Neural Dynamics and the Geometry of Population Activity

Abigail A. Russo

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

Abstract

Neural Dynamics and the Geometry of Population Activity

Abigail A. Russo

A growing body of research indicates that much of the brain's computation is invisible from the activity of individual neurons, but instead instantiated via population-level dynamics. According to this 'dynamical systems hypothesis', population-level neural activity evolves according to underlying dynamics that are shaped by network connectivity. While these dynamics are not directly observable in empirical data, they can be inferred by studying the structure of population trajectories. Quantification of this structure, the 'trajectory geometry', can then guide thinking on the underlying computation. Alternatively, modeling neural populations as dynamical systems can predict trajectory geometries appropriate for particular tasks. This approach of characterizing and interpreting trajectory geometry is providing new insights in many cortical areas, including regions involved in motor control and areas that mediate cognitive processes such as decision-making. In this thesis, I advance the characterization of population structure by introducing hypothesis-guided metrics for the quantification of trajectory geometry. These metrics, trajectory tangling in primary motor cortex and trajectory divergence in the Supplementary Motor Area, abstract away from task-specific solutions and toward underlying computations and network constraints that drive trajectory geometry.

Primate motor cortex (M1) projects to spinal interneurons and motoneurons, suggesting that motor cortex activity may be dominated by muscle-like commands. Observations during reaching lend support to this view, but evidence remains ambiguous and much debated. To provide a different perspective, we employed a novel behavioral paradigm that facilitates comparison between time-evolving neural and muscle activity. We found that single motor cortex neurons displayed many muscle-like properties, but the structure of population activity was not muscle-like. Unlike muscle activity, neural activity was structured to avoid 'trajectory tangling': moments where similar activity patterns led to dissimilar future patterns. Avoidance of trajectory tangling was present across tasks and species. Network models revealed a potential reason for this consistent feature: low trajectory tangling confers noise robustness. We were able to predict motor cortex activity from muscle activity by leveraging the hypothesis that muscle-like commands are embedded in additional structure that yields low trajectory tangling.

The Supplementary Motor Area (SMA) has been implicated in many higher-order aspects of motor control. Previous studies have demonstrated that SMA might track motor context. We propose that this computation necessitates that neural activity avoids 'trajectory divergence': moments where two similar neural states become dissimilar in the future. Indeed, we found that population activity in SMA, but not in M1, reliably avoided trajectory divergence, resulting in fundamentally different geometries: cyclical in M1 and helix-like in SMA. Analogous structure emerged in artificial networks trained without versus with context-related inputs. These findings reveal that the geometries of population activity in SMA and M1 are fundamentally different, with direct implications regarding what computations can be performed by each area.

The characterization and statistical analysis of trajectory geometry promises to advance our understanding of neural network function by providing interpretable, cohesive explanations for observed population structure. Commonality between individuals and networks can be uncovered and more generic, task-invariant, fundamental aspects of neural response can be explored.

Table of Contents

List of Figures

Acknowledgements

The PhD experience has been entirely more wonderful than I ever dared to bargain. I've been unfathomably fortunate to receive the mentorship and support I have, without which none of this would have been possible.

First and foremost, I want to thank my parents, Kathleen and Paul Russo for their constant love and support. Over the years, they have not only provided practical support, but they've also genuinely engaged with my science. Explaining my work to them has proved equivalent to explaining my work to a neuroscientist outside of my field. They ask truly insightful questions and see straight to the heart of the matter. It's clear to me that the traits I've learned are critical to being a good scientist- curiosity and a willingness to ask questions- run in the family. I would like to thank my brothers for their love, encouragement, and goofiness. I also want to thank the little ones in my life- my nieces, my nephews, and my dog Theo- for making sure I take time to play.

I want to thank my friends especially Dana Petermann, Michelle Farbaniec, Jenny Ouchveridze, Abby Finkelstein, Georgia Pierce, and Macayla Donegan. They have held my hand and talked me through many trying experiences, personal and professional, with patience and unwavering love.

I would like to thank my undergraduate research mentor Professor Donald Katz. My time in his lab introduced me to academic research and the wonderful experience I had there inspired my decision to go to graduate school.

I would especially like to thank Professor Eve Marder, without whom I would not be here today. During my time at Brandeis, she spent untold hours answering my questions, expressing an inexplicable confidence in my future success, and offering professional and personal wisdom that will continue to guide me for years to come. It is entirely her doing that I found myself at Columbia at just the right time to pick up a wonderful project in Mark's lab. She is an exceptional role model for everyone in the field and a guiding light whenever I find myself in the dark.

I would like to thank all members of the Churchland lab, past and present, who provided a welcoming, balanced, challenging, and thoughtful environment. On many occasions, lab members have dropped what they were doing to help me troubleshoot or work through an idea. I'd especially like to thank Sean Perkins who was instrumental early in the project and Yana Pavlova who provided the monkeys with unparalleled care and who provided me with endless amounts of chocolate. I would especially like to thank Brian London whose early work on the project jump started my PhD. His efforts enabled me to take on what would become a very computational project before I had written a single piece of code. In fact, he taught me to code, a skill which was crucial to my PhD and will undoubtedly serve me for years to come.

I would like to thank the members of the Theory Center especially my collaborators, Sean Bittner, Ramin Khajeh, and Jeff Seely, whose work has deepened our understanding of the data, clarified our thinking, and enriched the story we could tell.

I would like to thank my committee members- Roozbeh Kiani, Rui Costa, John Cunningham, and Larry Abbott. Committee meetings have always been a delightful, engaging process. They greatly

matured the work and refined my thinking. I would especially like to thank John and Larry whose creativity and exceptional intuition profoundly enhanced my work. They welcomed me into the world of computational neuroscience and in doing so, have steered my career.

I would like to express my deep gratitude to my research advisor, Mark Churchland. When I entered his lab, I hadn't taken a math class since high school and had hadn't written so much as a line of code. I doubted I was capable but he assured me I would learn along the way. Indeed, I did thanks to his exceptional mentorship. I exit graduate school a far more creative, confident, and mature scientist than I entered. I've seen this growth not only in myself but in all of his trainees. Due to Mark, and the culture of the lab he created, my time in graduate school has been challenging, light-hearted, fulfilling, and fun. It has been a great pleasure to spend time working and thinking with him.

for my parents, Kathleen and Paul Russo,

and Professor Eve Marder

Chapter 1 **Introduction**

A fundamental goal of neuroscience is to understand neural computation. What function is performed during a particular behavior and how is it instantiated by a network of neurons? In addressing these questions, we favor hypotheses of neural computation that are high-level, interpretable, and behaviorally-relevant. Yet it isn't always clear how such ideas translate into network-level instantiations. In this dissertation, I present my work attempting to bridge these two levels of understanding through studying the geometric properties of population activity.

Recent advances in neuroscience have been driven by the belief that neural computations are instantiated via population-level dynamics (Driscoll, Golub, & Sussillo, 2018). In this view, the activity of single neurons in a network reflect a modest number of population activity patterns; abstract, time-varying signals that cannot be observed directly, but represent the correlated activity of the neural population (Pandarinath, Ames, et al., 2018; Shenoy, Sahani, & Churchland, 2013). During behavior, these signals evolve in time by obeying consistent dynamic rules. Thus, although neural computations and dynamics are not directly observable, their signatures can be inferred by studying the structure of population-level neural activity.

Population-level structure can be characterized in numerous ways: shape and curvature of the trajectory follows over time, the speed of evolution, and distances between trajectories corresponding to different conditions. Such features can be broadly classified into two categories: the structure of trajectories within a given space and the differential exploration of neural subspaces across times and conditions. I posit that the characterizing these aspects of population structure inform the specific form of the active dynamics and how these dynamics might change across times and conditions.

While such characterization has clear value, here I argue that our knowledge of the underlying computation can be further deepened by abstracting away from a specific solution and toward more task-invariant properties and underlying constraints of the network. To this end, I present metrics of geometric properties that quantify how the neural state evolves across times and conditions and can be constructed to test specific hypotheses. Such metrics seek to measure the underlying drives and constraints that result in the observed population structure. In this way, various aspects of the population structure (observed in different tasks or individuals) might be understood as different manifestations of the same underlying geometric property. Unlike other methods for studying neural dynamics, geometric properties can be informative whether or not the underlying dynamics are linear and generally requires few assumptions (*e.g.* regarding the dimensionality of the system). Additionally, geometric properties facilitate comparison between empirical and artificial networks, and between network activity and behavior.

Overview of dissertation

In this **Chapter 1**, I will detail approaches for characterizing and interpreting population activity structure. First, I will describe dimensionality reduction and state-space visualization which will provide basic intuition for which features of the data warrant characterization. I will then describe how the shape of trajectories within a space over time and conditions, 'trajectory structure', can inform what underlying dynamics might be present. I also describe how differential neural covariance or 'subspace' exploration across time and tasks may reflect the functional connectivity of the network and provide a means for altering what dynamics are active and communication between neural regions. Next, I will describe methods for interpreting such population structure. I argue that characterizing families of artificial networks in terms of population structure can guide thinking on how structural motifs relate to the underlying computations. Finally, I will describe the study of geometric properties and argue that such analyses have the potential to reveal more fundamental aspects of the network that drive population structure.

0 presents an application of these approaches to primary motor cortex (M1). Neural activity was recorded during an extended movement task in which monkeys grasped a hand-pedal and cycled through a virtual environment for a prescribed duration. The dominant structure of motor cortical activity did not resemble that of muscle activity. Rather, it expressed repeating circular structure with an organized relationship between conditions. I argue that these features can be summarized with a geometric property: low 'trajectory tangling'. Indeed, optimization for low trajectory tangling drives the emergence of analogous geometry in artificial networks. Further modeling reveals that low trajectory tangling enables motor cortex to produce patterns of muscle activity in a noise robust fashion.

Chapter 3 presents another application of these approaches to the supplementary motor area (SMA), a high-order motor cortex. Neural activity was recorded during the same task as for motor cortex and the population structure was compared between regions. SMA activity was characterized by helical structure in contrast to the repeating circular structure observed in M1. These differences emerged in analogous, idealized network models that were trained with or without contextual information. The difference in structure for both empirical and model networks was summarized with a geometric property: low 'trajectory divergence'. Network modeling reveals that low trajectory divergence is necessary for networks that guide movement over long time-scales.

Chapter 4 offers some concluding remarks and future directions. I discuss important considerations for validating our interpretations of population structure and propose how future work might relate population structure to circuit properties and behavior.

Characterizing population structure

Visualizing and characterizing population activity can inspire hypotheses regarding the underlying computation. In this section, I outline tools for visualizing population activity and for characterizing trajectory structure and neural covariance across time and tasks. These motifs of population structure are useful for understanding the active dynamics and computation of the system.

Visualizing population activity

The desire to record more neurons simultaneously has driven numerous technological advances in neuroscience. As our technical ability increases, we require statistical tools for analyzing large-scale datasets. Dimensionality reduction protocols provide a means to explore such datasets and illuminate population response structure (Cunningham & Yu, 2014). If population dynamics exist, complexity at the level of single neurons will give way to simpler organizing principles at the level of the population. Any salient structure can then inspire subsequent analyses.

Principal components analysis (PCA), the dimensionality reduction protocol predominantly considered here, identifies lineally uncorrelated population activity patterns that are optimized to capture variance. While many other dimensionality reduction algorithms have been developed to extract hypothesis-guided features of the data (Churchland et al., 2012; Kobak et al., 2016; Pandarinath, O'Shea, et al., 2018), PCA remains a useful tool for exploring the basic features and the dominant population structure of a dataset. Notably, PCA can be valuable whether or not underlying dynamics are hypothesized. For example, visualizing the dominant patterns of muscle activity can yield intuition for what types of signals need to be generated by motor cortex.

Once population activity patterns have been identified, structure can be visualized by plotting signals verses one another in 'state-space'. Details on identifying computationally relevant structure will be provided in the next section. Here, I wish to emphasize that while state-space views can be a critical tool for inferring underlying dynamics, they are not inherently meaningful. Indeed, some computations are best understood when plotted verses time (even if dimensionality reduction is applied). In such cases, state-space views may be confusing or misleading. For example, neural activity during decision-making tasks that require evidence accumulation is often

best visualized when plotted verses time. In such cases, neural activity is often well-described by bounded integration. Activity is projected onto a single dimension that reflects the decision and is plotted verses time. In this view, decision variables may fluctuate on single trial and moment-to-moment changes of mind can be visualized yielding deeper insights into the decision-making process.

Further, some dynamics can be visualized equally well when plotted verses time as in state-space such as computations that are dominated by fixed points. When fixed-point dynamics dominate, neural activity reaches a plateau when visualized in time or settles into a localized point when visualized in state-space. Depending on whether the focus is to determine the time course of such stabilization or to determine how deviations from the ideal location affect behavior, either visualization may be preferred.

Generally, understanding structure in high-dimensional data is aided by exploration. Even if clear, *a priori* hypotheses are present about the population structure, skipping single-unit visualization all-together is ill-advised. Familiarity with single-unit responses will ensure that population signals are representative and capture meaningful signals in the data. Single-unit responses also occasionally suggest features of the population structure that can then be verified. Further, it is prudent to employ basic dimensionality reduction techniques before hypothesis-guided ones. In this way, one can gain intuition for which aspects of the population structure may be a consequence of simpler phenomenon and which are truly indicative of underlying computation and dynamics (Elsayed & Cunningham, 2017).

Identifying computationally-relevant trajectory structure

Neural dynamics cannot be observed directly or at the level of single-neuron responses. Instead, they can be inferred by studying 'trajectory structure': how neural population activity evolves across time and conditions with respect to one another. Features such as the shape and curvature of the trajectory, speed of evolution, and distances between trajectories corresponding to different conditions will all be considered structure (Williamson, Doiron, Smith, & Yu, 2019). Structure will be considered a 'motif' if its relationship to an underlying computation can be interpreted.

As in artificial networks (Sussillo & Barak, 2013), interpretability of neural population structure will be aided when the dynamics are low dimensional and linear. While this may not be broadly true, trajectory structure can be studied in locally linear portions of state-space while intuition is built. For example, it has been proposed that primary motor cortex may express separate dynamics during movement preparation and movement generation (Ames, Ryu, & Shenoy, 2014; Kao, 2018). During movement preparation, population activity converges onto a single point in state-space corresponding to the target location for a given condition. This target-specific preparatory state then seeds oscillatory structure during movement generation. In this way, population structure implies dynamics: converging trajectories imply a stable fixed point, diverging trajectories imply an unstable fixed point, oscillations about a single point imply rotational dynamics (Williamson et al., 2019).

The clear mapping from motor cortical population structure to its underlying dynamics has situated this region as a wonderful proving ground for tools and analyses for the study of neural dynamics. In this region, linear approximations provide a faithful summary of the full population response. Yet generalizing these approaches to other regions may require identifying computationally

relevant structure that is high dimensional and non-linear. This is a notoriously challenging proposition and interpreting the precise form of the dynamics (*i.e.* system identification) may not be possible. However, we can still identify computationally relevant motifs by comparing trajectory structure across conditions.

We can do this generally, without knowing the specific form of the dynamics or requiring linearity, if we assume that the system employs 'simple' dynamics (Sussillo, Churchland, Kaufman, & Shenoy, 2015). In network-terms, this assumption implies that few modes are active at a given moment in time and the underlying flow-field changes smoothly and slowly over state-space. In more general terms, this means that the system is 'well-behaved': neural activity will vary smoothly across conditions and variance across conditions will map onto behavioral variance. This also implies that neural activity that is intermediate between two conditions will yield intermediate behavior (Mante, Sussillo, Shenoy, & Newsome, 2013; Remington, Narain, Hosseini, & Jazayeri, 2018; Wang, Narain, Hosseini, & Jazayeri, 2018) and the network will be robust to noise (Russo et al., 2018; Sussillo et al., 2015).

