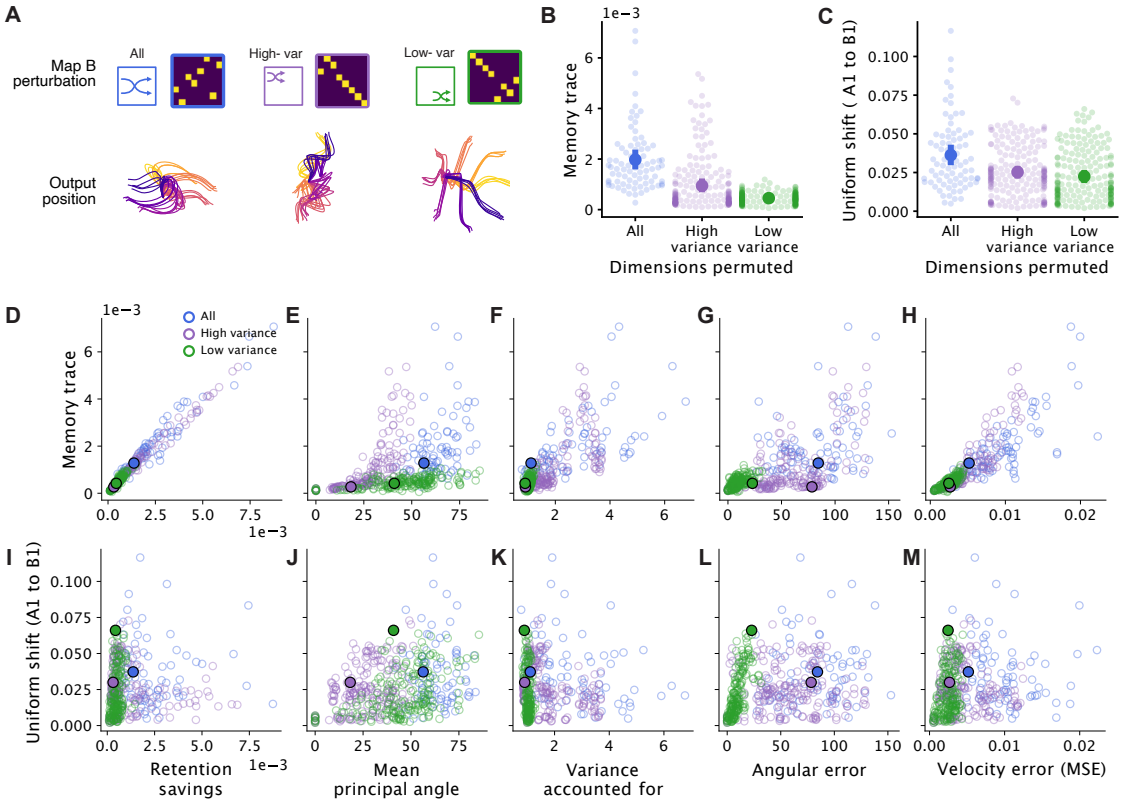


Figure S4.5: The magnitude of memory traces and uniform shifts compared to different measures. **A**) Networks were trained on different Map Bs representing within-manifold perturbations made from permutations in "all" manifold dimensions (blue, n=10 maps), only the first four "high-variance" dimensions (purple, n=24 maps), or only the last four "low-variance" dimensions (green, n=24 maps). Top: Permutation matrices for example maps for each category. Bottom: Output positions when Task A1 activity is projected into the new maps before learning in Task B1. **B**) Mean memory trace across targets. Big circle and error bars, means and 95% confidence intervals with bootstrapping across different maps for each random seed (n = 8 random seeds); small circles, different maps for different random seeds. **C**) Uniform shift from Task A1 to Task B1 across targets. **D-H**) Mean memory trace across targets compared to different measures: from left to right, retention savings, mean principal angle between Map A and Map B, variance accounted for (VAF) in Task A1 activity by Map B, Task A1→Map B angular error, and Task A1→Map B velocity MSE (Methods). Open circles, different maps for different random seeds; closed circles, examples for each permutation type shared across D-M. **I-M**) Uniform shift from Task A1 to Task B1 across targets compared to different measures.

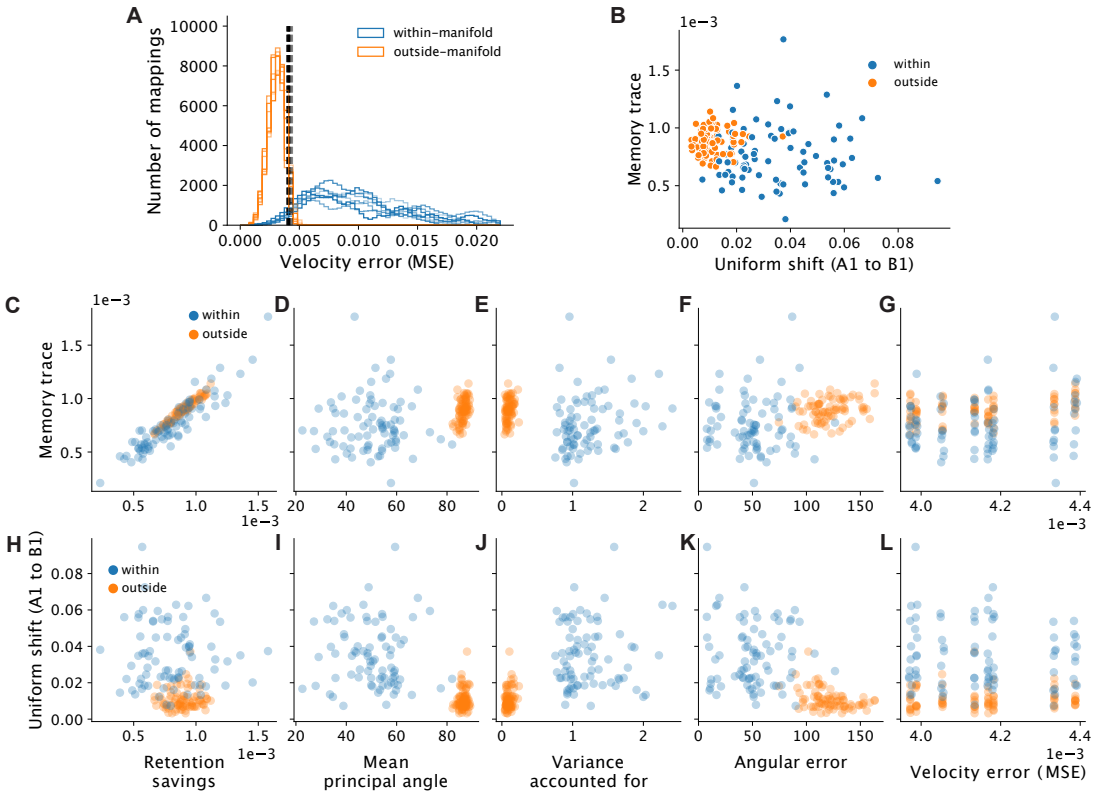


Figure S4.6: Relationship between memory trace and uniform shift and different factors once initial MSE is fixed. **A**) Task A1 activity was projected into thousands of candidate BCI maps representing within- and outside-manifold perturbations to obtain the initial Task A1→Map B MSE (Methods). Dashed lines, median for maps for each random seed. Opacities, different random seeds (n=8 random seeds). The 10 maps for each type of perturbation with MSEs closest to the median MSE across all maps for each seed were chosen for training. B) Uniform shift from Task A1 to Task B1 across targets compared to the mean memory trace across targets. Circles, mean for each map for each random seed. C-G) Mean memory trace across targets compared to the retention savings. **C-G**) Mean memory trace across targets compared to different measures: from left to right, retention savings, mean principal angle between Map A and Map B, variance accounted for (VAF) in Task A1 activity by Map B, Task A1→Map B angular error, and Task A1→Map B velocity MSE (Methods). Circles, different maps for different random seeds. **H-L**) Uniform shift from Task A1 to Task B1 across targets compared to different measures.