Trajectory structure across conditions can also inspire hypotheses regarding the underlying computation. Orderly organization along a behaviorally-relevant parameter suggests that the network participates in a computation that requires that parameter. In dorsal medial frontal cortex, trajectories are ordered according to interval duration in an interval timing task (Remington, Narain, et al., 2018). In prefrontal cortex, trajectories are ordered by context and stimulus coherence in a decision-making task (Mante et al., 2013). Similarly, lack of such organization indicates that the computation performed by that region is independent of that parameter (*e.g.* is downstream of that computation). For example, in the supplementary motor area but not primary

motor cortex, trajectories separate according to whether the movement is triggered externally or internally (A. H. Lara, Cunningham, & Churchland, 2018; Mushiake, Inase, & Tanji, 1991). While the SMA may participate in this computation, the collapse across context in M1 is consistent with the hypothesis that M1 computation focuses on low-level pattern generation, independent of how the movement was triggered.

Such insights are indebted to state-space views of population activity. Low-dimensional state-space visualization of population activity can reveal structure that was invisible at the level of single neurons. For example, views of single-neurons and even views of population activity plotted verses time yield a seemingly unrelated relationship between preparatory activity and movement related activity. Such relationships become well-defined in the population structure (Churchland, Cunningham, Kaufman, Ryu, & Shenoy, 2010). Yet ultimately, trajectory visualizations are a fundamentally impoverished view of the data. Only 2-3 dimensions can be visualized at a time but 10-20 dimensions are typically required to sufficiently capture neural variance. Further, smoothness across conditions is implied by the dynamical systems view, but does not alone indicate dynamics (Elsayed & Cunningham, 2017). Thus, to truly understand neural computation, we must turn to analyses that take high-dimensional structure into account.

Characterizing neural covariance

Neural networks, artificial and empirical, are highly interconnected resulting in correlations between single-unit activity (Cohen & Kohn, 2011). Because single neurons do not act independently, population activity does not span the full-dimensional space it hypothetically could. Activity is instead constrained to lie on a low-dimensional surface, termed the "neural manifold"

(Gallego, Perich, Miller, & Solla, 2017). Further, it has been demonstrated that neural covariance changes under different stimuli and behavioral conditions. Therefore, the "subspace" explored by neural activity during a given task may reflect *functional* connectivity that is indicative of the active neural computation rather than a hard constraint imposed by fixed network connectivity. Here, I review and integrate these interpretations of neural covariance.

The interpretation that neural manifolds reflect fixed network connectivity is supported by the apparent stability of the manifold across tasks (Gallego et al., 2018). In this work, monkeys performed a handful of motor tasks and neural activity was found to occupy the same manifold. It should be noted however, that the *a priori* expectation here is not necessarily clear. That is, if two tasks are sufficiently similar (e.g. reaching at different speeds, curved vs straight reaches), they may be driven by the same underlying dynamics and neural activity thus ought to occupy the same space across tasks (Churchland et al., 2012). Thus, the expectation that different tasks ought to occupy different subspaces is predicated on studying tasks are "sufficiently" different, a notion that is ill-defined at the moment. As will be described below, it has been demonstrated that neural correlations do change dramatically as a function of behavioral conditions. Perhaps stronger evidence that neural manifolds reflect fixed network connectivity comes from BMI studies of learning (Sadtler et al., 2014). Here, monkeys were trained to control a cursor via brain control. After monkeys learned the task, the mapping between neural activity and cursor control was manipulated so as to maintain (within-manifold) or disrupt (out-of-manifold) neural correlation. Out-of-manifold mappings were generally much harder to learn but could be acquired over the course of several months, a timeline which accords with changing synaptic connectivity (Oby et al., unpublished data).

The study of neural manifolds is new and rapidly growing. Going forward, it will be critical to distinguish between three non-exclusive explanations for these observations. As commonly proposed, manifolds may reflect the synaptic connectivity of the network. In this case, manifolds may be changed but slowly, on the time-course consistent with forming new synapses. It also may be that manifolds represent hard constraints on the system that have yet to be discovered. If this is the case, then a subset of out-of-manifold patterns of activity are simply impossible for the network to produce. Finally, and perhaps most intriguingly, the space explored by the network may reflect the "functional connectivity" of the network. Considerations from artificial networks informs that the effective connectivity of a network can be dramatically and rapidly changed with inputs. This apparent change in connectivity (and in neural correlation) could allow the network to perform very different computations using very different active dynamics. To distinguish this possibility, I will refer to neural spaces explored under different behavioral conditions as "neural subspaces", the term commonly used in this literature.

There is a growing body of evidence suggesting that neural networks, both empirical and artificial, exploit different neural subspaces to implement distinct dynamics across behaviors (Kaufman, Churchland, Ryu, & Shenoy, 2014; Machens, Romo, & Brody, 2010; Mante et al., 2013; Miri et al., 2017; Raposo, Kaufman, & Churchland, 2014; Russo et al., 2018). For example, neural activity in motor cortex occupies orthogonal subspaces during movement preparation and movement generation (Elsayed, Lara, Kaufman, Churchland, & Cunningham, 2016). This may allow the same circuit of neurons to perform two very different computations that require very different underlying dynamics (a stable fixed-point in the case of movement preparation and rotational dynamics in the case of movement generation). This strategy may also provide an explanation for how the same neurons can be active during both stages yet muscle activity is only produced during movement

generation. More broadly, neural circuits may leverage distinct subspaces whenever two computations require "sufficiently" different underlying dynamics. Indeed, this is suggested by data analyzed in **0** of this thesis (Russo et al., 2018). Briefly, monkeys perform a pedaling task that requires they cycle in either a 'forward' or 'backward' direction which require vastly different patterns of muscle activity. Indeed, motor cortical activity during this task occupies non-overlapping subspaces during forward and backward pedaling.

While it is clear that neural activity tends to occupy a low-dimensional space, it still remains an open question as to whether this observation is a true constraint due to the underlying connectivity or is a consequence of insufficient sampling of the system (i.e. under a wide enough variety of behaviors). In either case, summarizing neural activity in terms of the space explored is incomplete: the trajectories may also be constrained in how they move through space (Russo et al., 2018). As I will argue in later sections, there may be more fundamental constraints that shape both the spaces explored and the structure of the trajectories through that space. But first, it is worth verifying that these motifs of population structure (trajectory structure and neural covariance) are computationally relevant. In the next section, I will describe how training artificial neural networks can help to check and validate our intuitions for population structure motifs.

Predicting and interpreting population structure

The previous sections described data-driven approaches for characterizing population-level features of empirical neural activity. Here, we turn our attention to hypothesis-driven methods for interpreting and predicting these features (Williamson et al., 2019). Artificial networks can be

trained to accomplish a specific task without prescribing the form of the solution. Motifs of the solutions they find can then be characterized and compared to the data. By training ensembles of networks, we can begin to understand the solution space and what different network activity structures might have in common (Prinz, 2010). We can learn from what types of models naturally express population structure motifs that match the data and what constraints push an artificial network into a realistic regime (Sussillo et al., 2015). Further, we can use intuition from the study of dynamics to identify statistical features of both artificial and empirical data that might reveal more fundamental aspects of the computation from which the observed population structure motifs arise.

Studying population structure in artificial networks

Given a high-level hypothesis about the function of a network, what can one *a priori* expect neural activity to look like? Given hypotheses born of observed neural activity, how can one validate interpretations of structural motifs? The study of artificial neural networks has proved an invaluable tool to address such questions. Before we begin, we require an important first step: translating high-level, language-based hypotheses into a goal that can be instantiated by a network. That is, we determine precisely what inputs the network will receive and what outputs we require. The computation is then defined in terms of transforming inputs into outputs. As will be described more fully in the next section, this process of reframing hypotheses in network-terms can itself provide clarifying insight into how such computations might be instantiated by population activity.

Once such clarifications have been made, a network can be trained to perform the desired task. Because networks are trained to perform a certain task without specifying how to perform it, we

are given relatively unbiased insight into how population structure should look to accomplish the prescribed goal (Sussillo & Barak, 2013). Such insight can be deepened further by training a family of networks with different random initializations (Prinz, 2010; Russo et al., 2018). This will provide a range of viable solutions and give more confidence that any consistent population structure motifs are stable, real features resulting from an underlying computation. In **Chapter 3** of this thesis, I present two such families of networks that were both trained to produce the same output but received different inputs. I then characterized the population structure of all networks and found that within each family, very similar population structure motifs were expressed but there were striking differences between the two families.

Once the space of potential solutions is characterized, we can compare artificial networks and empirical data. If the models use entirely different strategies (as evidenced by very different population structure), it may be that the model itself is fundamentally flawed. Indeed, good models of the data should replicate motifs of population structure and such features can be leveraged to rule out models (Elsayed et al., 2016; Williamson et al., 2019). Alternatively, the model may not be flawed but rather under-constrained. For example, unregularized networks may find overly elaborate solutions and constraints need to be added to encourage realistic solutions. For example, models trained to produce patterns of muscle activity became both more like the neural data in terms of population structure and also more noise robust when regularization was added to encourage the model to use simple solutions (Sussillo et al., 2015). Apart from the practical utility of matching data, the emergence of a natural solution as a function of such constraints suggests that the neural network might undergo similar constraints. Indeed, simpler solutions may naturally be more robust to noise- a desirable property for artificial and empirical neural networks alike.

Metrics of geometric properties

Trajectory structure and neural covariance are important and computationally-relevant features of empirical data. Artificial and empirical neural networks that putatively share a high-level computational goal, share motifs of population structure. Indeed, the appearance of realistic population structure motifs is generally the goal of training such networks. Yet, artificial networks are often treated as 'black boxes' and a deep understanding of how that population structure instantiates the computation is lacking (Driscoll et al., 2018). Here, I propose methods for pursuing this line of questioning: hypothesis-driven metrics of geometric properties.

Rather than characterizing motifs of population structure, the goal becomes to identify the underlying properties of the population geometry that are expressed by the characterized motifs. In doing so, we can begin to abstract away from particular solutions and toward more fundamental, stable properties of the network. A key question becomes, what fundamental properties of the network (shaped by inherent constraints or by the task the network was trained to perform) would necessitate the observed structural motifs? We seek answers to this question in network-terms that begin to open the black box and bridge language-level hypotheses of computation and observed population activity.

In seeking such 'motif-driving' geometric properties, it is helpful to compare motifs that co-occur across empirical datasets and across network instantiations. For example, in **Chapter 2** of this thesis, I characterize several motifs of motor cortical population structure that were present in both monkeys and across artificial networks. Population trajectories expressed oscillatory structure that rotated in the same direction across conditions. Trajectories were simple and circular unlike the

elongated, complex muscle activity. Across conditions, motor cortical responses also explored non-overlapping subspaces. These motifs were present in artificial networks trained to produce the empirical muscle activity. We argue that, taken together, these motifs are driven by the same underlying geometric property- the avoidance of 'trajectory tangling': moments where trajectories cross in state-space.

We propose that low trajectory tangling is an inherent and general constraint of motor cortex due to strong internal dynamics. Identifying properties of population geometry can also reveal properties that are more closely tied to the specific computation the network performs. In **Chapter 3**, I describe such an instance in the SMA. Again, we begin by characterizing motifs of the population structure in this region and in analogous, idealized artificial networks. We find that these motifs can be understood cohesively as being driven by the need to have low 'trajectory divergence'. Further, we propose that this geometric property is necessary to guide movement on extended time-scales.

Thus, by characterizing population structure motifs, we can identify the underlying geometric properties that drive them. We now wish to extend and validate these findings to better interpret their functional relevance. First, we wish to validate that the geometric property of interest is indeed driving the observed motifs. This can be accomplished by jointly optimizing for the geometric property and the hypothesized output of the system to determine if population structure motifs emerge naturally (**Chapter 2**). Notably, this strategy can only be used if there is a high-degree of confidence that the output of the system is properly identified. Second, we wish to clarify the functional relevance of the geometric property. This can be accomplished by using 'trajectory constrained' modeling to enforce that networks express a prescribed degree of the geometric

property of interest. These networks are thus parameterized along the property interest and can be probed for computationally-relevant properties such as robustness to noise (**0** and **Chapter 3**).

In summary, this approach promises to advance our understanding of neural network function by providing interpretable, cohesive explanations for observed population geometry. Commonality between individuals and networks can be uncovered and more generic, task-invariant, fundamental aspects of neural response can be explored.

Chapter 2 **Motor cortex embeds muscle-like commands in an untangled population response**[1]

Primate motor cortex projects to spinal interneurons and motoneurons, suggesting that motor cortex activity may be dominated by muscle-like commands. Extensive observations during reaching lend support to this view, but evidence remains ambiguous and much-debated. To provide a different perspective, we employed a novel behavioral paradigm that affords extensive

[1] This chapter was published as Motor cortex embeds muscle-like commands in an untangled population response, Nature Neuroscience (2018) with co-authors Sean R. Bittner[1,2], Sean M. Perkins[2,3], Jeffrey S. Seely[1,2], Brian M. London[4], Antonio H. Lara[1,2], Andrew Miri[1,2,7], Najja J. Marshall[1,2], Adam Kohn[8], Thomas M. Jessell[1,2,5,6,7], Laurence F. Abbott[1,2,5,9,11], John P. Cunningham[2,10,11,12], and Mark M. Churchland[1,2,5,10*]

1. Dept. of Neuroscience, Columbia University Medical Center, New York, NY 10032, USA.
2. Zuckerman Institute, Columbia University, New York, NY 10027, USA.
3. Dept. of Biomedical Engineering, Columbia University, New York, NY 10027, USA.
4. SeatGeek, New York, NY
5. Kavli Institute for Brain Science, Columbia University Medical Center, New York, NY 10032, USA.
6. Howard Hughes Medical Institute
7. Depts. of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY 10032, USA
8. Dept. of Ophthalmology and Visual Sciences, Dominick Purpura Department of Neuroscience, Albert Einstein College of Medicine, Yeshiva University, Bronx, New York, USA.
9. Dept. of Physiology and Cellular Biophysics, Columbia University Medical Center, New York, NY 10032, USA.
10. Grossman Center for the Statistics of Mind, Columbia University, New York, NY 10027, USA.
11. Center for Theoretical Neuroscience, Columbia University Medical Center, New York, NY 10032, USA.
12. Dept. of Statistics, Columbia University, New York, NY 10027, USA.

comparison between time-evolving neural and muscle activity. We found that single motor cortex neurons displayed many muscle-like properties, but the structure of population activity was not muscle-like. Unlike muscle activity, neural activity was structured to avoid 'tangling': moments where similar activity patterns led to dissimilar future patterns. Avoidance of tangling was present across tasks and species. Network models revealed a potential reason for this consistent feature: low tangling confers noise robustness. Finally, we were able to predict motor cortex activity from muscle activity by leveraging the hypothesis that muscle-like commands are embedded in additional structure that yields low tangling.

Introduction

For fifty years, a central question in motor physiology has been whether motor cortex activity resembles muscle activity, and if not, why not (Evarts, 1968)? Primate motor cortex is as close as one synapse to the motoneurons (Rathelot & Strick, 2009) and single action potentials in corticospinal neurons can measurably impact muscle activity (Cheney & Fetz, 1980; Schieber & Rivlis, 2007) suggesting that motor cortex may encode muscle-like commands (Ajemian et al., 2008; Herter, Korbel, & Scott, 2009; Morrow, Pohlmeyer, & Miller, 2009; Sergio, Hamel-Paquet, & Kalaska, 2005; Todorov, 2000). Yet motor cortical responses often differ from patterns of muscle force, motivating the hypothesis that motor cortex might primarily encode movement velocity or direction (Georgopoulos, Schwartz, & Kettner, 1986; Moran & Schwartz, 1999b; Schwartz, 1994, 2007). Alternatively, it has been proposed that non-muscle-like response features may reflect network or feedback dynamics (Churchland & Cunningham, 2014; Churchland et al., 2012; Kaufman et al., 2016; Lillicrap & Scott, 2013; Maier, Shupe, & Fetz, 2005; Michaels, Dann,

& Scherberger, 2016; Rokni & Sompolinsky, 2012; Seely et al., 2016; Shenoy et al., 2013; Sussillo et al., 2015). Many studies, largely focused on reaching, have produced little consensus (Aflalo & Graziano, 2007; Fetz, 1992; Georgopoulos, Naselaris, Merchant, & Amirikian, 2007; Moran & Schwartz, 2000; Mussa-Ivaldi, 1988; Reimer & Hatsopoulos, 2009; Scott, 2008).

The ubiquity of reaching tasks has naturally promoted analysis of directional tuning (e.g., Ajemian et al., 2008; Georgopoulos, Kalaska, Caminiti, & Massey, 1982; Kakei, Hoffman, & Strick, 1999; Lillicrap & Scott, 2013; Scott, 1997) the interpretation of which remains debated (Georgopoulos et al., 2007; Moran & Schwartz, 2000; Mussa-Ivaldi, 1988; Sanger, 1994). More generally, reaching tasks tend to prompt hypotheses where neurons encode parameters relevant to reaching (Burnod et al., 1992; Georgopoulos et al., 1982; Georgopoulos et al., 1986; Moran & Schwartz, 1999b) or reflect reach-appropriate dynamics (Churchland & Cunningham, 2014; Churchland et al., 2012). A few studies (Hatsopoulos, Xu, & Amit, 2007; Moran & Schwartz, 1999a; Schwartz, Moran, & Reina, 2004) examined primate motor cortex during extended drawing or tracing movements, but also focused largely on directional properties (although see Fitzsimmons, Lebedev, Peikon, & Nicolelis, 2009; Foster et al., 2014). Given that the defining feature of movement is change with time, progress may benefit from detailed comparisons of time-evolving patterns of neural and muscle activity. To afford such comparisons, an ideal task would achieve the traditional goal of dissociating kinematics from muscle activity (Kakei et al., 1999; Scott, 1997), but do so in the temporal rather than spatial domain. This has been achieved during reaches (Churchland & Shenoy, 2007; Sergio et al., 2005) but more extended movements may improve the power of such comparisons.

Unlike in sensory systems where responses strongly reflect incoming stimuli, time-evolving responses in the motor system may reflect computations performed by internal and feedback

dynamics. A growing body of work seeks to understand neural responses in terms of signals that a recurrent or feedback-driven neural network would need to perform the relevant task (Hennequin, Vogels, & Gerstner, 2014; Li, Daie, Svoboda, & Druckmann, 2016; Lillicrap & Scott, 2013; Mante et al., 2013; Michaels et al., 2016; Sussillo & Barak, 2013). Although multiple network solutions are typically possible, broad principles can still apply and yield explanatory power. For example, the simple constraint of a smooth dynamical flow-field explains aspects of neural dynamics during reaching (Sussillo et al., 2015).

In the present study, we leveraged a 'cycling' task that evoked extended movements with simple kinematics driven by temporally complex patterns of muscle activity. We found that single neurons and muscles shared many temporal response properties. Yet the neural population as a whole was dominated by signals that were not muscle-like, and could not be explained by velocity / direction coding. To seek an alternative explanation, we focused on a basic principle of recurrent and feedback-driven networks: the present network state strongly influences the future state. Thus, two similar patterns of activity, observed at different moments, should not lead to highly dissimilar patterns in the near future. We refer to violations of this principle as 'trajectory tangling'. Moments of high tangling imply either a potential instability in network dynamics or a moment when the system must rely on external commands.

Tangling was often high for muscle population trajectories. This was expected. Muscles reflect descending commands and need not avoid tangling. In contrast, tangling was very low for motor cortex population trajectories. This effect was observed not only during cycling but during a reaching task, and in rodent during reach-to-grasp and locomotion. However, low tangling was anatomically specific and was not observed for primary visual or somatosensory cortex. We found that the dominant signals in motor cortex were those that naturally reduced tangling. Using an

optimization approach, we could quantitatively predict the neural population response based on only two principles: the need to encode muscle-like commands and the need to ensure low tangling. Network simulations confirm that low trajectory tangling is computationally beneficial. Networks with lower tangling are more noise robust. In summary, our data reveal a potentially general property of motor cortex: muscle-like signals are present but are relatively modest 'ripples' riding on top of larger signals that confer minimal tangling. Thus, the dominant signals in motor cortex may serve not a representational function – encoding specific variables – but rather a computational function: ensuring that outgoing commands can be generated reliably.

Results

Task and behavior

We trained two rhesus macaque monkeys to grasp a hand-pedal and cycle an instructed number of revolutions for juice reward. Cycling produced movement through a virtual landscape. Landscape color indicated whether forward virtual motion required 'forward' cycling (**Figure** 2.1**A**) or 'backward' cycling (**Figure** 2.1**B**). During each trial, the monkey progressed from one stationary target to another. Target acquisition required a stationary pedal with the target 'under' the first person perspective ((**Figure** 2.1**A,B**). The first target was acquired with a pedal orientation either straight up ('top-start') or straight down ('bottom-start'). Inter-target distance determined the required number of revolutions: 0.5, 1, 2, 4, or 7 cycles. Monkeys performed all combinations of two cycling directions, two starting orientations, and five distances. Cycling required overcoming simulated inertia and viscosity while countering the weight of an arm extended in front of a vertically oriented body. These requirements differ from those during locomotion, and had to be learned.

Behavior was highly stereotyped; note similarity of virtual-world-position traces across trials in (**Figure** 2.1**C,D**. Nevertheless, small trial-to-trial variations in cycling speed caused accumulating misalignment of kinematics with time. We therefore temporally scaled trials so that virtual-world-position traces were closely matched. Doing so revealed considerable temporal structure in neural and electromyographic (EMG) responses (**Figure** 2.1**E,F**). To summarize such structure, we computed average firing rate (**Figure** 2.1**G**) or muscle activation (**Figure** 2.1**H**) across trials. We used a narrow filter (25 ms Gaussian kernel) relative to the timescale of behavior (~500 ms cycling period) to preserve fine temporal features.

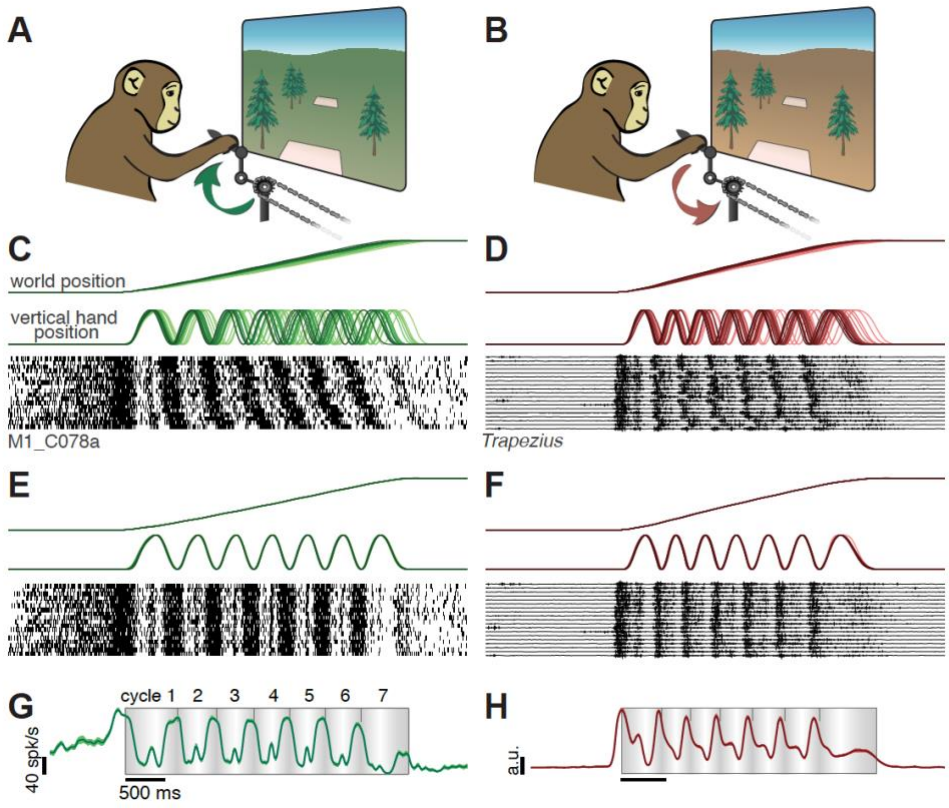

Figure 2.1 Behavioral and physiological responses during cycling

A. Schematic of the task during forward cycling. A green landscape indicated that virtual progress required cycling 'forward'. **B.** An orange landscape indicated that progress required cycling 'backward'. **C.** Behavioral data and spikes from one neuron during an example session. Data are for a single condition: forward / seven-cycle / bottom-start (monkey C). Trials are aligned to movement onset, and ordered from fastest to slowest. **D.** Behavioral data and raw trapezius EMG for one condition: backward / seven-cycle / bottom-start (monkey D). **E.** Data from C after temporal scaling to align trials. **F.** Data from D after temporal scaling. **G.** Trial-averaged and filtered neural activity for the example neuron in C,E. Envelopes show standard error of the mean (SEM; often within the trace width). Shading tracks vertical hand position: lightest at top and darkest at bottom. Small tick-marks indicate each cycle's completion. **H.** Rectified, filtered and trial-averaged EMG for the example in D,F.

We also computed trial-averaged responses (with SEMs) for key kinematic parameters such as hand velocity. Consistent with the circular pedal motion, vertical and horizontal hand velocity exhibited approximately sinusoidal profiles (**Figure** 2.2**A,B**). Top- and bottom-start movements differed in phase but were otherwise similar during middle cycles. The temporal profile of hand

velocity was repeated across middle cycles, and was slightly slower during initial / terminal cycles as angular velocity ramped up and down.

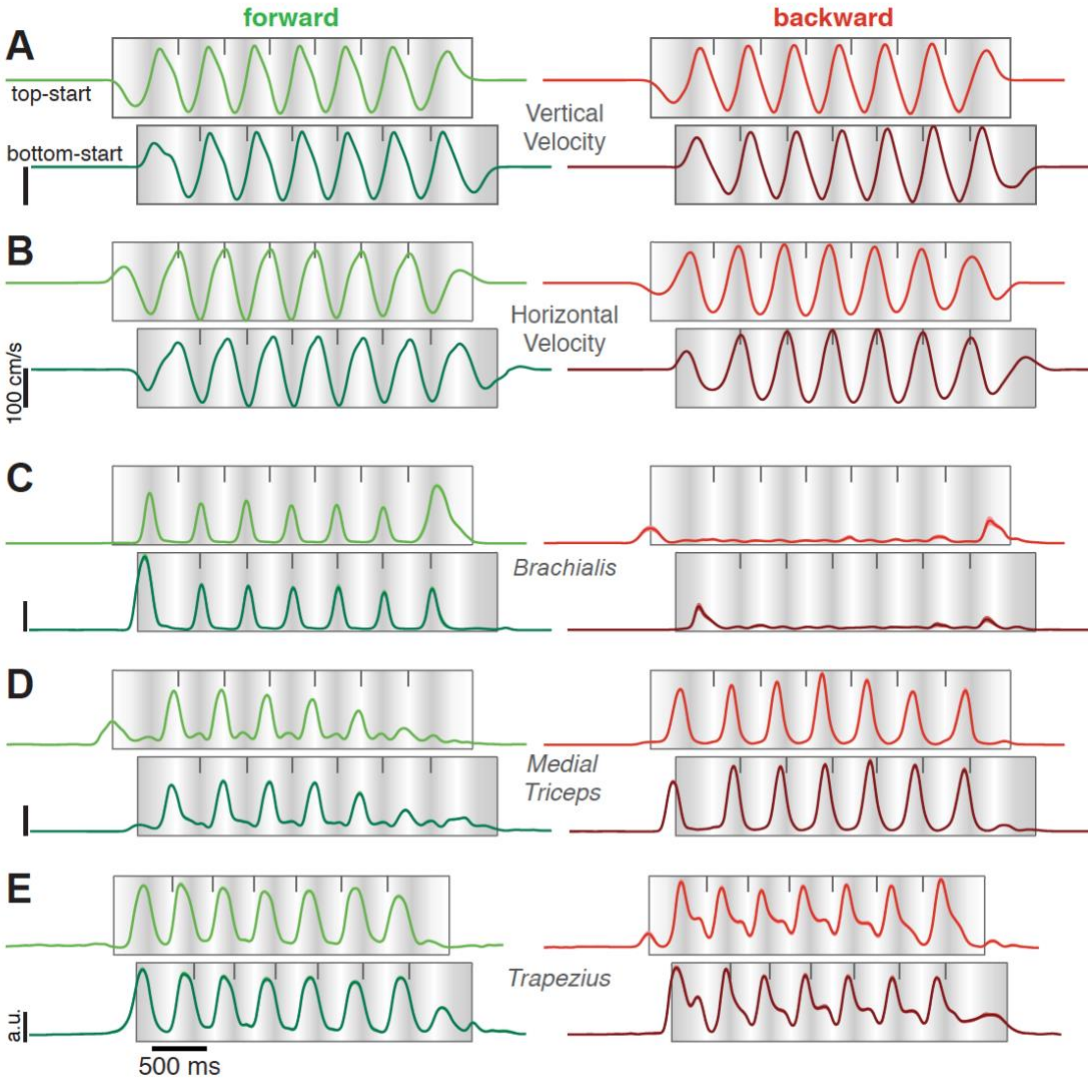

Figure 2.2 Kinematics and muscle activity during cycling

A. Vertical hand velocity, averaged across trials from a typical session (monkey C). Same format as in Fig 1G. Data are shown for seven-cycle movements for forward cycling (*green*, left column) and backward cycling (*red*, right column), and for both top-start and bottom-start movements. The latter have been shifted a half-cycle to visually align hand position between top- and bottom-start movements (*light shading* indicates the top of each cycle). Flanking traces show the SEM but are generally narrower than the trace width. Small tick-marks indicate the completion of each cycle. **B.** Horizontal hand velocity from the same session, plotted using the same format. **C.** EMG activity of *brachialis* muscle (monkey C) plotted using the same format. Flanking traces (barely visible) show the SEM. **D.** EMG activity of the *medial triceps* muscle (monkey C). **E.** EMG activity of the *trapezius* muscle (monkey D).

Intramuscular EMG recordings (35 and 29 sites in monkey D and C) concentrated on muscles that moved the shoulder and elbow and to a lesser degree the wrist (which had limited mobility given the pedal design). Muscle activity (**Figure** 2.2**C-E**) generally followed intuitions from biomechanics. For example, the *triceps* muscle extends the elbow, moving the hand away from the body. Accordingly, *triceps* activity (**Figure** 2.2**D**) peaked near each cycle's apex (*white shading*) when cycling forward and near its bottom (*dark shading*) when cycling backward. Some muscle responses were roughly sinusoidal and resembled kinematics, yet deviations from sinusoidal were common (*e.g.*, **Figure** 2.2**E**).