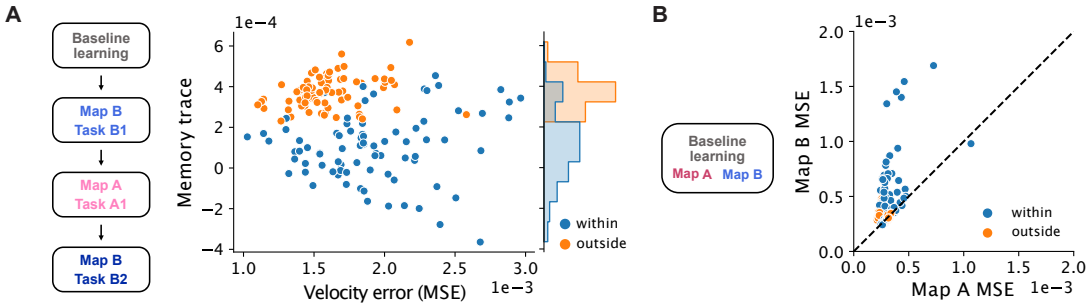


Figure S4.7: Learning within- and outside-manifold maps under different training regimes. A) Networks were trained with the same Map A and Map B pairings as in the main text (Fig 5), but the within- or outside perturbation (Map B) was used as the familiar map while the intuitive map (Map A) was used as the new map, such that learning proceeded from baseline learning, to Map B, to Map A, and back to Map B. Memory trace compared to Task A1→Map B MSE. Circles, mean for each map (n=10 maps per perturbation type) for each random seed (n=8 random seeds). **B)** Networks were trained with the same Map A and Map B pairings, but they were trained simultaneously during baseline learning after network initialization. Map A MSE compared to Map B MSE after baseline learning. Circles, mean for each map for each random seed; dashed line, identity line. Note that Map B had worse performance than Map A, especially for within-manifold perturbations.

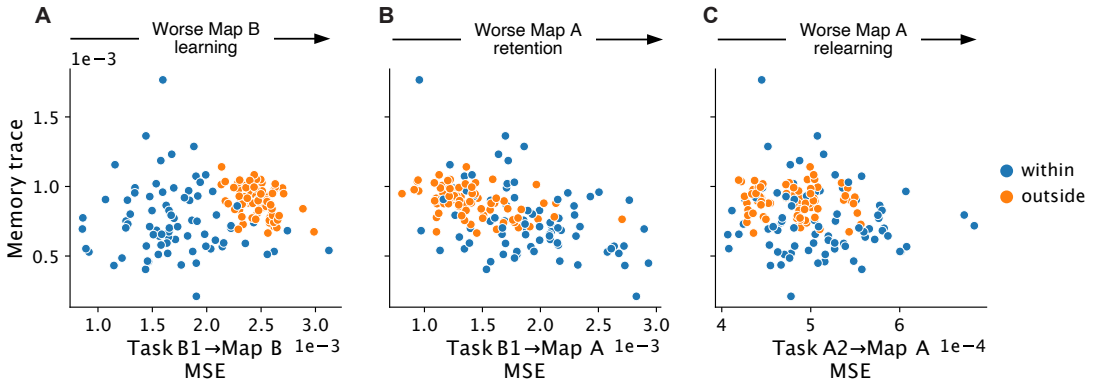


Figure S4.8: Larger memory traces for outside-manifold learning are not due to performance tradeoffs. A) Memory traces compared to Task B1→Map B MSE. Greater memory traces for outside-manifold perturbations were not due to better initial learning of Map B in Task B1. Circles, mean for each map (n=10 maps per perturbation type) for each random seed (n=8 random seeds). **B**) Memory traces compared to Task B1→Map A MSE. Greater memory traces for outside-manifold perturbations were not due to worse retention of Map A in Task B1. **C**) Memory traces compared to Task A2→Map A MSE. Greater memory traces for outside-manifold perturbations were not due to worse relearning of Map A in Task A2.

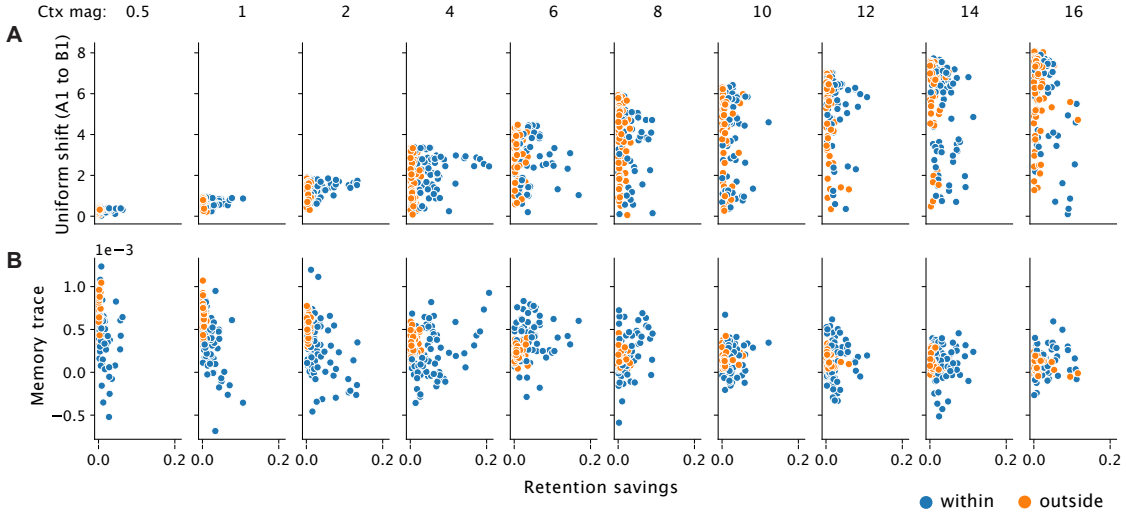


Figure S4.9: Context disrupts the correlation between memory traces and savings. A) Uniform shift from Task A1 to Task B1 across targets compared to mean retention savings for different context magnitudes. Circles, mean for each map (n=10 maps per perturbation type) for each random seed (n=8 random seeds). **B)** Mean memory trace across targets compared to mean retention savings for different context magnitudes.