Single-neuron responses

Well-isolated single neurons (103 and 109, monkeys D and C) were sequentially recorded from motor cortex, including sulcal and surface primary motor cortex and the immediately adjacent aspect of dorsal premotor cortex (potential differences within this population are explored later). Recordings were localized to the region where microstimulation activated the muscles from which we recorded. Cycling evoked strong responses; nearly all neurons that could be isolated were task-modulated. Peak firing rates ranged from 16-184 spikes/s (monkey D, mean: 69 spikes/s) and 16-185 spikes/s (monkey C, mean: 76 spikes/s). Neurons displayed a variety of intricate response patterns (**Fig 3**). These patterns were statistically reliable. SEMs were small and the same pattern could be seen repeatedly across middle cycles for both top- and bottom-start conditions.

Inspection revealed three features shared between muscles and neurons. First, responses often deviated from the sinusoidal profile of kinematics (*e.g.*, **Figure** 2.2**E**-backward; **Figure** 2.3**A**-forward). Second, responses during initial / terminal cycles often displayed differences in amplitude or temporal profile compared to middle cycles (*e.g.*, **Figure** 2.2**D**-forward;

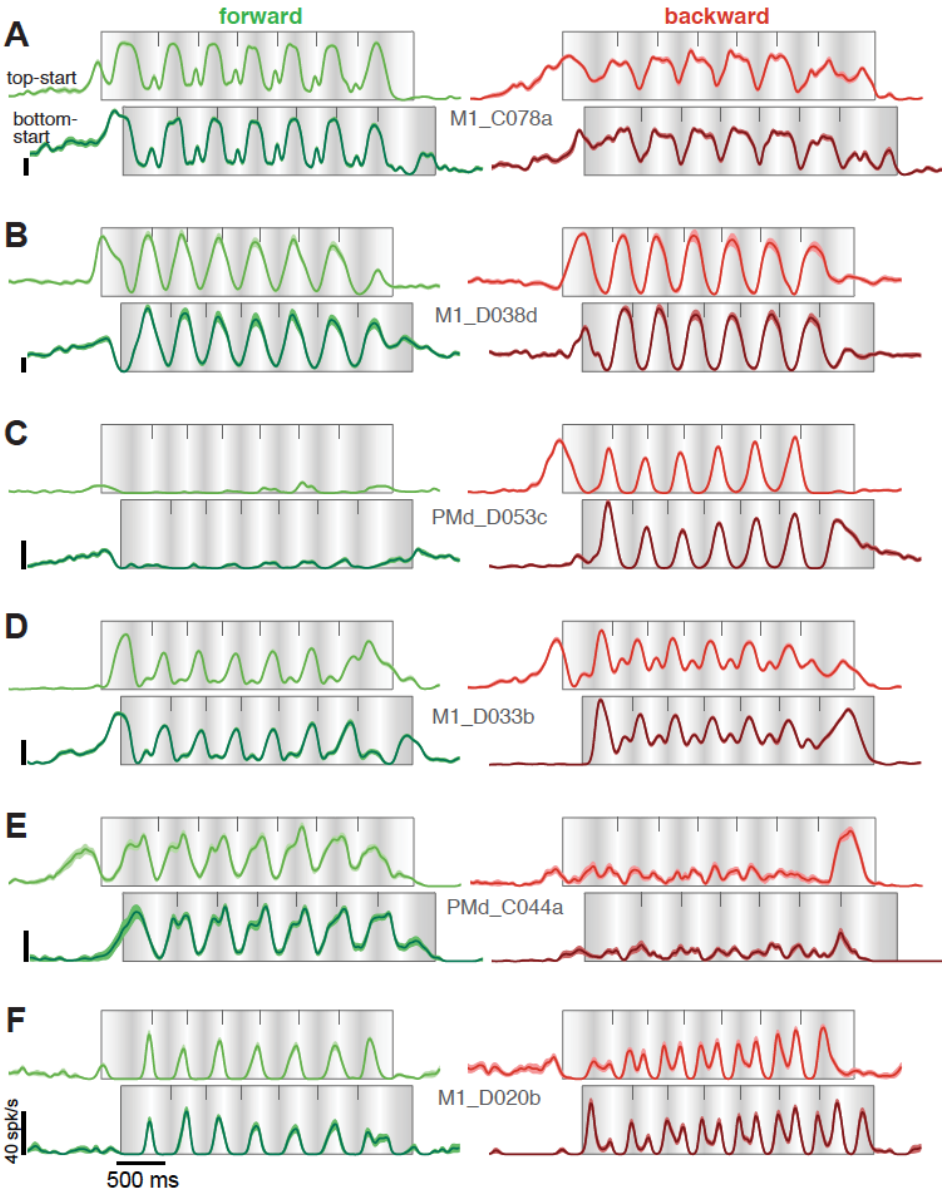

Figure 2.3 Firing rates of six example neurons recorded from motor cortex

Same format as for Figure 2. Flanking envelopes show the SEM (median of 15 trials per condition per cell). Cell names indicate area (M1 versus PMd) and monkey (D and C). All vertical calibrations are 40 spikes/s.

Figure 2.3**D**-forward; **Figure** 2.3**E**-backward). This effect presumably relates to the unique force patterns required to start and stop. Third, responses could differ between forward and backward cycling in both amplitude (*e.g.*, **Figure** 2.2**C**, **Figure** 2.3**C**) and structure (*e.g.*, **Figure** 2.2**E**, **Figure** 2.3**A**,**F**).

Consistent with these shared features, muscle responses could be successfully decoded from the neural population using a linear model (Leave-one-out-cross-validated R^2 = .80 and .78) consistent with prior studies (Griffin, Hudson, Belhaj-Saif, McKiernan, & Cheney, 2008; Morrow et al., 2009; Schieber & Rivlis, 2007). This is potentially impressive, given that a linear model is almost certainly too simplistic. This finding might suggest that motor cortex activity primarily reflects muscle-like commands. However, decoding neural activity from muscle activity was less successful (Leave-one-out-cross-validated R^2 = .54 and .50). This discrepancy in fit quality was not simply due to neural recordings being 'noisier' (having higher sampling error) than muscle recordings. The same discrepancy was observed when neural responses were de-noised using dimensionality reduction techniques (*Methods*). Thus, while muscle-like signals can be found in the neural data, there exist additional, non-muscle-like neural response patterns.

Non-muscle-like signals dominate the neural population response

To characterize population responses, we applied principal component analysis (PCA), a standard unsupervised algorithm that identifies the dominant signals in multi-dimensional data (**Figure** 2.4). Each signal is a weighted combination of individual-neuron responses, with those weights (the PCs) optimized such that a small number of signals faithfully summarizes the full population response. We first examine the signals captured by the top two PCs. Plotting these signals versus one another yields a state-space trajectory (**Figure** 2.4**C**). Each point on the trajectory (*e.g., the orange dot* in **Figure** 2.4**C**) corresponds to the neural state at one moment (*dashed line* in **Figure** 2.4**A,B**). A two-dimensional trajectory provides only a partial summary of the neural state, but the resulting visualization can still be informative and inspire hypotheses. Neural trajectories for monkey D are shown during both forward and backward cycling (**Figure** 2.4**E**, *top* and *bottom*

subpanels). Top-start and bottom-start trajectories are superimposed. For monkey C, trajectories during forward and backward cycling are also superimposed. For illustrative purposes, data are shown only for seven-cycle conditions (as in **Figs. 1-3**). Middle cycles (3-5) are highlighted in color.

Neural trajectories followed repeating orbits throughout the middle cycles. Rotating orbits are expected during cycling, in contrast to reaching (Churchland et al., 2012), and simply reflect what can be observed in single neurons: middle-cycle responses tend to repeat. Muscle trajectories also followed repeating orbits (**Figure** 2.4**D,G**). Despite this basic similarity, neural and muscle trajectories behaved differently. Muscle trajectories counter-rotated: they orbited in opposing directions for forward and backward cycling. Counter-rotation is expected given the reversal of required force patterns. For example, forward cycling requires lifting before pushing and backward cycling requires pushing before lifting. In contrast, neural trajectories co-rotated: they orbited in the same direction for forward and backward cycling. Furthermore, muscle trajectories tended to depart from circular: the orbit often possessed a kidney- or saddle-like shape. In contrast, neural trajectories were more circular or elliptical. Thus, the dominant signals in the neural population differ from those in the muscle population.

Potential explanations and caveats

A potential explanation for non-muscle-like patterns in motor cortex is that they encode directional signals such as hand velocity (e.g., Moran & Schwartz, 1999b). This explanation initially seems appealing given the present data. For example, the neural trajectory during backward cycling for monkey D (**Figure** 2.4**E**, *bottom*) visually resembles the corresponding velocity trajectory (**Figure** 2.4**F**,

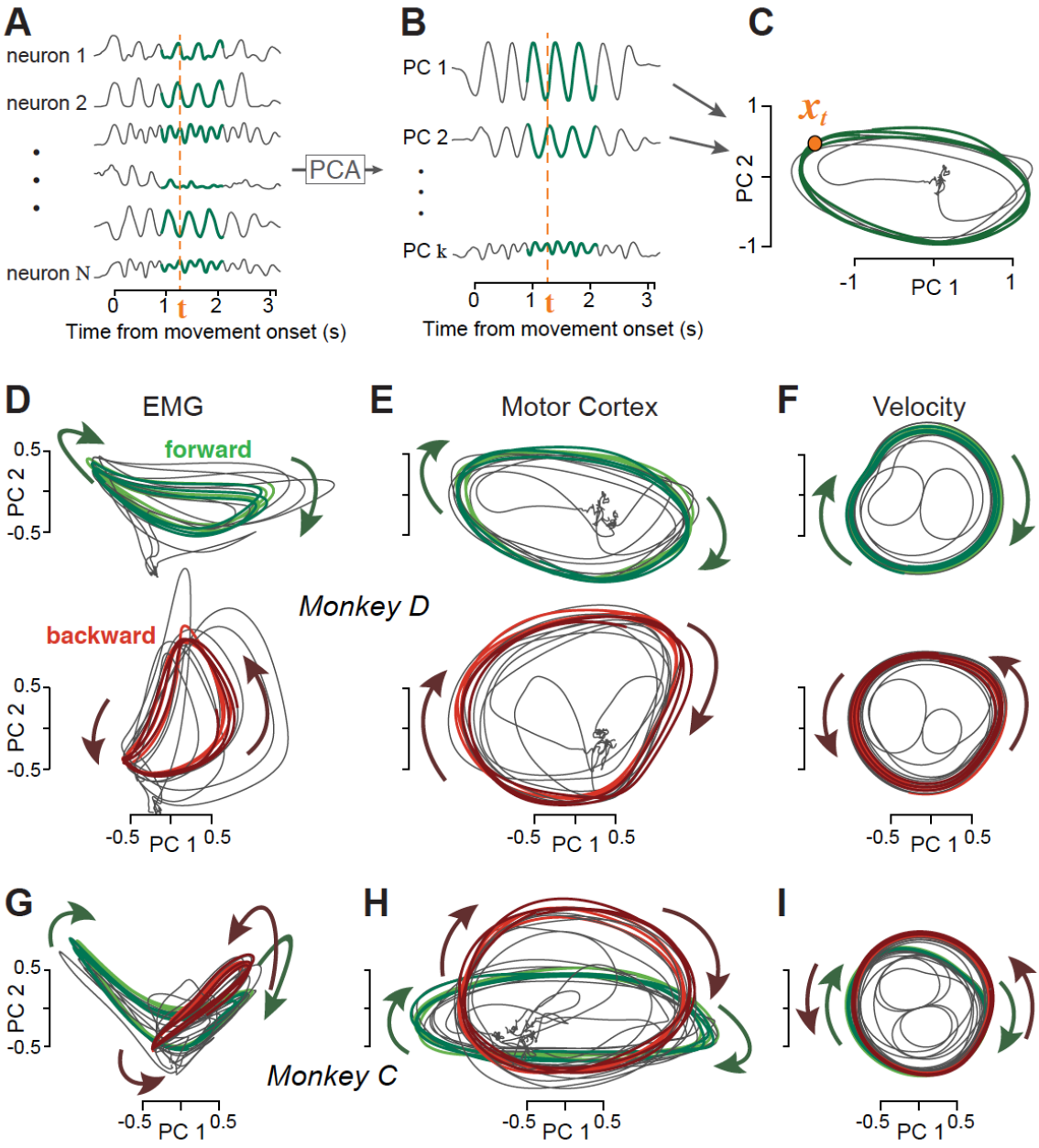

Figure 2.4 Visualization of population structure via PCA.

A. PCA operates on a population of responses (six of 103 neurons are shown, monkey D). *Green traces* highlight the middle three 'steady state' cycles, which were used to find the PCs for the present analyses (subsequent analyses consider all times for all conditions). Data are shown for only one condition – forward cycling starting at the bottom – but PCs were computed based on both forward and backward cycling and both top- and bottom-start conditions. **B.** Projections onto the PCs capture the dominant signals in the data. *Orange dashed lines* highlight the 'neural state' at a single time. That state can be summarized either using the full vector of firing rates (A) or a reduced-dimensional vector containing the values of the projections onto the top PCs (B). **C.** Neural trajectories revealed by plotting the

projection onto the second PC versus the projection onto the first PC (~35% of the total variance is captured in these two dimensions). This is equivalent to projecting the 103-dimensional neural trajectory onto the two dimensions defined by the PCs. *Orange dot* corresponds to the neural state at the same time as in A and B. **D.** Muscle trajectories captured by projecting the muscle population response onto its first two PCs (monkey D). Trajectories are shown for forward cycling (*green*) and backward cycling (*red*). Each panel overlays trajectories for top-start and bottom-start conditions (*lighter* and *darker colored traces* respectively). The same PCs were used to project data for both forward and backward cycling. **E.** Corresponding neural trajectories for the same monkey and conditions. **F.** Corresponding hand-velocity trajectories. Trajectories were produced by applying PCA to horizontal and vertical hand velocity traces across multiple sessions. This is exceedingly similar (but for a change of axes) to simply plotting average vertical velocity versus average horizontal velocity. **G,H,I**. PCA-based muscle, neural, and velocity trajectories for monkey C. Same format as D,E,F, but trajectories for forward and backward cycling are overlaid.

bottom). However, velocity trajectories necessarily counter-rotate between forward and backward cycling (the same would be true of hand direction, position, or other kinematic variables). The dominant signals in the neural data do just the opposite. Combined with the fact that single-neuron response profiles typically do not resemble hand velocity or position traces, it seems unlikely that a simple representation of kinematic parameters can explain the dominant signals in the neural data.

An alternative explanation is that the dominant neural signals may constitute descending commands to the muscles, yet may look non-muscle-like because they will be heavily modified by spinal circuitry. Cortical commands are likely integrated / low-pass filtered by the spinal cord (Shalit, Zinger, Joshua, & Prut, 2012) and may encode muscle synergies rather than individual-muscle activations (Hart & Giszter, 2010). However, any commands related to force are almost certain to reverse between forward and backward cycling due to the reversal of required force patterns. Thus, the dominant signals in the neural data are not readily explained in terms of either muscle-command encoding or kinematic encoding. Of course, this does not rule out the possibility that muscle-like commands (or kinematic commands) are encoded in dimensions beyond the top two PCs. Indeed, we will suggest below that muscle-like commands likely are encoded. Yet one

is tempted to question the assumption that the dominant signals encode commands of any sort. Might there exist an alternative explanation?