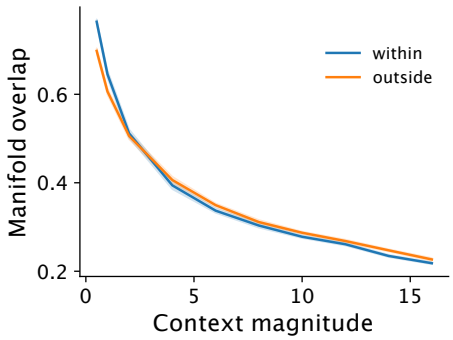


Figure S4.10: Manifold overlap for different context magnitudes. Manifold overlap between the activity after learning the new map in Task B1 and the original intuitive manifold for within- and outside-manifold perturbations. Line and shaded surfaces, mean and 95% confidence interval across 10 maps for each perturbation across networks of different seeds (n=8 random seeds).

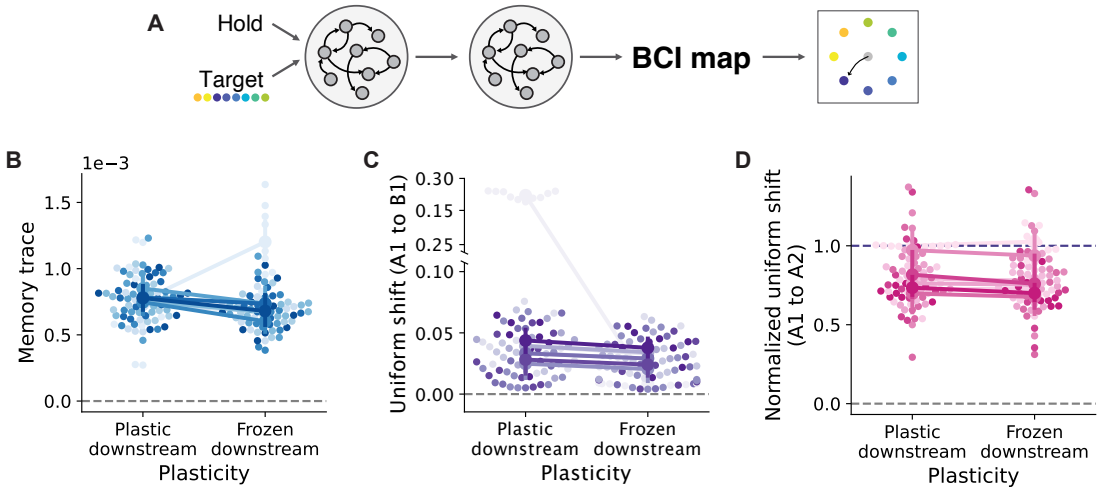


Figure S4.11: Memory traces and sustained uniform shifts arise when learning is confined to upstream of motor cortex. A) Schematic for network models with an additional RNN module added upstream. Activity from the downstream module is directly used to control the BCI, similar to how activity from the motor cortex is often used to control BCIs. The additional upstream module models brain regions upstream of the motor cortex. We examined how memory traces and uniform shifts differed when learning was restricted to the upstream module (frozen downstream weights) compared to when learning was allowed in the downstream module (plastic downstream weights). **B**) Mean memory trace across targets. Big circle and error bars, means and 95% confidence intervals with bootstrapping for 10 different maps for each random seed (n=8 random seeds). Small circles, different maps; colors, different random seeds. Note that memory traces were consistently greater than zero (grey dotted line), even when learning is restricted upstream. **C**) Uniform shift from Task A1 to Task B1 across targets. **D**) Uniform shift from Task A1 to Task A2 across targets, normalized to the uniform shift from Task A1 to Task B1. Note that the uniform shifts were sustained since they were consistently greater than zero (grey dotted line) and close to the shift from Task A1 to Task B (purple dotted line).

Chapter 5

Conclusion

> Don't adventures ever have an end? I suppose not. Someone else always has to carry on the story.
>
> *J.R.R. Tolkien*

5.1 Summary

This thesis explored how the brain implements stable yet flexible behavior through population latent dynamics. We first examined how the same behavior can be stably produced across different individuals of the same species. By combining experimental recordings with RNN modeling, we showed that population latent dynamics were preserved across individuals performing the same task. This preservation of latent dynamics was evident across multiple domains: in monkeys and mice, in simpler and more complex tasks, in covert and overt behavior, and in the motor cortex and striatum. The prevalence of these preserved dynamics, along with their ability to decode movement parameters, highlighted their importance in maintaining stable motor control and shaping behavior in general.

While the ability to reliably produce behavior is crucial for a species' survival, the ability to flexibly learn and adapt new skills is also integral for an individual's survival. We next

examined how *de novo* learning affects the flexibility for motor adaptation. Pretraining RNNs with various motor repertoires demonstrated that different learning experiences change the structure of the latent dynamics. Larger repertoires facilitated subsequent adaptation, but only when the structure of the inputs during learning, the structure of the latent dynamics, and the structure of the perturbation were all congruent. Thus, the flexibility of motor adaptation was determined by the structure of the latent dynamics created through *de novo* learning.

Flexible changes are necessary for motor learning, but animals need to accommodate these neural changes without losing stable control of familiar behaviors. We finally examined how learning a new task changes the latent dynamics used for a familiar task. By using RNNs to model a BCI, we found that two experimentally-observed signatures of motor memories—uniform shifts and memory traces—emerged naturally as a result of sequential learning. These signatures were both related to savings and indicated how information from different tasks could be incorporated within the same population activity. When we added contextual inputs to separate the activity of different tasks, these signatures disappeared even though savings remained. As a result, these signatures reflected flexible changes that preserved stable motor control, though effective dynamics could still emerge without them.

5.2 Limitations and future work

Balancing stable and flexible behavior

Our findings showed that the latent dynamics underlying motor control is highly stable, even proving to be preserved on an evolutionary scale across animals of the same species. However, these dynamics still need to be malleable to facilitate motor learning. Throughout an animal's lifetime, it can maintain stable motor control even while continuously acquiring new skills. How the brain achieves this lifelong continual learning is a key area of research, especially since artificial systems still struggle to reproduce this capability[21,22].

In this thesis, we demonstrated both the stability (Chapter 2) and the flexibility (Chapter 3) of latent dynamics, and sought to reconcile these dual aspects by examining potential signatures of motor memories that may allow flexible changes while preserving stable control (Chapter 4). These signatures were related to savings and continual learning, but they were not causative since they disappeared under different learning conditions. Thus, these signatures have limited use as indicators of continual learning, underscoring the need for further research into how the brain balances stability and flexibility for behavior.

As discussed in Chapter 4, there are many potential mechanisms of continual learning that we have not explored in this thesis. Some mechanisms were not explored due to the limited scope of the sequential BCI learning experiment that we model in Chapter 4. In the experiment, sequential learning was examined as adaptations to within- and outside-manifold perturbations, which require shorter time-scales than *de novo* learning, and only two different tasks were learned in sequence. Mechanisms such as complementary learning systems[207], which recruit other brain regions to promote memory consolidation, may be required to adjust population activity over longer timescales. Refinement can eventually result in stable patterns even if the initial dynamics look different. Furthermore, as the number of learned tasks increases, additional mechanisms may be required to integrate new tasks into a more complex neural manifold. In Chapter 3, we showed that neural networks could flexibly adapt to perturbations, but importantly, this flexibility was constrained by the structure of the neural space created by different motor repertoires. Also, while our artificial neural network models did not have any additional constraints apart from the tasks they needed to learn, this is not the case for real neural networks. The brain naturally faces many more biophysical constraints, which we hypothesized is responsible for preserved latent dynamics across animals (Chapter 2). Mechanisms that impose these constraints may ensure stability of behavior and limit the ways that dynamics can change to balance stability and flexibility in continual learning. As a result, understanding how latent dynamics support these mechanisms will be an important direction for future research.

The role of computational models

Throughout this thesis, we used computational models to investigate the neural underpinnings of motor control and motor learning. However, models are inherently simplified from the real-world system that that they are trying to model. In our case, our models have several limitations that constrain our findings. Although we mostly used single RNNs to model motor cortical populations, the motor cortex does not operate in isolation: it receives inputs from many other brain regions (Figure 1.1), and is involved in feedback loops through additional brain regions, muscles, and the external environment. While we modeled some of these interactions by incorporating external inputs and feedback from motor output, we did not account for all these contributions to motor control. Notably, we modeled learning as changes in synaptic weights within the recurrent network, but motor learning also involves other brain regions, particularly the cerebellum for adaptation[4–6]. Our models also assumed smooth, gradual learning via gradient descent, which does not capture discontinuous processes such as memory consolidation during sleep or rapid improvements through cognitive insight.

These simplifications limit the conclusions that we can draw for the role of neural manifolds in motor control and learning, since these manifolds also reflect constraints from multiple sources. Neural connectivity within a region imposes intrinsic constraints, but manifold structure is also shaped by external inputs, as shown in Chapter 3 and Chapter 4. Biomechanical constraints imposed by the physical body may also shape manifolds, but likely through historical co-evolution with the brain, rather than through immediate effects: neural circuits and the resulting dynamics can remain relatively similar even after substantial changes to the body like limb loss[209] or developmental changes[210]. As a result, we cannot fully understand how neural manifolds stay stable or change flexibly without modelling all of these factors.

Despite their intrinsic departure from biological realism, computational models can still provide insights in neuroscience that would be difficult to obtain otherwise. While the architecture of the RNN models throughout Chapters 2 to 4 were relatively similar, the

models were used in three distinct ways, each leveraging different strengths of computational modeling. In Chapter 2, we used RNN models to isolate the contribution of behavior to preserved population dynamics. By training multiple networks to produce the same outputs, we could determine whether behavioral similarity alone was sufficient to lead to the same dynamics. This use of models as controlled experiments allowed us to disentangle factors–neural dynamics and behavior—that are normally confounded in biological data. In Chapter 3, we took advantage of the ability to duplicate networks and subject them to different training sequences, allowing us to understand how different prior experiences affect adaptation to different perturbations. Experimentally, this would require tracking an animal's entire motor repertoire, which remains infeasible. Even if this was feasible, without the ability to rewrite an individual's history, the results would have to be aggregated from multiple animals undergoing different learning sequences. Instead, these models enabled rapid, tightly controlled comparisons across networks sharing identical initial conditions but distinct learning histories. In Chapter 4, we used RNNs to unify observations from different experimental setups in the same system, allowing us to synthesize findings about memory traces and uniform shifts that came from separate studies. Models can thus serve as a common ground for integrating results across laboratories and experimental conditions.

Our ability to fully manipulate these models allowed us to address questions that are inaccessible in biological systems, by isolating confounding factors, rewriting experimental steps, and reconciling observations across experimental contexts. Such models allow us to identify important variables and make predictions in a rapid and reproducible manner; however, their impact on understanding biological systems is maximized when they are closely developed alongside experimental work. Throughout this thesis, we aimed to connect computational results to experimental findings wherever possible, and observed that the models reproduced several features of the experimental data. Future progress will benefit from iterative computational modeling that is tightly integrated with experiments.

Preserved latent dynamics in other domains

In this thesis, we have established the prevalence of preserved latent dynamics and showed that analyzing population-level neural activity provides valuable insights into motor control and learning. While we showed that preserved latent dynamics exist across multiple domains, future work should explore the extent of preservation across a broader range of conditions to deepen our understanding.

In Chapter 2, we focused on latent dynamics in behavioral tasks. During overt behavior, animals generated similar motor outputs using similar latent dynamics. While the finding of preserved dynamics extended to covert behavior, it likely reflected similar underlying computations across animals. Given standardized experimental conditions and training procedures, animals probably learned to produce the same covert computations for these relatively straightforward behavioral tasks.

However, as shown in Chapter 3, different learning conditions can shape the structure of the latent dynamics, particularly through different inputs. In the motor cortex, inputs can either come from external cues via sensory pathways or from other brain regions as more processed, potentially cognitive, information. Investigating how distinct experimental learning paradigms change the latent dynamics would be interesting for future work, particularly for cognitive tasks that can be solved with various strategies. Different training processes may bias certain cognitive strategies, even if overt behavior is the same. By examining how well latent dynamics align across animals performing the same cognitive task, we may be able to distinguish between specific cognitive computations used by the brain.

Since latent dynamics reflect the brain's underlying computations, we can examine how these computations change by quantifying their alignment under different circumstances. First, we can align dynamics underlying the same behavior within the same animal. Observing how the dynamics shift as an animal acquires new skills can provide further insight into the flexibility and stability of behavior. While Chapters 3 and 4 explored this concept,

we purposely used simple behaviors to maximize interpretability and limited our analysis to sequential learning of two tasks. Future studies that incorporate more complex learning sequences and behavior would be valuable. Similarly, tracking how this alignment changes as an animal develops can reveal insights into how neural computations evolve through different life stages. Moving to comparisons across animals, future work can identify areas of misalignment across healthy animals and those with neurological disorders to improve our understanding of disease-related changes in the brain. Finally, expanding beyond individuals of the same species, we can align dynamics across individuals of closely related (e.g., mice and rats) or evolutionarily distant (e.g. mice and monkeys) species to explore how neural computations have adapted over evolutionary timescales.

Applications for BCIs

Our results hold implications for BCI development. Understanding when neural activity remains stable or changes over time is important for designing more robust and durable BCIs. Our finding that latent dynamics are preserved across animals suggest that we can create BCI decoders that generalize across subjects. Already, we showed that we could train decoders on the latent dynamics on one animal, and decode motor output with high accuracy using the aligned dynamics of another animal (Chapter 2). This suggests that we can train BCI decoders on one subject, and readily deploy them on another subject with minimal additional training. Such an approach has the potential to save resources, time, and effort, which are especially important for more vulnerable patient populations.

5.3 Conclusion

Through experimental recordings and neural network models, this thesis explored the stability and flexibility of population latent dynamics in motor control and motor learning. In three subsections, it examined 1) the stability of latent dynamics across animals during motor control, 2) the flexibility of latent dynamics for motor adaptation following motor

learning, and 3) the emergence of motor memories for flexible yet stable control during sequential learning. We believe that this work has established the importance of population latent dynamics in behavior and will motivate additional research into how they evolve and drive computations in the brain.

References

[1] Luppino, G. & Rizzolatti, G. The organization of the frontal motor cortex. *Physiology* **15**, 219–224 (2000).