Smooth dynamics predict low trajectory tangling

Recent physiological and theoretical investigations suggest that the neural state in motor cortex obeys smooth dynamics (Churchland et al., 2012; Hall, de Carvalho, & Jackson, 2014; Michaels et al., 2016; Seely et al., 2016; Sussillo et al., 2015). Smooth dynamics imply that neural trajectories should not be 'tangled': similar neural states, either during different movements or at different times for the same movement, should not be associated with different derivatives. We quantified trajectory tangling using

$$Q(t) = \max_{t'} \frac{\|\dot{\boldsymbol{x}}_t - \dot{\boldsymbol{x}}_{t'}\|^2}{\|\boldsymbol{x}_t - \boldsymbol{x}_{t'}\|^2 + \varepsilon}$$

Equation 2.1

where $\boldsymbol{x}_t$ is the neural state at time t (*i.e.*, a vector containing the neural responses at that time), $\dot{\boldsymbol{x}}_t$ is the temporal derivative of the neural state, $\|\cdot\|$ is the Euclidean norm, and ε is a small constant that prevents division by zero (*Methods*). $Q(t)$ becomes high if there exists a state at a different time, t', that is similar but associated with a dissimilar derivative. We take the maximum to ask whether the state at time t ever becomes tangled with any other state. This maximum is taken with t indexing across time during all conditions. $Q(t)$ can be analogously assessed for the muscle trajectories.

We chose tangling as a straightforward measure of whether a given trajectory could have been produced by a smooth dynamical flow-field. Given limits on how non-smooth dynamics can be, moments of very high tangling are incompatible with a fixed flow-field. Furthermore, even

moderately high tangling implies potential instabilities in the underlying flow-field (**Figure 2.S1** and **Supplemental Note**). High tangling thus implies that the system must rely on external commands rather than internal dynamics, or that the system is flirting with instability. Although other metrics are possible, tangling has the practical benefit that it can be computed directly from the trajectories without needing to know (or fit) a flow-field.

For the reasons above, a network that relies heavily on intrinsic dynamics should avoid tangling. In contrast, when population activity primarily reflects external commands (as for the muscles or a population of sensory neurons) high tangling is both benign and, with enough observations, likely. For example, co-contraction of the *biceps* and *triceps* at one moment might need to be quickly followed by *biceps* activation and *triceps* relaxation. At a later moment or during a different movement, co-contraction might instead need to be followed by *biceps* relaxation and *triceps* activation. This would constitute an instance of tangling because the same state (co-contraction) is followed by different subsequent states. Do such moments of high tangling indeed occur for the muscles? If so, are they mirrored or avoided in the neural responses?

The state for a given time is a location on a state-space trajectory. The derivative is the direction in which the trajectory is headed. Two states are thus tangled if they are nearby but associated with different trajectory directions. For visualization, we consider a subset of the data: the middle five cycles of seven-cycle movements projected onto two dimensions (**Figure** 2.5**A,B**). Of course, two-dimensional projections only partially reflect the true population state; activity spans multiple dimensions. As a practical choice, we computed tangling in eight dimensions (results were robust with respect to this choice – see below). Muscle trajectories (**Figure** 2.5**A**) show three features suggestive of high tangling. First, muscle trajectories counter-rotate when cycling forward versus backward, yielding opposing derivatives for similar states. Second, muscle trajectories often

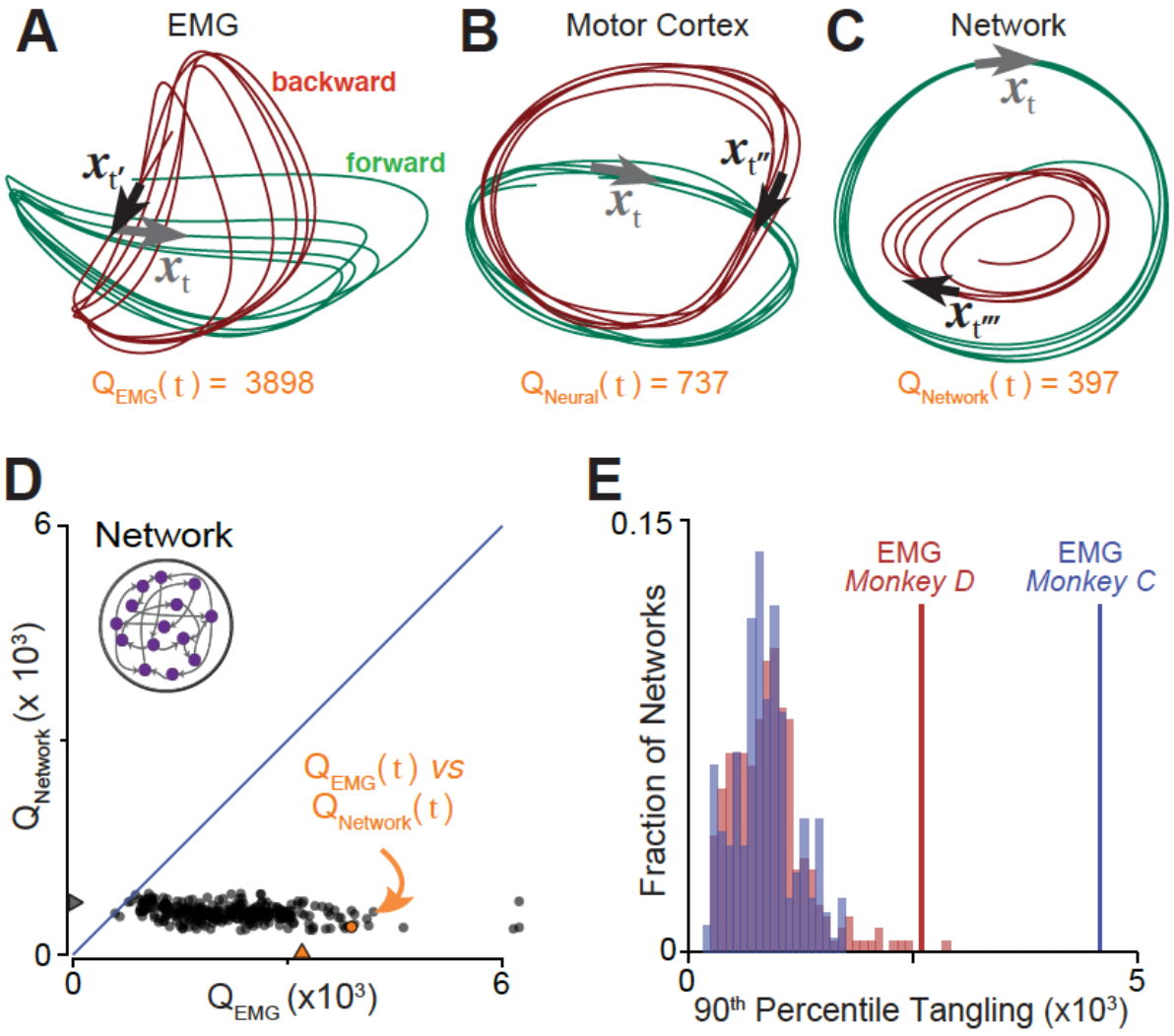

Figure 2.5 Illustration and validation of the trajectory tangling metric

A. Muscle trajectories during the middle five cycles for two conditions: seven-cycle / bottom-start / forward (*green*) and seven-cycle / bottom-start / backward (*red*). Arrows illustrate a pair of highly tangled states. Arrows point in the direction of the derivative (the path of the trajectory). Time t is a time that resulted in a high value of $Q_{EMG}(t)$. Time t' is the 'associated time' that resulted in that tangling value – *i.e.*, that maximizes $\frac{\|\dot{x}_t - \dot{x}_{t'}\|^2}{\|x_t - x_{t'}\|^2 + \varepsilon}$. In this example, time t' occurs during a different condition (forward rather than backward cycling). Tangling was computed in eight dimensions. **B.** Same as A but for neural trajectories. Time t is the same time as in A, and time t'' is the associated time used to compute $Q_{Neural}(t)$. **C.** Same but for network trajectories from an artificial recurrent network. The network was trained to produce the activity of all muscles for the times / conditions illustrated in A. **D.** Scatterplot, with one point per time / condition, of network-trajectory tangling versus muscle-trajectory tangling. *Orange arrow* denotes tangling for time t, corresponding to the time for which tangling was assessed in panels A and C. **E.** The consistency of the effect in panel D is demonstrated across 463 networks, each trained to produce the pattern of muscle activity from monkey D (*red*) or monkey C (*blue*). Tangling is summarized by the 90th percentile value (which highlights how high tangling can become). Lines denote 90th percentile tangling for the empirical muscle populations.

crossed themselves at right angles, resulting in similar states with very different derivatives. Third, non-circular trajectories sometimes cause create nearby muscle states moving in rather different directions. These features indeed lead to occasional moments of high tangling. For example, the *gray arrow* shows the muscle state and its derivative at a chosen time t. There exists another state, at time t', at a similar location in state-space but with a very different derivative (*black arrow*).

Neural trajectories (**Figure** 2.5**B**) appear potentially less tangled. Co-rotation prevents trajectories from continuously opposing one another between forward and backward cycling. Even within a condition, trajectories are closer to circular with fewer sharp bends. There are moments where trajectories cross in these two dimensions, but this did not result in high tangling because trajectories were separated in other dimensions. Notably, at moments when muscle trajectories became highly tangled, neural trajectories did not. For example, the muscle state at time t was strongly tangled while the neural state at that same time was much less tangled.

Before comparing tangling across all times/conditions, we wished to confirm that the tangling metric behaves as intended when the ground truth is known. We examined trajectories from a simulated recurrent neural network trained to produce muscle activity for the subset of data plotted in **Figure** 2.5**A**. The network output closely approximated those muscle signals, yet the dominant signals internal to the network did not (compare **Figure** 2.5**C** with **Figure** 2.5**A**). We plotted $Q_{\mathrm{Network}}(t)$ versus $Q_{\mathrm{EMG}}(t)$ for every time during both simulated conditions (**Figure** 2.5**D**). Network-trajectory tangling was consistently lower than muscle-trajectory tangling, despite producing muscle trajectories as an output. We repeated this analysis for multiple simulated networks, using different weight initializations and meta-parameters. Across multiple training initializations, the degree of network-trajectory tangling was variable (distributions in **Figure** 2.5**E**) but was nearly always lower than muscle-trajectory tangling.

Neural- versus muscle-trajectory tangling

For motor cortex, we compared Q_{Neural} and Q_{EMG} for all times across all twenty conditions. At least four results are possible. First, if motor cortex activity is a straightforward code for muscle activity, Q_{Neural} and Q_{EMG} should have a linear relationship with a slope near unity. Second, if motor cortex reflects unknown variables, and/or if tangling captures nothing fundamental, Q_{Neural} and Q_{EMG} may show no clear relationship. Third, if neural activity is more complex, intricate, or 'noisier' than muscle activity, Q_{Neural} could tend to be greater than Q_{EMG}. Finally, Q_{Neural} could be systematically reduced relative to Q_{EMG}, as for the simulated networks.

The data obeyed the final prediction (**Figure** 2.6**A,B**). The neural state was less tangled than the corresponding muscle state in 99.9% and 96.6% of cases (monkey D and C). The rare exceptions occurred when tangling was low for both. Strikingly, muscle-trajectory tangling could be quite high with no accompanying increase in neural-trajectory tangling. Statistically, distributions of Q_{Neural} and Q_{EMG} were indeed different (paired t-test, $p<10^{-10}$ for each monkey). The difference in tangling was robust to analysis choices: it did not depend on the use of PCA versus 'raw' data (**Figure 2.S2**), on the number of PCs analyzed (**Figure 2.S3**), on whether we matched dimensionality or variance explained (**Figure 2.S3**), or on the relative number of neurons versus muscles (**Figure 2.S3**). The large difference between Q_{Neural} and Q_{EMG} contrasts with the fact that visual inspection does not readily reveal whether individual recordings are neural or muscular (compare **Figure** 2.3 with **Figure** 2.2). Yet the tangling metric readily distinguished between even small populations of neurons versus muscles (**Figure** 2.6**C**).

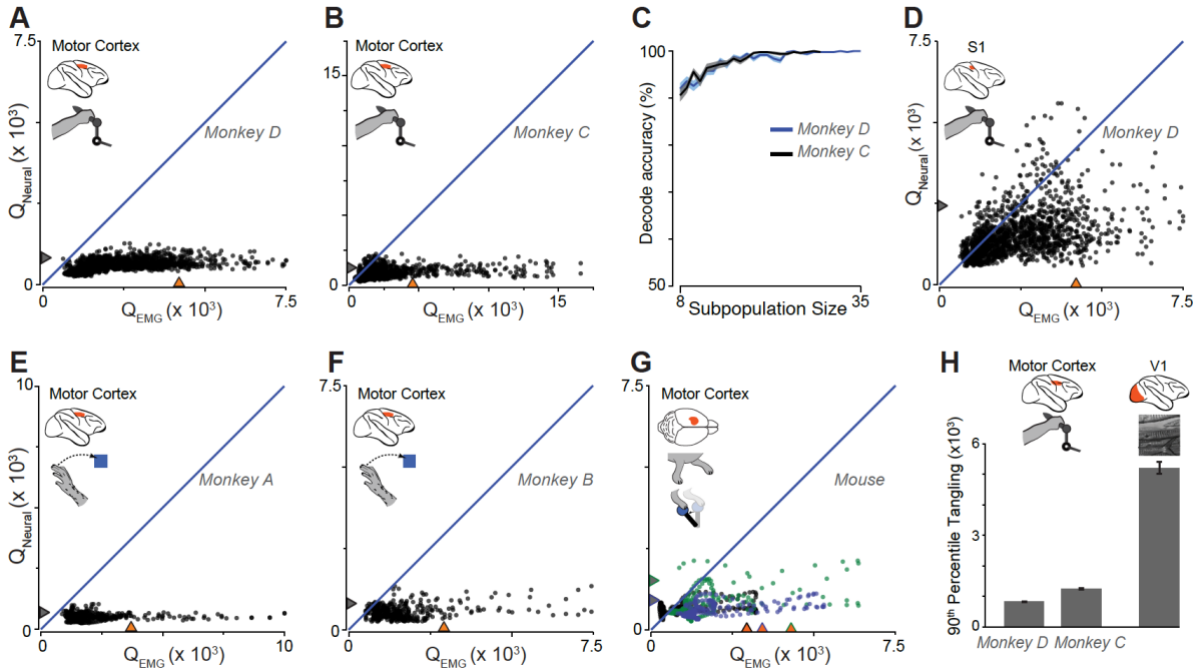

Figure 2.6 Trajectory tangling for multiple datasets

A. Scatterplot of motor-cortex-trajectory tangling versus muscle-trajectory tangling (monkey D). Each point shows tangling for one moment (one time during one condition). Points are shown for all times during movement (sampled every 25 ms) for all twenty conditions. *Blue line* indicates unity slope. *Gray / orange triangles* indicate 90th percentile tangling. **B.** Same as A but for monkey C. **C.** Neural versus muscle populations could be distinguished based on tangling. For a given number of recordings, we drew that many neurons and muscles and computed tangling for each subpopulation. 500 such draws were made for each subpopulation size. The vertical axis gives the percentage of instances where the neural subpopulation was correctly identified based on lower tangling. Flanking standard errors are based on binomial statistics. **D.** S1 neural-trajectory tangling versus muscle-trajectory tangling (monkey D). **E.** Motor-cortex-trajectory tangling versus muscle-trajectory tangling during reaching (monkey A). Each point corresponds to one time during one of eight conditions. **F.** Same as E but for monkey B. **G.** Scatterplot of motor-cortex-trajectory tangling versus muscle-trajectory tangling in three mice (*black*, *blue*, and *green* symbols) during both locomotion and lever pulling. Illustration in inset by E. Daubert. **H**. Comparison of motor-cortex-trajectory tangling and visual-cortex-trajectory tangling. Because V1 data contains no corresponding muscle activity, tangling is quantified by the 90th percentile values. Motor cortex data are from the cycling task as in panels A and B. V1 data were recorded using natural scenes. Error bars show the standard error computed via bootstrap: the distribution of tangling values was resampled 200 times, and we computed the sampling distribution of the 90th percentile values

Tangling across tasks, species, and areas

Is low neural- versus muscle-trajectory tangling specific to cycling or a more general property of motor cortex? We leveraged recently collected data (Elsayed et al., 2016) from two monkeys performing a center-out reach task. The same result was observed: Q_{Neural} was greatly reduced relative to Q_{EMG} (**Figure** 2.6**E,F**). We also compared Q_{Neural} and Q_{EMG} in mice during an experiment with two behaviors: reaching to pull a joystick and walking on a treadmill (Miri et al., 2017). We observed a slightly weaker yet similar effect (**Figure** 2.6**G**) to that seen in primates. Thus, low trajectory tangling in motor cortex appears to be a general property.