[2] Shadmehr, R. & Krakauer, J. W. A computational neuroanatomy for motor control. *Experimental Brain Research* **185**, 359–381 (2008).

[3] Penfield, Wilder. & Boldrey, Edwin. Somatic motor and sensory representation i nthe cerebral cortex of man as studied by electrical stimulation. *Brain* **60**, 389–443 (1937).

[4] Galea, J. M., Vazquez, A., Pasricha, N., Orban de Xivry, J.-J. & Celnik, P. Dissociating the roles of the cerebellum and motor cortex during adaptive learning: the motor cortex retains what the cerebellum learns. *Cerebral Cortex* **21**, 1761–1770 (2011).

[5] Taylor, J. A., Klemfuss, N. M. & Ivry, R. B. An explicit strategy prevails when the cerebellum fails to compute movement errors. *Cerebellum* **9**, 580–586 (2010).

[6] Rabe, K. *et al.* Adaptation to visuomotor rotation and force field perturbation is correlated to different brain areas in patients with cerebellar degeneration. *Journal of Neurophysiology* **101**, 1961–1971 (2009).

[7] Mazzoni, P., Hristova, A. & Krakauer, J. W. Why don't we move faster? parkinson's disease, movement rigor, and implicit motivation. *Journal of Neuroscience* **27**, 7105–7116 (2007).

[8] Dudman, J. T. & Krakauer, J. W. The basal ganglia: from motor commands to the control of vigor. *Current Opinion in Neurobiology* **37**, 158–166 (2016).

[9] Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. & Haith, A. M. Motor learning. *Comprehensive Physiology* **9**, 613–663 (2019).

[10] Kleim, J. A. *et al.* Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning. *Journal of Neuroscience* **24**, 628–633 (2004).

[11] Xu, T. *et al.* Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* **462**, 915–919 (2009).

[12] Rioult-Pedotti, M.-S., Friedman, D., Hess, G. & Donoghue, J. P. Strengthening of horizontal cortical connections following skill learning. *Nature Neuroscience* **1**, 230–234 (1998).

[13] Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).

[14] Perich, M. G., Gallego, J. A. & Miller, L. E. A neural population mechanism for rapid learning. *Neuron* **100**, 964–976.e7 (2018).

[15] Feulner, B. *et al.* Small, correlated changes in synaptic connectivity may facilitate rapid motor learning. *Nature Communications* **13**, 5163 (2022).

[16] Oby, E. R. *et al.* New neural activity patterns emerge with long-term learning. *Proceedings of the National Academy of Sciences* **116**, 15210–15215 (2019).

[17] Golub, M. D. *et al.* Learning by neural reassociation. *Nature Neuroscience* **21**, 607–616 (2018).

[18] Yang, C. S., Cowan, N. J. & Haith, A. M. De novo learning versus adaptation of continuous control in a manual tracking task. *eLife* **10**, e62578 (2021).

[19] Haith, A. M., Yang, C. S., Pakpoor, J. & Kita, K. De novo motor learning of a bimanual control task over multiple days of practice. *Journal of Neurophysiology* **128**, 982–993 (2022).

[20] Gabriel, G. A., Mushtaq, F. & Morehead, J. R. De novo sensorimotor skill learning through reuse of movement components (2024).

[21] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks* **113**, 54–71 (2019).

[22] Mundt, M., Hong, Y., Pliushch, I. & Ramesh, V. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *Neural Networks* **160**, 306–336 (2023).

[23] Thach, W. T. Correlation of neural discharge with pattern and force of muscular activity, joint position, and direction of intended next movement in motor cortex and cerebellum. *Journal of Neurophysiology* **41**, 654–676 (1978).

[24] Georgopoulos, A. P., Kalaska, J. F., Caminiti, R. & Massey, J. T. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience* **2**, 1527–1537 (1982).

[25] Morrow, M. M., Jordan, L. R. & Miller, L. E. Direct comparison of the task-dependent discharge of M1 in hand space and muscle space. *Journal of Neurophysiology* **97**, 1786–1798 (2007).

[26] Gibson, A. R., Houk, J. C. & Kohlerman, N. J. Relation between red nucleus discharge and movement parameters in trained macaque monkeys. *The Journal of Physiology* **358**, 551–570 (1985).

[27] Evarts, E. V. Relation of pyramidal tract activity to force exerted during voluntary movement. *Journal of Neurophysiology* **31**, 14–27 (1968).

[28] Fetz, E. E. Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and Brain Sciences* **15**, 679–690 (1992).

[29] Scott, S. H. Inconvenient truths about neural processing in primary motor cortex. *The Journal of Physiology* **586**, 1217–1224 (2008).

[30] Oby, E. R., Ethier, C. & Miller, L. E. Movement representation in the primary motor cortex and its contribution to generalizable EMG predictions. *Journal of Neurophysiology* **109**, 666–678 (2013).

[31] Churchland, M. M. & Shenoy, K. V. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of Neurophysiology* **97**, 4235–4257 (2007).

[32] Stevenson, I. H. & Kording, K. P. How advances in neural recording affect data analysis. *Nature Neuroscience* **14**, 139–142 (2011).

[33] Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* **17**, 1500–1509 (2014).

[34] Williamson, R. C., Doiron, B., Smith, M. A. & Yu, B. M. Bridging large-scale neuronal recordings and large-scale network models using dimensionality reduction. *Current Opinion in Neurobiology* **55**, 40–47 (2019).

[35] Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* **17**, 440–448 (2014).

[36] Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical preparatory activity: Representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010).

[37] Vyas, S. *et al.* Neural population dynamics underlying motor learning transfer. *Neuron* **97**, 1177–1186.e3 (2018).

[38] Sauerbrei, B. A. *et al.* Cortical pattern generation during dexterous movement is input-driven. *Nature* **577**, 386–391 (2020).

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

[40] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).

[41] Nogueira, R., Rodgers, C. C., Bruno, R. M. & Fusi, S. The geometry of cortical representations of touch in rodents. *Nature Neuroscience* **26**, 239–250 (2023).

[42] Smoulder, A. L. *et al.* A neural basis of choking under pressure. *Neuron* **112**, 3424–3433.e8 (2024).

[43] Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience* **23**, 260–270 (2020).

[44] Gallego, J. A. *et al.* Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature Communications* **9**, 4233 (2018).

[45] Miri, A. *et al.* Behaviorally selective engagement of short-latency effector pathways by motor cortex. *Neuron* **95**, 683–696.e11 (2017).

[46] Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

[47] Dadarlat, M. C., Canfield, R. A. & Orsborn, A. L. Neural plasticity in sensorimotor brain–machine interfaces. *Annual Review of Biomedical Engineering* **25**, 51–76 (2023).

[48] Chapin, J. K., Moxon, K. A., Markowitz, R. S. & Nicolelis, M. A. L. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience* **2**, 664–670 (1999).

[49] Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience* **18**, 1025–1033 (2015).

[50] Hennequin, G., Vogels, T. P. & Gerstner, W. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron* **82**, 1394–1406 (2014).

[51] Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A. & Scherberger, H. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the National Academy of Sciences* **117**, 32124–32135 (2020).

[52] Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

[53] O'Shea, D. J. *et al.* Direct neural perturbations reveal a dynamical mechanism for robust computation. URL `https://www.biorxiv.org/content/10.1101/2022.12.16.520768v1`. Pages: 2022.12.16.520768 Section: New Results.

[54] Safaie, M. *et al.* Preserved neural dynamics across animals performing similar behaviour. *Nature* **623**, 765–771 (2023).

[55] Harris, K. D. & Shepherd, G. M. G. The neocortical circuit: themes and variations. *Nature Neuroscience* **18**, 170–181 (2015).

[56] Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications* **10**, 3770 (2019).

[57] Hiesinger, P. R. *The Self-Assembling Brain* (Princeton University Press, 2021).

[58] Mitchell, K. *Innate* (Princeton University Press, 2018).

[59] Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

[60] Marshel, J. H. *et al.* Cortical layer–specific critical dynamics triggering perception. *Science* eaaw5202 (2019).

[61] Okun, M. *et al.* Diverse coupling of neurons to populations in sensory cortex. *Nature* **521**, 511–515 (2015).

[62] Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annual Review of Neuroscience* **43**, 249–275 (2020).

[63] Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nature Reviews Neuroscience* 1–13 (2021).

[64] Santhanam, G. *et al.* Factor-analysis methods for higher-performance neural prostheses. *Journal of Neurophysiology* **102**, 1315–1330 (2009).

[65] Grillner, S. Evolution of the vertebrate motor system — from forebrain to spinal cord. *Current Opinion in Neurobiology* **71**, 11–18 (2021).

[66] Cisek, P. Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics* **81**, 2265–2287 (2019).

[67] Tuttle, A. H., Philip, V. M., Chesler, E. J. & Mogil, J. S. Comparing phenotypic variation between inbred and outbred mice. *Nature Methods* **15**, 994–996 (2018).

[68] Keller, D., Erö, C. & Markram, H. Cell densities in the mouse brain: A systematic review. *Frontiers in Neuroanatomy* **12**, 83 (2018).

[69] Häusser, M., Spruston, N. & Stuart, G. J. Diversity and dynamics of dendritic signaling. *Science* **290**, 739–744 (2000).

[70] Nakanishi, S. Molecular diiversity of glutamate receptors and implications for brain function. *Science* (1992).

[71] Kutsuwada, T. *et al.* Molecular diversity of the NMDA receptor channel. *Nature* **358**, 36–41 (1992).

[72] Brennan, C. & Proekt, A. A quantitative model of conserved macroscopic dynamics predicts future motor commands. *eLife* **8**, e46814 (2019).

[73] Goaillard, J.-M., Taylor, A. L., Schulz, D. J. & Marder, E. Functional consequences of animal-to-animal variation in circuit parameters. *Nature Neuroscience* **12**, 1424–1430 (2009).

[74] Kennedy, A. *et al.* Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature* **586**, 730–734 (2020).

[75] Thura, D., Cabana, J.-F., Feghaly, A. & Cisek, P. Unified neural dynamics of decisions and actions in the cerebral cortex and basal ganglia. *bioRxiv* 2020.10.22.350280 (2020).

[76] Perich, M. G. & Rajan, K. Rethinking brain-wide interactions through multi-region 'network of networks' models. *Current Opinion in Neurobiology* **65**, 146–151 (2020).

[77] Jiang, X., Saggar, H., Ryu, S. I., Shenoy, K. V. & Kao, J. C. Structure in neural activity during obbserved and executed movements is shared at the neural population level, not in single neurons. *Cell Reports* **32**, 108006 (2020).

[78] Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience* **18**, 1025–1033 (2015).

[79] Gallego-Carracedo, C., Perich, M. G., Chowdhury, R. H., Miller, L. E. & Gallego, J. A. Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner (2021).

[80] Elsayed, G. F. & Cunningham, J. P. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience* **20**, 1310–1318 (2017).

[81] Glaser, J. I. *et al.* Machine learning for neural decoding. *eNeuro* (2020).

[82] Glaser, J. I., Perich, M. G., Ramkumar, P., Miller, L. E. & Kording, K. P. Population coding of conditional probability distributions in dorsal premotor cortex. *Nature Communications* **9**, 1788 (2018).

[83] Lawlor, P. N., Perich, M. G., Miller, L. E. & Kording, K. P. Linear-nonlinear-time-warp-poisson models of neural activity. *Journal of Computational Neuroscience* **45**, 173–191 (2018).

[84] Jurado-Parras, M.-T. *et al.* The dorsal striatum energizes motor routines. *Current Biology* **30**, 4362–4372.e6 (2020).

[85] Thura, D. & Cisek, Paul. The basal ganglia do not select reach targets but control the urgency commitment. *Neuron* **95**, 1160–1170.e5 (2017).

[86] Cruz, B. F. *et al.* Action suppression reveals opponent parallel control via striatal circuits. *Nature* **607**, 521–526 (2022). Number: 7919.

[87] Park, J. *et al.* Motor cortical output for skilled forelimb movement is selectively distributed across projection neuron classes. *Science Advances* **8**, eabj5167 (2022).

[88] Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience* **9**, 357–381 (1986).

[89] Pruszynski, J. A. *et al.* Primary motor cortex underlies multi-joint integration for fast feedback control. *Nature* **478**, 387–390 (2011).

[90] Perich, M. G. *et al.* Motor cortical dynamics are shaped by multiple distinct subspaces during naturalistic behavior. *bioRxiv* 2020.07.30.228767 (2020).

[91] Machens, C. K., Romo, R. & Brody, C. D. Functional, But Not Anatomical, Separation of "What" and "When" in Prefrontal Cortex. *Journal of Neuroscience* **30**, 350–360 (2010).

[92] Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

[93] Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience* **21**, 102–110 (2018).

[94] Sun, X. *et al.* Cortical preparatory activity indexes learned motor memories. *Nature* 1–6 (2022).

[95] Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: Permitting preparation without movement. *Nature Neuroscience* **17**, 440–448 (2014).

[96] Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical areas interact through a communication subspace. *Neuron* **102**, 249–259.e4 (2019).

[97] Chen, H.-T., Manning, J. R. & van der Meer, M. A. A. Between-subject prediction reveals a shared representational geometry in the rodent hippocampus. *Current Biology* (2021).

[98] Herrero-Vidal, P., Rinberg, D. & Savin, C. Across-animal odor decoding by probabilistic manifold alignment. *Advances in Neural Information Processing Systems* **34**, 20360–20372 (2021).

[99] Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* 1–9 (2022). Publisher: Nature Publishing Group.

[100] Gallego, J. A., Makin, T. R. & McDougle, S. D. Going beyond primary motor cortex to improve brain–computer interfaces. *Trends in Neurosciences* **45**, 176–183 (2022).

[101] Pandarinath, C. & Bensmaia, S. J. The science and engineering behind sensitized brain-controlled bionic hands. *Physiological Reviews* **102**, 551–604 (2022).

[102] Dyer, E. L. *et al.* A cryptography-based approach for movement decoding. *Nature Biomedical Engineering* **1**, 967–976 (2017).