We also examined responses in the proprioceptive region (area 3a) of primary somatosensory cortex (S1) during cycling. This region is immediately adjacent to motor cortex, and individual-neuron responses are surprisingly similar to those in motor cortex (**Figure 2.S4**). Yet tangling was not as consistently low in S1 (**Figure** 2.6**D**) as it was in motor cortex (**Figure** 2.6**A**, same task and monkey). At moments where the muscle state became highly tangled, the S1 state often also became quite tangled. All three tangling distributions were significantly different: $p < 10^{-10}$ when comparing muscle and S1 populations; $p < 10^{-10}$ when comparing S1 and motor cortex populations (paired t-test).

We also considered a primary visual cortex (V1) population responding to natural-scene movies. V1 trajectories were much more tangled than motor cortex trajectories (**Figure** 2.6**H**; $p<10^{-10}$ and $p<10^{-10}$, two-sample t-test comparing V1 with motor cortex for monkey D and C). Across datasets (motor cortex, muscle, S1, V1) there was no clear relationship between dimensionality and tangling (**Figure 2.S5**). Instead, tangling was highest the muscles, and for cortical areas where sensory input is expected to have the largest impact. This is consistent with the fact that sensory

input (unless it can be predicted from outgoing commands) can readily cause the same state to be followed by different future states (*e.g.*, no constraint prevents image A from being followed by image B on one occasion, and by image C on another occasion).

Noise-robust networks display low tangling

For a recurrent or feedback-driven network, it is intuitive that high tangling must be avoided. If the flow-field has some degree of smoothness, nearby states cannot be associated with very different derivatives. Thus, moments of high tangling cannot be produced without relying on disambiguating external inputs. Yet motor cortex trajectories avoided even moderate tangling. This is not strictly necessary even in the idealized case of a fully autonomous dynamical system. For example, some recurrent networks did show moderate tangling (right tail of the distribution in **Figure** 2.5**E**) yet still functioned. Might the very low empirical tangling confer some computational advantage? Formal considerations support that possibility: even moderate tangling implies potential dynamical instabilities (**Supplemental Note**).

To explore potential advantages of low tangling, we considered neural networks trained to generate a simple idealized output: $\cos t$ for one muscle and $\sin 2t$ for a second muscle (**Figure** 2.7**A**, *top*). The resulting output trajectory was thus a figure-eight (*left* sub-panel). It is not possible for a network's internal trajectory to follow a pure figure-eight; the center-most state is very highly tangled. Tangling can be reduced by employing a third dimension such that the trajectory is: $[\cos t\,;\ \sin 2t\,;\ \beta \sin t]$. Even a modest value of β reduces tangling enough (*middle* sub-panel) that the trajectory can be produced. As a network follows that three-dimensional trajectory, the figure-eight trajectory can still be 'read out' via projection onto two of the axes (with the third dimension falling in the null space of the readout (Druckmann & Chklovskii, 2012; Kaufman et al., 2014)).

Is there an advantage to further decreases in tangling (*right* sub-panel)? We examined noise tolerance across networks whose internal trajectories were $[\cos t\,;\ \sin 2t\,;\ \beta \sin t]$ with different values of β. This necessitated the unusual step of training networks not only to produce a desired output, but also to follow a specified internal trajectory (*Methods*).

Networks with high trajectory tangling failed to produce the figure-eight output trajectory in the presence of even small amounts of noise (**Figure** 2.7**B**). Networks with low trajectory tangling were much more noise robust. We performed a similar analysis with trajectories that encoded the empirical muscle trajectories, but with varying degrees of tangling (found using the optimization approach in the next section). Again, low tangling provided noise robustness (**Figure 2.S6**). This was true both for networks that generated a single internal trajectory, and networks that generated different 'forward' and 'backward' trajectories based on inputs. Intuitively, when tangling is low it is less likely that noise will perturb the network onto a nearby but inappropriate part of the trajectory. More formally, low tangling aids local stability (**Figure 2.S1; Supplemental Note**).

While the example in **Figure** 2.7**A,B** is intentionally simplified, it illustrates a feature that may help interpret the empirical neural trajectories. Note that $\beta = 1$ yields a weakly-tangled trajectory that encodes the desired figure-eight output in one projection and is a circle in another projection (**Figure** 2.7**A**, *right* sub-panel). Although we created this shape via construction, it is a natural shape to introduce: a circle is the least-tangled rhythmic trajectory.

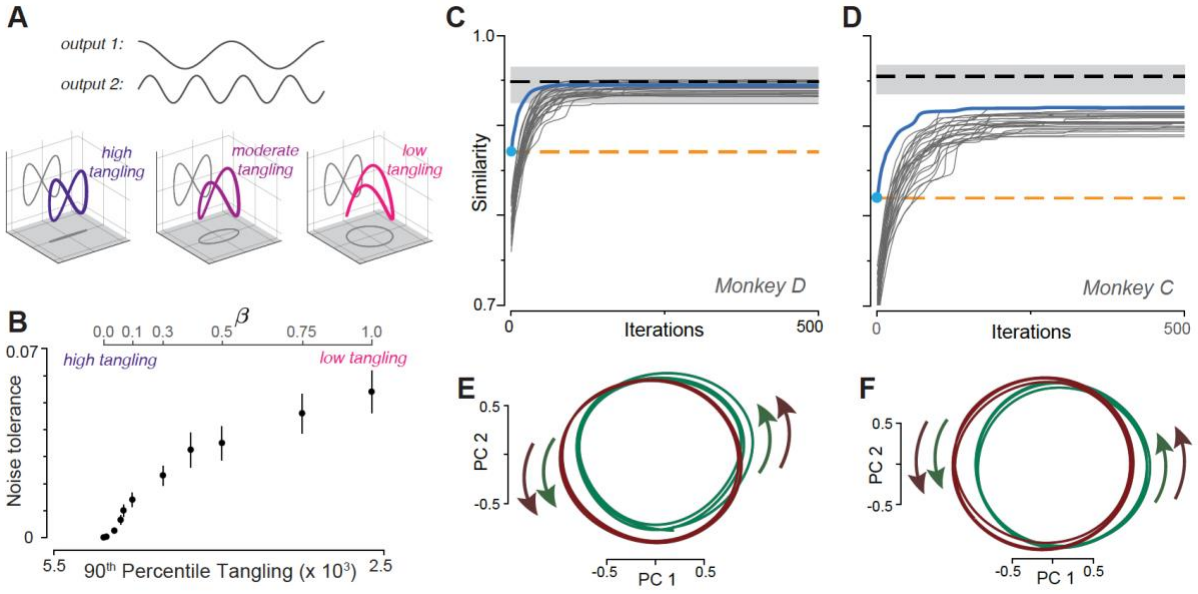

Figure 2.7 Leveraging the observation of low trajectory tangling to predict the neural population response.

A. Illustration of how the same output can be embedded in a larger trajectory with varying degrees of tangling. *Top gray traces*: A hypothetical desired two-dimensional output $[\cos t\,;\ \sin 2t]$. Plotted in state space, the output trajectory is a figure-eight, and contains a central point that is maximally tangled. Adding a third dimension ($\beta \sin t$) reduces tangling at that central point. The figure-eight can still be decoded via projection onto two dimensions, in which case the third dimension falls in the null-space of the decode. **B.** Noise robustness of recurrent networks trained to follow the internal trajectory $[\cos t\,;\ \sin 2t\,;\ \beta \sin t]$. By varying β, we trained a set networks that could all produce the same figure-eight output, but had varying degrees of trajectory tangling. For each value of β we trained 20 networks, each with a different random weight initialization. Noise tolerance was the largest magnitude of state noise for which the network still produced the figure-eight output. Plotted are the mean and SEM of the noise tolerance versus network tangling for each value of β. Note that the x-axis has been flipped such that tangling decreases from left to right. **C.** Similarity of the predicted and empirical motor-cortex population responses (monkey D). *Blue trace*: prediction yielded by optimizing the cost function in Equation 2. Optimization was initialized with the empirical muscle trajectories. *Cyan dot* indicates similarity at initialization, which is simply the similarity of empirical neural and muscle trajectories. *Gray traces:* Same as blue trace but initialized with Gaussian noise added to muscle trajectories. Multiple initializations were tested yielding a family of predictions. *Black dashed line* shows upper benchmark as described in the text. *Gray shading* indicates 95% confidence interval on the upper benchmark computed across multiple random divisions of the population. *Orange dashed line* shows a lower benchmark: similarity of muscle and neural data (this necessarily intersects the *cyan dot*). **D.** Same but for monkey C. **E.** Projection of a representative predicted population response (after optimization was complete) onto the top two principal components. Data are for monkey D. *Green* / *red* traces show trajectories for three cycles of forward / backward cycling respectively. **F.** Same but for monkey C.

Hypothesis-based prediction of neural responses

The results above suggest a hypothesis: motor cortex may embed outgoing commands (which, if muscle-like, would be quite tangled) in a larger trajectory such that the full orbit is minimally tangled. Inspired by optimizations that successfully predicted V1 responses (Olshausen & Field, 1996), we employed an optimization approach to predict the dominant patterns of motor cortex activity. Optimization found a predicted neural population response, $\hat{X}$, that could be linearly decoded to produce the empirical muscle activity Z, yet was minimally tangled. Specifically:

$$\hat{X} = \underset{X}{\operatorname{argmin}} \left(\left\| Z - ZX^{\dagger}X \right\|_F^2 + \lambda \sum_t Q_X(t) \right)$$

Equation 2.2

where each column of the matrix Z describes the muscle population response for one time and condition. The first term of the cost function ensures that neural activity 'encodes' muscle activity; $ZX^{\dagger}X$ is the optimal linear reconstruction of Z from X († indicates the pseudo-inverse; $\|\cdot\|_F$ indicates the Frobenius norm). This formulation should not be taken to imply that the true neural-to-muscle mapping is linear, merely that the predicted neural activity should yield a reasonable linear readout of muscle activity, consistent with empirical findings (Griffin et al., 2008; Morrow et al., 2009; Schieber & Rivlis, 2007). The second term of the cost function encourages low trajectory tangling. The predicted neural population response thus balances optimal encoding of muscle activity with minimal tangling.

We applied optimization using muscle data that included three middle cycles of forward cycling and three middle cycles of backward cycling. Thus, we are attempting to simultaneously predict two 'steady state' neural trajectories. We used canonical correlation to assess the similarity

between predicted and actual neural responses. Canonical correlation finds linear transformations of two datasets such that they are maximally correlated. We employed a variant of canonical correlation that enforces orthonormal matrix transformations. Unity similarity thus indicates two datasets are the same but for a rotation, isotropic scaling, or offset. We initialized optimization with $\hat{X}_{init} = Z$, corresponding to the baseline hypothesis that neural activity is a 'pure' code for muscle activity. This resulted in a reasonably high initial similarity (**Figure** 2.7**C,D**, *cyan dot*) because muscle activity shares many basic features with neural activity (*e.g.*, the same fundamental frequency).

During optimization, we insisted that the predicted neural population response, $\hat{X}$, have the same dimensionality as the muscle population response, Z (both were ten-dimensional). Matching dimensionality is a conservative choice that aids interpretation. Because optimization cannot add dimensions, some muscle-like features must be lost in order to gain features that reduce tangling. Similarity will therefore increase only if the features gained during optimization are more realistic / prominent than the features that are lost.

Similarity between predicted and empirical populations increased with optimization (**Figure** 2.7**C,D** *blue*), reaching a similarity roughly halfway between the 'pure muscle encoding' hypothesis and

perfect similarity. To provide a rough benchmark of good similarity, we computed the average similarity between two random halves of the empirical neural population (*black dashed trace* with 95% confidence intervals). Similarity approached this benchmark for both monkeys. To test the consistency of this result we repeated optimization, each time initializing with the empirical patterns of muscle activity plus temporally smooth noise in each of the ten dimensions. Similarity to the data always increased (*gray traces*). This analysis also revealed that the addition of random

structure decreased initial similarity (*gray traces* start below the *blue trace*). This underscores that increasing similarity requires the addition of structure matching that in the neural data, rather than any arbitrary structure.

Each initialization resulted in a slightly different solution (the optimized $\hat{X}$). We were thus able to ask which solutions were common and whether the nature of those solutions explains the increased similarity with the empirical data. For all 200 solutions (100 per monkey), optimization produced near-circular trajectories. When comparing between forward and backward, two classes of solution emerged. The less common (31/100 for monkey D and 13/100 for monkey C) involved dominant circular trajectories in planes that were nearly orthogonal (first principal angle > 85°) for forward and backward. The most common (69/100 and 87/100 for monkey D and C) involved at least some overlap between these planes. In such cases, trajectories were almost always co-rotational (67/69 and 85/87 for monkey D and C) in the top two PCs. Two typical solutions are shown in **Figure** 2.7**E,F**. Co-rotations dominate because, when two trajectories exist in a common subspace, tangling is lowest if they co-rotate (if they exist in orthogonal planes, co-rotation versus counter-rotation is not defined). Similar structure was seen for the empirical data: the planes that best captured neural trajectories during forward and backward cycling overlapped (principal angles were 72° and 61° for monkey D, and 73° and 40° for monkey C) and showed co-rotation in the top two PCs (as in **Figure** 2.4**E,H**). Thus, the hypothesis embodied in **Equation** 2.2 not only increased quantitative similarity, it also reproduced the dominant features of the neural data: nearly circular trajectories that exist in distinct but overlapping planes, and that co-rotate in the projection capturing the most variance.

Alternative predictions

We performed a variety of optimizations corresponding to cost functions embodying other hypotheses (**Figure 2.S7**). Optimizations that sought to reduce the norm of activity or to increase sparseness (standard forms of regularization) led to decreases in similarity. Optimizing for local smoothness (one aspect of low tangling) increased similarity but not as much as optimizing for low tangling itself. Thus, similarity increased only when optimization reduced tangling, and increased most when low tangling was directly optimized.

However, low tangling *per se* was not necessarily sufficient to increase similarity. We created simulated populations where the response of each unit was either the response of a muscle or the derivative of that response. This reflects the hypothesis that neurons might represent both muscle activity and the change in muscle activity (Evarts, 1968). By construction, these simulated populations had fairly low tangling (**Figure 2.S8A**). Yet, they did not particularly resemble the neural population. Quantitatively, similarity increased modestly for monkey D (roughly half as much as when optimizing for low tangling directly) and decreased for monkey C. The dominant signals in these simulated populations also did not show the same dominant circular structure seen in the neural data (**Figure 2.S8B**). The mismatch can be understood by noting that differentiation increases the prevalence of high-frequency features. This does not lead to a match with the dominant circular structure at the fundamental frequency in the empirical data. In summary, optimizing directly for low tangling introduced features that were both particularly effective in reducing tangling and matched features in the data. Reducing tangling in a more 'incidental' fashion did not produce these realistic features.