[103] Barra, B. *et al.* Epidural electrical stimulation of the cervical dorsal roots restores voluntary upper limb control in paralyzed monkeys. *Nature Neuroscience* **25**, 924–934 (2022).

[104] Degenhart, A. D. *et al.* Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature Biomedical Engineering* 1–14 (2020).

[105] Wen, S. *et al.* Rapid adaptation of brain–computer interfaces to new neuronal ensembles or participants via generative modelling. *Nature Biomedical Engineering* 1–13 (2021).

[106] Jude, J., Perich, M. G., Miller, L. E. & Hennig, M. H. Robust alignment of cross-

session recordings of neural population activity by behaviour via unsupervised domain adaptation (2022). 2202.06159.

[107] Cowley, B. R. *et al.* Slow Drift of Neural Activity as a Signature of Impulsivity in Macaque Visual and Prefrontal Cortex. *Neuron* **108**, 551–567.e8 (2020).

[108] Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364**, eaav3932 (2019). Publisher: American Association for the Advancement of Science.

[109] Latimer, K. W. & Freedman, D. J. Low-dimensional encoding of decisions in parietal cortex reflects long-term training history. *Nature Communications* **14**, 1010 (2023). Number: 1 Publisher: Nature Publishing Group.

[110] Kleim, J. A. *et al.* Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning. *Journal of Neuroscience* **24**, 628–633 (2004).

[111] Xu, T. *et al.* Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* **462**, 915–919 (2009).

[112] Perich, M. G. & Miller, L. E. Altered tuning in primary motor cortex does not account for behavioral adaptation during force field learning. *Experimental Brain Research* **235**, 2689–2704 (2017).

[113] Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).

[114] Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods* **10**, 64–67 (2013). Number: 1 Publisher: Nature Publishing Group.

[115] Trautmann, E. M. *et al.* Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308.e4 (2019).

[116] Osborne, J. E. & Dudman, J. T. RIVETS: A mechanical system for in vivo and in vitro electrophysiology and imaging. *PLOS ONE* **9**, e89007 (2014).

[117] Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

[118] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

[119] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[120] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

[121] McKinney, W. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, 56 – 61 (2010).

[122] Bach, F. R. & Jordan, M. I. Kernel independent component analysis. *Journal of Machine Learning Research* **3**, 1–48 (2002).

[123] Paszke, A. *et al.* Automatic differentiation in pytorch. In *NIPS-W* (2017).

[124] Gower, J. C. Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975).

[125] Kobak, D. *et al.* Demixed principal component analysis of neural population data. *eLife* **5**, e10989 (2016).

[126] Michaels, J. A., Dann, B. & Scherberger, H. Neural population dynamics during reaching are better explained by a dynamical system than representational tuning. *PLOS Computational Biology* **12**, e1005175 (2016).

[127] Feulner, B. *et al.* Small, correlated changes in synaptic connectivity may facilitate rapid motor learning. *Nature communications* **13**, 5163 (2022).

[128] Chang, J. C., Perich, M. G., Miller, L. E., Gallego, J. A. & Clopath, C. De novo motor learning creates structure in neural activity that shapes adaptation. *Nature Communications* **15**, 4084 (2024).

[129] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2017).

[130] Krakauer, J. W., Pine, Z. M., Ghilardi, M. F. & Ghez, C. Learning of visuomotor transformations for vectorial planning of reaching trajectories. *Journal of Neuroscience* **20**, 8916–8924 (2000).

[131] Shadmehr, R. & Mussa-Ivaldi, F. A. Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* **14**, 3208–3224 (1994).

[132] Sohn, H., Meirhaeghe, N., Rajalingham, R. & Jazayeri, M. A network perspective on sensorimotor learning. *Trends in Neurosciences* **44**, 170–181 (2021).

[133] Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annual Review of Neuroscience* **43**, 249–275 (2020).

[134] Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nature Reviews Neuroscience* **22**, 359–371 (2021).

[135] Yu, B. M. *et al.* Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology* **102**, 614–635 (2009).

[136] Nieh, E. H. *et al.* Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).

[137] Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).

[138] Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience* **22**, 1512–1520 (2019).

[139] Fortunato, C. *et al.* Nonlinear manifolds underlie neural population activity during behaviour. *bioRxiv* (2024).

[140] Hermansen, E., Klindt, D. A. & Dunn, B. A. Uncovering 2-D toroidal representations in grid cell ensemble activity during 1-D behavior (2022).

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

[142] Feulner, B. & Clopath, C. Neural manifold under plasticity in a goal driven learning behaviour. *PLOS Computational Biology* **17**, e1008621 (2021).

[143] Skinner, B. F. *The behavior of organisms: an experimental analysis* (Appleton-Century, Oxford, England, 1938).

[144] Krueger, K. A. & Dayan, P. Flexible shaping: How learning in small steps helps. *Cognition* **110**, 380–394 (2009).

[145] Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. *ICML* 1–8 (2009).

[146] Vyas, S., O'Shea, D. J., Ryu, S. I. & Shenoy, K. V. Causal role of motor preparation during error-driven learning. *Neuron* **106**, 329–339.e4 (2020).

[147] Losey, D. M. *et al.* Learning leaves a memory trace in motor cortex. *Current Biology* **34**, 1519–1531.e4 (2024).

[148] Logiaco, L. & Escola, G. S. Thalamocortical motor circuit insights for more robust hierarchical control of complex sequences. *bioRxiv* (2020).

[149] Bond, K. M. & Taylor, J. A. Structural learning in a visuomotor adaptation task is explicitly accessible. *eNeuro* **4**, ENEURO.0122–17.2017 (2017).

[150] Day, K. A., Roemmich, R. T., Taylor, J. A. & Bastian, A. J. Visuomotor learning generalizes around the intended movement. *eNeuro* **3**, ENEURO.0005–16.2016 (2016).

[151] Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nature Reviews Neuroscience* **9**, 292–303 (2008).

[152] Marshall, N. J. *et al.* Flexible neural control of motor units. *Nature Neuroscience* **25**, 1492–1504 (2022).

[153] Schwartz, A. B. Direct cortical representation of drawing. *Science* **265**, 540–542 (1994).

[154] Schwartz, A. B. Motor cortical activity during drawing movements: population representation during sinusoid tracing. *Journal of Neurophysiology* **70**, 28–36 (1993).

[155] Braun, D. A., Aertsen, A., Wolpert, D. M. & Mehring, C. Motor task variation induces structural learning. *Current Biology* **19**, 352–357 (2009).

[156] Heald, J. B., Lengyel, M. & Wolpert, D. M. Contextual inference underlies the learning of sensorimotor repertoires. *Nature* **600**, 489–493 (2021).

[157] Krakauer, J. W., Mazzoni, P., Ghazizadeh, A., Ravindran, R. & Shadmehr, R. Generalization of motor learning depends on the history of prior action. *PLOS Biology* **4**, e316 (2006).

[158] Verstynen, T. & Sabes, P. N. How each movement changes the next: An experimental and theoretical study of fast adaptive priors in reaching. *Journal of Neuroscience* **31**, 10050–10059 (2011).

[159] Kalaska, J. F., Cohen, D. A., Hyde, M. L. & Prud'homme, M. A comparison of movement direction-related versus load direction- related activity in primate motor cortex, using a two-dimensional reaching task. *Journal of Neuroscience* **9**, 2080–2102 (1989).