Signals introduced by optimization yield incidental correlations

The optimization based on **Equation** 2.2 added structure that reduced tangling. That structure is unconnected to kinematics or other task parameters; optimization was blind to all such parameters. Nevertheless, the predicted neural population response appeared to encode kinematics to a greater degree than would a pure code for muscle activity. We used linear regression to decode a set of kinematic parameters (horizontal and vertical position and velocity) from the activity of the muscle population. Fits were reasonable (R^2=0.86 and 0.88 for monkey D and C) but improved (R^2=0.97 and 0.94) when we instead decoded kinematics from the predicted neural population response. This performance was nearly identical to that observed when decoding kinematics from the empirical neural population (R^2=0.98 and 0.93). The ability to decode horizontal and vertical velocity might initially seem surprising: the dominant signals in the neural data co-rotated in the top two PCs – inconsistent with a velocity representation. However, the presence of more than two dimensions with sinusoidal structure ensured that velocity could be read out reasonably accurately. Despite these excellent decodes, generalization performance was poor: generalization R^2 was near-zero (or even negative) when fitting kinematics for one direction and predicting for the other. This was true whether decoding was based on the predicted or empirical neural response. While poor generalization does not exclude the possibility that the empirical population encodes kinematic signals, we saw no direct evidence for this hypothesis. As noted above, we also rarely observed neurons whose firing rates resembled kinematic parameters.

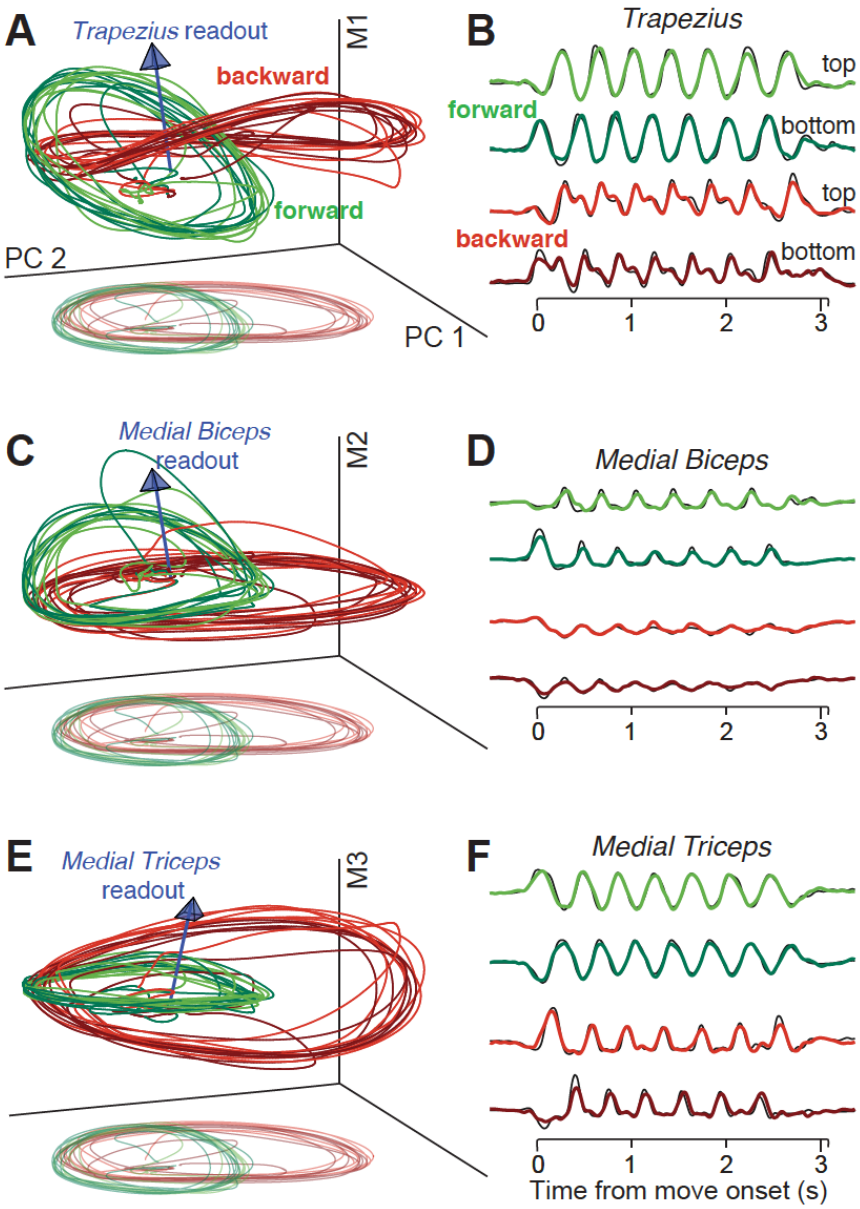

Figure 2.8 Muscle-like signals coexist with signals that contribute to low tangling.

Data are for monkey D. **A.** Three-dimensional subspace capturing trajectories that encode *trapezius* activity; *i.e.*, can be linearly read out to approximate *trapezius* activity. *Blue arrow* indicates the readout direction, defined by the weights identified via linear regression. Axes correspond to the first two PCs and a third dimension that ensures the space spans the readout direction. Trajectories are shown for four conditions: forward (*green*) and backward (*red*) seven-cycle movements, starting at the top and bottom (*lighter* and *darker traces*). *Lighter 'shadow' traces* at bottom show the projection onto just the first two PCs (perspective has been added). **B.** Projections, for the four conditions plotted in A, onto the readout direction. *Thin black trace* plots the true activity of the *trapezius*. Axis spans the time of movement. **C,D.** Same as A,B but for the *medial biceps*. Only the third (vertical) axis is different. **E,F.** Same but for the *medial triceps*.

Muscle-like signals are embedded in trajectories with low tangling

The optimization results lead to the hypothesis that the dominant population-level signals in motor cortex function to yield low tangling, and that muscle-like signals may be encoded by relatively modest 'ripples' in dimensions that point off the plane of dominant circular structure. A rough analogy would be a phonograph, where the direction that encodes a temporally complex output is orthogonal to the dominant motion of the record. Can such structure be viewed directly in the empirical data? We projected the neural population response onto triplets of dimensions (**Figure** 2.8). The first and second dimensions were always the first two PCs. The third was based on the readout direction of a particular muscle, defined by the set of weights found via linear regression (*arrow* in **Figure** 2.8**A** plots the readout direction for the *trapezius*). The third dimension was then the vector that was orthogonal to the first two PCs, and allowed the three dimensions to span the readout direction.

Consider first a triplet of dimensions that span the *trapezius* readout direction (**Figure** 2.8**A**). Trajectories trace out circular paths in the top PCs. Ripples in a third dimension yield the fine temporal structure that matches *trapezius* activity **Figure** 2.8**B**). The overall trajectory thus has the joint properties of encoding *trapezius* activity while exhibiting low tangling. Similar structure was observed for other muscles (**Figure** 2.8**C,E**, **Figure 2.S9** shows data for monkey C).

The dimensions that encode muscle activity captured only modest variance. In the examples in **Figure** 2.8, each muscle-readout dimension captured ~10% as much variance as each of the top two PCs. The vertical dimensions in **Figure** 2.8**A,C,E** are thus shown on an expanded scale for visualization. Similar structure was present for the network model in **Figure** 2.5**C** and also for the predicted population responses in **Figure** 2.7**E,F**: the activity of each 'encoded' muscle constituted a set of ripples upon dominant circular structure that yielded low tangling.

In addition to the dimensions from which muscle-like signals can be read out, there exist other dimensions (not visible in **Figure** 2.8) that provide separation between neural trajectories during forward and backward cycling. Low tangling may require such separation, else forward and backward trajectories would have to encode very different patterns of muscle activity despite following similar paths. Indeed, forward and backward neural trajectories were on average much better separated than the corresponding muscle trajectories (**Figure 2.S10**). This difference in separation was large but not as profound as the difference in tangling. Thus, low neural-trajectory tangling (relative to muscle-trajectory tangling) results from a variety of factors: more circular trajectories, increased separation between forward and backward trajectories, and greater alignment of flow-fields (*e.g.*, co-rotation in the dominant dimensions).

Tangling in sulcal motor cortex

The results above support the hypothesis that population activity in motor cortex is less tangled than the outputs of that population. If so, tangling might be predicted to be moderately higher in sulcal motor cortex, where some neurons (cortico-motoneurons) make mono-synaptic connections onto motor neurons (Rathelot & Strick, 2009), and signals related to outgoing muscle-like commands may be enriched. This is worth investigating both as an additional test of the hypothesis and because our measurements of muscle activity are only a proxy for the output of motor cortex. Ideally, we would be able to compute tangling for a subpopulation of identified cortico-motoneurons. In the absence of such recordings, we considered the subpopulation of sulcal recordings as a whole, and compare with a subpopulation from the most anterior region from which we recorded: the aspect of dorsal premotor cortex contiguous with surface primary motor cortex. Cortico-motoneurons are largely absent from this anterior region (Rathelot & Strick, 2006). The

subpopulation of sulcal neurons did indeed show modestly but significantly higher tangling during both cycling and reaching (**Figure 2.S11**).

Discussion

Are the dominant signals in motor cortex representational or computational?

We found that the dominant signals in motor cortex were not muscle-like. This result echoes findings during reaching, where aspects of neural responses depart from expectations under a muscle-encoding framework (Evarts, 1968; Heming et al., 2016; Kakei et al., 1999; Moran & Schwartz, 1999b; Scott, 1997, 2008; Todorov, 2000). The dominance of non-muscle-like signals is more patent during cycling; non-muscle-like signals are apparent simply via inspection of projections onto the top PCs.

A traditional explanation for non-muscle-like signals is that they represent higher-level movement parameters. The present results are inconsistent with the most common proposal: a representation of direction or velocity. Under that proposal, trajectories should have been co-planar and counter-rotated between forward and backward cycling. We also found that single-neuron responses rarely resembled velocity profiles. Our data do not rule out the possibility that neural activity encodes yet-to-be-determined set of kinematic parameters (perhaps in addition to muscle-like signals). However, our results urge caution when considering such hypotheses. For example, reducing tangling via optimization increases the degree to which activity appears (incorrectly) to represent kinematic parameters. More broadly, it may often be possible *post hoc* to select kinematic parameters that resemble the neural dominant signals. As one example, a representation of horizontal position and velocity would produce ellipses that co-rotate during forward / backward cycling. However, this 'horizontal kinematics' hypothesis would require a high relative position sensitivity to ensure a circular trajectory. A high position sensitivity is inconsistent with

observations during reaching, where correlations are strongest with reach velocity and direction (Ashe & Georgopoulos, 1994). In summary, in this study as in others, there will always be correlations that are incidental rather than fundamental (Churchland & Shenoy, 2007; Fetz, 1992; Mussa-Ivaldi, 1988; Reimer & Hatsopoulos, 2009; Todorov, 2000). While it remains possible that kinematic parameters are represented, we saw no compelling evidence for this idea. The dominant signals were already naturally explained by the hypothesis that tangling should be minimized.

Our results thus suggest that the dominant signals in cortex may play a computational rather than a representational function. Specifically, the dominant signals may fall partly or largely in the null-space of communication with downstream structures, yet may be critical for ensuring reliable generation of the commands that are communicated. Put differently, motor cortex is part of a larger dynamical system (spanning many areas, including the spinal cord, and incorporating sensory feedback) that culminates in the generation of muscle commands. Such a system as a whole is likely to contain non-output signals. It does logically follow that motor cortex itself must show either non-output signals or low tangling; motor cortex could be downstream of the relevant dynamics or reflect only a small part of the overall network state. Yet empirically, motor cortex displayed very low tangling.

Differences and commonalities across tasks

During both cycling and reaching (Churchland et al., 2012) neural trajectories follow circular paths that rotate in a concordant direction, a feature not seen in the muscle population during either task. This shared feature may reflect the combination of two facts. First, a circle is the least-tangled rhythmic trajectory. Second, muscle activity during both tasks involves rhythmic aspects. This is trivially true during cycling. It is more subtly true during reaching, where multiphasic patterns of

muscle activity are readily constructed from a quasi-oscillatory basis (Churchland & Cunningham, 2014; Churchland et al., 2012). Rotational trajectories are thus a natural way of encoding muscle activity while maintaining low tangling. This interpretation agrees with the recent finding that a network model, trained to produce muscle activity during reaching, reproduced the rotational neural trajectories (Sussillo et al., 2015). This occurred only if the network was regularized to encourage smooth dynamics, a regularization which would implicitly encourage low tangling.

Yet we stress that rotational structure *per se* is unlikely to be the fundamental principle shared across tasks. There are many ways of adding structure that can reduce tangling. Even if certain motifs are common, the optimal way to reduce tangling will be task-dependent. Thus, we propose that the deeper connection across tasks will not be a specific form of dynamics, but dynamics that yield low tangling.

We also note that different tasks may involve motor cortex sending different classes of output commands. For some tasks, the details of muscle activity may be largely determined by spinal circuitry, while other tasks (especially learned or dexterous tasks) may require more direct control of the musculature. The latter is potentially true during cycling, and some of our analyses thus assumed a roughly linear relationship between neural and muscle activity. However, the hypothesized computational principle – embed outgoing commands in structure that minimizes tangling – would apply even if commands were only somewhat muscle-like (*e.g.*, if they were transformed considerably by the spinal cord). Indeed, it would apply even if descending commands are high-level, as may have been the case in mice during locomotion.

Tangling across areas

Trajectory tangling was very low for motor cortex, considerably higher for S1, and higher still for the muscles. Tangling was also high for V1. The degree of tangling may depend on how fully activity in that area reflects global dynamics. Motor cortex may show particularly low tangling because it processes many relevant sources of information. It is not only a major output of the primate motor system, but responds robustly and rapidly to sensory inputs (Herter et al., 2009) and lies at the nexus of cerebellar and basal-ganglia feedback loops (Middleton & Strick, 2000). Other areas, even those that participate in the same task, may or may not exhibit low tangling depending on how fully they reflect the overall network state. In particular, S1 responses are likely dominated by sensory feedback and may very incompletely reflect the broader dynamics of motor control. Even within motor cortex, tangling was modestly higher within the sulcus, where activity may be more dominated by output commands. Although V1 presumably does exhibit some dynamics, activity is likely dominated by visual inputs which can produce high tangling. These comparisons echo our recent finding that population structure can be fundamentally different depending on whether an area is hypothesized to primarily reflect population dynamics versus external variables (Seely et al., 2016).

Might tangling differ within a population, even for the same task? Might the motor system, over the course of learning or development, adopt network trajectories that are increasingly less tangled? When a new skill is learned, is performance better if subjects achieve lower tangling? Are pathological conditions associated with increased tangling? Such questions illustrate that many aspects of motor cortex activity may be best understood not in terms of representations of external parameters, but in terms of the computational strategies that allow outputs to be accurately and reliably generated.

Methods

Experimental apparatus

Subjects were two adult male rhesus macaques (monkeys D and C). Animal protocols were approved by the Columbia University Institutional Animal Care and Use Committee. Experiments were controlled and data collected under computer control (Speedgoat Real-time Target Machine). During experiments, monkeys sat in a customized chair with the head restrained via a surgical implant. Stimuli were displayed on a monitor in front of the monkey. A tube dispensed juice rewards. The left arm was loosely restrained using a tube and a cloth sling. With their right arm, monkeys manipulated a pedal-like device. The device consisted of a cylindrical rotating grip (the pedal), attached to a crank-arm, which rotated upon a main axel. That axel was connected to a motor and a rotary encoder that reported angular position with 1/8000 cycle precision. In real time, information about angular position and its derivatives was used to provide virtual mass and viscosity, with the desired forces delivered by the motor. The delay between encoder measurement and force production was 1 ms.