[160] Churchland, M. M., Santhanam, G. & Shenoy, K. V. Preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach. *Journal of Neurophysiology* **96**, 3130–3146 (2006).

[161] Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).

[162] Humphreys, P. C., Daie, K., Svoboda, K., Botvinick, M. & Lillicrap, T. P. BCI learning phenomena can be explained by gradient-based optimization (2022).

[163] Rathelot, J.-A. & Strick, P. L. Subdivisions of primary motor cortex based on cortico-motoneuronal cells. *Proceedings of the National Academy of Sciences* **106**, 918–923 (2009).

[164] Perich, M. G. *et al.* Inferring brain-wide interactions using data-constrained recurrent neural network models (2021).

[165] Perich, M. G. & Rajan, K. Rethinking brain-wide interactions through multi-region "network of networks" models. *Current opinion in neurobiology* **65**, 146–151 (2020).

[166] Beiran, M., Dubreuil, A., Valente, A., Mastrogiuseppe, F. & Ostojic, S. Shaping dynamics with multiple populations in low-rank recurrent networks. *Neural Computation* **33**, 1572–1615 (2021).

[167] Xiao, J. Premotor neuronal plasticity in monkeys adapting to a new dynamic environment. *The European Journal of Neuroscience* **22**, 3266–3280 (2005).

[168] Reis, J. *et al.* Noninvasive cortical stimulation enhances motor skill acquisition over multiple days through an effect on consolidation. *Proceedings of the National Academy of Sciences* **106**, 1590–1595 (2009).

[169] Gentner, R. *et al.* Encoding of motor skill in the corticomuscular system of musicians. *Current biology: CB* **20**, 1869–1874 (2010).

[170] Gréa, H. *et al.* A lesion of the posterior parietal cortex disrupts on-line adjustments during aiming movements. *Neuropsychologia* **40**, 2471–2480 (2002).

[171] Mutha, P. K., Sainburg, R. L. & Haaland, K. Y. Left parietal regions are critical for adaptive visuomotor control. *Journal of Neuroscience* **31**, 6972–6981 (2011).

[172] Park, J. *et al.* Conjoint specification of action by neocortex and striatum. *Neuron* **113**, 620–636.e6 (2025).

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

[174] Wolff, S. B. E., Ko, R. & Ölveczky, B. P. Distinct roles for motor cortical and thalamic inputs to striatum during motor skill learning and execution. *Science Advances* **8**, eabk0231 (2022).

[175] Driscoll, L., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs (2022).

[176] Codol, O., Michaels, J. A., Kashefi, M., Pruszynski, J. A. & Gribble, P. L. MotorNet: a Python toolbox for controlling differentiable biomechanical effectors with artificial neural networks. *eLife* **12** (2024).

[177] Delp, S. L. *et al.* OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering* **54**, 1940–1950 (2007).

[178] Todorov, E., Erez, T. & Tassa, Y. MuJoCo: A physics engine for model-based control. *IEEE* 5026–5033 (2012).

[179] Russo, A. A. *et al.* Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron* **107**, 745–758.e6 (2020).

[180] Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M. & Cunningham, J. P. Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications* **7**, 13239 (2016).

[181] Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2** (2008).

[182] Gallego-Carracedo, C., Perich, M. G., Chowdhury, R. H., Miller, L. E. & Gallego, J. Á. Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner. *eLife* **11**, e73155 (2022).

[183] Russo, A. A. *et al.* Motor cortex embeds muscle-like commands in an untangled population response. *Neuron* **97**, 953–966.e8 (2018).

[184] Chang, J. C., Clopath, C. & Gallego, J. A. Neural signatures of motor memories emerge in neural network models. *bioRxiv* (2025).

[185] Wise, S. P. & Murray, E. A. Role of the hippocampal system in conditional motor learning: mapping antecedents to action. *Hippocampus* **9**, 101–117 (1999).

[186] Hikosaka, O., Nakamura, K., Sakai, K. & Nakahara, H. Central mechanisms of motor skill learning. *Current opinion in neurobiology* **12**, 217–222 (2002).

[187] Morton, S. M. & Bastian, A. J. Cerebellar contributions to locomotor adaptations during splitbelt treadmill walking. *Journal of Neuroscience* **26**, 9107–9116 (2006).

[188] Imamizu, H. *et al.* Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* **403**, 192–195 (2000).

[189] Komiyama, T. *et al.* Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature* **464**, 1182–1186 (2010).

[190] Shadmehr, R. & Holcomb, H. H. Neural correlates of motor memory consolidation. *Science* **277**, 821–825 (1997).

[191] Park, S.-W., Dijkstra, T. & Sternad, D. Learning to never forget—time scales and specificity of long-term memory of a motor skill. *Frontiers in Computational Neuroscience* **7**, 111 (2013).

[192] Miyachi, S., Hikosaka, O. & Lu, X. Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Experimental Brain Research* **146**, 122–126 (2002).

[193] Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annual Review of Neuroscience* **36**, 337–359 (2013).

[194] Feulner, B., Perich, M. G., Miller, L. E., Clopath, C. & Gallego, J. A. A neural implementation model of feedback-based motor learning. *Nature Communications* **16**, 1805 (2025).

[195] Gurnani, H., Liu, W. & Brunton, B. W. Feedback control of recurrent dynamics constrains learning timescales during motor adaptation (2024).

[196] Ogasa, K. *et al.* Decision uncertainty as a context for motor memory. *Nature Human Behaviour* 1–14 (2024).

[197] Hennig, J. A. *et al.* How learning unfolds in the brain: toward an optimization view. *Neuron* (2021).

[198] Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences* **115**, E10313–E10322 (2018).

[199] Shea, J. B. & Morgan, R. L. Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory* **5**, 179–187 (1979).

[200] Wärnberg, E. & Kumar, A. Perturbing low dimensional activity manifolds in spiking neuronal networks. *PLOS Computational Biology* **15**, e1007074 (2019).

[201] Zhang, T. *et al.* A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost. *Science Advances* **9**, eadi2947 (2023).

[202] Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**, 3521–3526 (2017).

[203] Duncker, L., Driscoll, L. N., Shenoy, K. V., Sahani, M. & Sussillo, D. Organizing recurrent network dynamics by task-computation to enable continual learning. *NeurIPS* 11 (2020).

[204] Tsuda, B., Tye, K. M., Siegelmann, H. T. & Sejnowski, T. J. A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proceedings of the National Academy of Sciences* **117**, 29872–29882 (2020).

[205] Zeng, G., Chen, Y., Cui, B. & Yu, S. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence* **1**, 364–372 (2019).

[206] Masse, N. Y., Grant, G. D. & Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences* **115**, E10467–E10475 (2018).

[207] Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences* **20**, 512–534 (2016).

[208] Björck, Å. & Golub, G. H. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **27**, 579–594 (1973).

[209] Schone, H. R. *et al.* Stable cortical body maps before and after arm amputation. *Nature Neuroscience* **28**, 2015–2021 (2025).

[210] Gomez-Marin, A. & Ghazanfar, A. A. The Life of Behavior. *Neuron* **104**, 25–36 (2019).