Horizontal and vertical hand position were computed based on angular position and the length of the crank-arm (64 mm). To minimize extraneous movement, the right wrist rested in a brace attached to the hand pedal. The motion of the pedal was thus almost entirely driven by the shoulder and elbow, with the wrist moving only slightly to maintain a comfortable posture. Wrist movements were monitored via two reflective spheres attached to the brace, which were tracked optically (Polaris system; Northern Digital, Waterloo, Ontario, Canada) and used to calculate wrist angle. The small wrist movements were highly stereotyped across cycles. Visual monitoring (via

infrared camera) confirmed the same was true of the arm as a whole (*e.g.*, the lateral position of the elbow was quite stereotyped across revolutions). Eye position and pupil dilation were monitored but are not analyzed here.

Task

The monitor displayed a virtual landscape, generated by the Unity engine (Unity Technologies, San Francisco). Surface texture and landmarks to each side provided visual cues regarding movement through the landscape. Movement was along a linear 'track'. One rotation of the pedal produced one arbitrary unit of movement. Targets on the landscape surface indicated where the monkey should stop for juice reward.

Each trial of the task began with the appearance of an initial target. To begin the trial, the monkey had to cycle to and to acquire the initial target (*i.e.*, stop on it and remain stationary) within 5 seconds. Acquisition of the initial target yielded a small reward. After a 1000 ms hold period, the final target appeared at a prescribed distance. Following a randomized (500-1000 ms) delay period, a go-cue (brightening of the final target) was given. The monkey then had to cycle to acquire the final target. After remaining stationary in the final target for 1500 ms, the monkey received a large reward.

Successfully completing a trial necessitated satisfying a variety of constraints. Cycling had to begin between within 650 ms after the go cue. Once cycling began, the final target had to be reached within a distance-dependent time limit. The trial was aborted if this time elapsed (<0.01% of trials for both monkeys), or if cycling speed dropped below a threshold before entering the final target (~1.5% of trials in monkey D and ~1.7% in monkey C). The trial was also aborted if the monkey moved past the final target (~1.5% / 0.6% of trials), or if the monkey acquired the final

target and then moved while waiting for the reward (~0.6% / 0.3%). These constraints, combined with the monkeys' natural desire to receive reward quickly, produced movements that were both brisk and quite consistent across trials. The primary difference in behavior across trials was modest variation in overall movement duration (as illustrated in **Figure** 2.1). In rare cases, behavior on a successful trial differed notably from typical behavior for that condition. Such trials were removed prior to analysis.

The task included 20 conditions distinguishable by final target distance (half-, one-, two-, four-, and seven-cycles), initial starting position (top or bottom of the cycle), and cycling direction. Salient visual cues (landscape color) indicated whether cycling must be 'forward' (the hand moved away from the body at the top of the cycle) or 'backward' (the hand moved toward from the body at the top of the cycle) to produce forward virtual progress. Trials were blocked into forward and backward cycling. Other trials types were interleaved using a block-randomized design. We collected a median of 15 trials / condition for both monkeys

Neural recordings during cycling

After initial training, we performed a sterile surgery during which monkeys were implanted with a head restraint and recording cylinders. Cylinders (Crist Instruments, Hagerstown, MD) were placed surface normal to the cortex, centered over the border between caudal PMd and primary motor cortex, located according to a previous magnetic resonance imaging scan. The skull within the cylinder was left intact and covered with a thin layer of dental acrylic. Electrodes were introduced through small (3.5 mm diameter) burr holes drilled by hand through the acrylic and skull, under ketamine / xylazine anesthesia. Neural recordings were made using conventional

single electrodes (Frederick Haer Company, Bowdoinham, ME) driven by a hydraulic microdrive (David Kopf Instruments, Tujunga, CA).

Sequential recording with conventional electrodes (as opposed to simultaneous recording with an array) allowed us to acquire recordings from a broader range of sites, including sulcal sites inaccessible to many array techniques. Recording locations were guided via microstimulation, light touch, and muscle palpation protocols to confirm the trademark properties of each region. For motor cortex, recordings were made from primary motor cortex (both surface and sulcal) and the adjacent (caudal) aspect of dorsal premotor cortex. For most analyses, these recordings are analyzed together as a single motor cortex population (although see **Figure 2.S11**). Motor cortex recordings were restricted to regions where microstimulation elicited responses in shoulder, upper arm, chest and forearm. For one monkey, we also recorded from area 3a (proprioceptive primary motor cortex). These recordings (44 neurons) were made from the deeper aspects of the posterior bank of the central sulcus, where microstimulation did not produce movement.

Neural signals were amplified, filtered, and manually sorted using Blackrock Microsystems hardware (Digital Hub and 128-channel Neural Signal Processor). A total of 277 isolations were made across the two monkeys. Nearly all neurons that could be isolated in motor cortex were responsive during cycling. A modest number (21) of isolations were discarded due to low signal-to-noise ratios or insufficient trial counts. No further selection criteria were applied. On each trial, the spikes of the recorded neuron were filtered with a Gaussian (25 ms standard deviation; SD) to produce an estimate of firing rate versus time. These were then averaged across trials as described below.

EMG recordings

Intra-muscular EMG was recorded from the major muscles of the arm, shoulder, and chest using percutaneous pairs of hook-wire electrodes (30mm x 27 gauge, Natus Neurology) inserted ~1 cm into the belly of the muscle for the duration of single recording sessions. Electrode voltages were amplified, bandpass filtered (10-500 Hz) and digitized at 1000 Hz. To ensure that recordings were of high quality, signals were visualized on an oscilloscope throughout the duration of the recording session. Recordings were aborted if they contained significant movement artifact or weak signal. That muscle was then re-recorded later. Offline, EMG records were high-pass filtered at 40 Hz and rectified. Finally, EMG records were smoothed with a Gaussian (25 ms SD, same as neural data) and trial averaged (see below). Recordings were made from the following muscles: the three heads of the *deltoid*, the two heads of the *biceps brachii*, the three heads of the *triceps brachii*, *trapezius*, *latissimus dorsi*, *pectoralis*, *brachioradialis*, *extensor carpi ulnaris*, *extensor carpi radialis*, *flexor carpi ulnaris*, *flexor carpi radialis*, and *pronator*. Recordings were made from 1-8 muscles at a time, on separate days from neural recordings. We often made multiple recordings for a given muscle, especially those that we have previously noted can display responses that vary with recording location (*e.g.*, the *deltoid*).

Trial alignment and averaging

To preserve response features, it was important to compute the average firing rate across trials with nearly identical behavior. This was achieved by 1) training to a high level of stereotyped behavior, 2) discarding rare aberrant trials, and 3) adaptive alignment of individual trials prior to averaging. Because of the temporally extended nature of cycling movements, standard alignment procedures (*e.g.*, locking to movement onset) often misalign responses later in the movement. For

example, a seven-cycle movement lasted ~3500 ms. By the last cycle, a trial 5% faster than normal and a trial 5% slower than normal would thus be misaligned by 350 ms, or over half a cycle.

To ensure response features were not lost to misalignment, we developed a technique to adaptively align trials within a condition. First, trials were aligned on movement onset. Individual trials were then scaled so that all trials had the same duration (set to be the median duration across trials). Because monkeys usually cycled at a consistent speed (within a given condition) this brought trials largely into alignment: *e.g.*, the top of each cycle occurred at nearly the same time for each trial. The adaptive alignment procedure was used to correct any remaining slight misalignments. The time-base for each trial was scaled so that the position trace on that trial closely matched the average position of all trials. This involved a slight non-uniform stretching, and resulted in the timing of all key moments – such as when the hand passed the top of the cycle – being nearly identical across trials. This ensured that high-frequency temporal response features (*e.g.*, the small peak in **Figure** 2.1**G**) were not lost to averaging.

All variables of interest (firing rate, hand position, hand velocity, EMG, etc.) were computed on each trial before adaptive alignment. Thus, the above procedure never alters the magnitude of these variables, but simply aligns when those values occur across trials. The adaptive procedure was used once to align trials within a condition on a given recording session, and again to align data across recording sessions. This allowed, for example, comparison of neural and muscle responses on a matched time-base.

Other experimental datasets

Recordings from primate motor cortex during reaching have been described and analyzed previously (Elsayed et al., 2016; A.H. Lara, Elsayed, Cunningham, & Churchland, 2017). Briefly,

two male rhesus monkeys (A and B) performed center-out reaches in eight target directions on a fronto-parallel screen. This task employed three 'contexts' in which reach initiation was prompted by different cues. That manipulation was incidental to the present analysis: we analyzed only movement-related responses, which were empirically very similar across the three contexts. We therefore simply computed the trial-averaged time-varying firing rate (smoothed with a 20 ms SD Gaussian) across all reaches for each of the eight directions. Trials were aligned to movement onset and we analyzed the period from 100 ms before movement onset until 100 ms after the average time of movement offset. Neural populations included 101 and 129 neurons (monkey A and B) recorded from the arm region of motor cortex (including sulcal and surface primary motor cortex and the adjacent aspect of dorsal premotor cortex). During this same task, activity was recorded from the muscles of the upper arm (*deltoid, trapezius, biceps, brachialis, pectoralis, latissimus dorsi* muscles) using the same procedures described above (13 and 10 recordings for monkey A and B; smoothed with a 20 ms SD Gaussian). The median number of analyzed trials per direction was 48 (monkey A) and 60 (monkey B).

Data from primate V1 were recorded using natural-movie stimuli from an anaesthetized adult monkey (*Macaca fascicularis*) implanted with a 96-electrode silicon 'Utah' array (Blackrock Microsystems, Salt Lake City, UT) in left-hemisphere V1 as previously described(Seely et al., 2016). These data were recorded in the laboratory of Adam Kohn. Procedures were approved by the Animal Care and Use Committees at Albert Einstein College of Medicine (protocol #20150303). The left eye was covered. Receptive field centers (2–4 degrees eccentric) were determined via brief presentations of small drifting gratings. Stimuli, which spanned the receptive fields, were 48 natural movie clips (selected from YouTube) with 50 repeats each. The frame rate was ~95 Hz. Each stimulus lasted 2.63 s (100 movie frames followed by 150 blank frames). Spikes

from the array were sorted offline using MKsort (available at https://github.com/ripple-neuro/mksort/). A total of 108 single units and stable multi-unit isolations were included. It is unclear how anesthesia might affect trajectory tangling of this neural population. However, responses to stimuli were robust and only stimulus-evoked aspects of the responses were analyzed. Data from mouse motor cortex have been described and analyzed previously (Miri et al., 2017). Briefly, three head-fixed mice performed a task that included both a reach-to-grasp sub-task and natural treadmill walking (10 cm/s), performed in separate blocks. Multiple neurons / muscles were recorded simultaneously, but were also accumulated across days to allow analysis of larger populations. The populations for each mouse were analyzed separately. Neural recordings were made with independently movable tetrode micro-drives, lowered over the course of two weeks to primarily target layer 5. A total of 890 well-isolated units from three animals were recorded across 11 behavioral sessions. Muscle activity from the forelimb was recorded from electrodes chronically implanted in the *trapezius*, *pectoralis*, *biceps*, *triceps*, *extensor digitorum communis*, and *palmaris longus*. For two mice, recordings were made from all six of these muscles. For one mouse, recordings could only be made from four. Each muscle was recorded across eleven sessions. PCA thus extracted the top EMG signals across 66 total records for two mice and 44 for the other. Spike-trains and muscle activity were smoothed with a Gaussian filter (20 ms SD) and averaged across trials.

Preprocessing and PCA

Because PCA seeks to capture variance, it can be disproportionately influenced by differences in firing rate range (*e.g.*, a neuron with a range of 100 spikes/s has 25 times the variance of a similar neuron with a range of 20 spikes/s). This concern is larger still for EMG, where the scale is

arbitrary and can differ greatly between recordings. The response of each neuron / muscle was thus normalized prior to application of PCA. EMG data were fully normalized: $response \coloneqq response/range(response)$, where the range is taken across all recorded times and conditions. Neural data were 'soft' normalized: $response \coloneqq response/(range(response) + 5)$. We standardly (Churchland et al., 2012; Seely et al., 2016) use soft normalization to balance the desire for PCA to explain the responses of all neurons with the desire that weak responses not contribute on an equal footing with robust responses. In practice, nearly all neurons had high firing rate ranges during cycling, making soft normalization nearly identical to full normalization.

Following preprocessing, neural data were formatted as a 'full-dimensional' matrix, X^{full}, of size $n \times t$, where n is the number of neurons and t indexes across all analyzed times and conditions. We similarly formatted muscle data as a matrix, Z^{full}, of size $m \times t$, where m is the number of muscles. Unless otherwise specified, analyzed times were from 100 ms before movement onset to 100 ms after movement offset, for all conditions. Because PCA operates on mean-centered data, we mean-centered X^{full} and Z^{full} so that every row had a mean value of zero. PCA was used to find X, a reduced-dimensional version of X^{full} with the property that $X^{full} \approx VX$, where V are the PCs ('neural dimensions' upon which the data are projected). PCA was similarly used to find Z, the reduced-dimensional version of Z^{full}. For most analyses, we employed eight PCs, such that X and Z were of size $8 \times t$. Eight PCs captured 70% and 68% (monkey D and C) of the neural data variance, and 94% and 88% of the muscle data variance.

Regression

Decoding of muscle activity from neural activity was accomplished via a linear model: $Z^{full} = BX^{full}$. B was found using ridge regression. Performance was assessed using generalization R^2, using Leave-One-Out Cross Validation. Regularization strength was chosen to maximize Leave-One-Out Cross Validation performance, though in practice a broad range of regularization strengths provided similar performance. We also attempted to decode neural activity from muscle activity using the model $X^{full} = BZ^{full}$. Decoding neural activity from muscle activity was less successful than decoding muscle activity from neural activity. Although our neural recordings generally had very good signal-to-noise, we considered that poor decoding of neural activity from muscle activity (relative to decoding muscle activity from neural activity) could potentially result because neural responses tend to have higher sampling error than muscle responses. We therefore re-ran the regression above after de-noising the neural data by replacing each neuron's response with its reconstruction using the top thirty PCs. The same discrepancy was observed.

In a subsequent analysis, we decoded kinematic parameters from both predicted and empirical population activity. The predicted population response pertained only to the three middle cycles of seven-cycle movements. Thus, all decoding of kinematic parameters involved only those three cycles. Decoding employed ridge regression as described above. Regularization strength was chosen to improve generalization performance without overly sacrificing test performance. Kinematics were mean centered, and regressed against the ten dimensions of the predicted population response, or the projection of the empirical data onto the top ten PCs. Matching dimensionality ensured that it is appropriate to compare R^2 and generalization R^2 values when regressing against the predicted versus empirical population. Generalization performance was

tested by fitting to data for one direction (*e.g.*, forward cycling) and generalizing to the other (*e.g.*, backward cycling).

Tangling

Tangling was computed as described in the results (**Equation** 2.1). The neural state, $\boldsymbol{x}_t$ was an 8×1 vector comprised of the t^{th} column of X, where X is of size $8 \times t$. Muscle tangling was computed analogously, based on Z. Essentially identical results were found if we used X^{full} and Z^{full} (**Figure 2.S2**) but this was less computationally efficient and did not allow matched dimensionality between neurons and muscles. We computed the derivative of the state as $\dot{\boldsymbol{x}}_t = (\boldsymbol{x}_t - \boldsymbol{x}_{t-\Delta t})/\Delta t$, where Δt was 1 ms. When computing tangling, we employed the squared distance between derivatives, $\|\dot{\boldsymbol{x}}_t - \dot{\boldsymbol{x}}_{t'}\|^2$, because its magnitude more intuitively tracks the difference in trajectory direction. For example, if the angle between derivatives doubles from 90° to 180°, the norm grows by only 41%, but the squared norm is doubled. The constant ε was set to 0.1 times the average squared magnitude of $\boldsymbol{x}_t$ across all t. Results were essentially identical across an order of magnitude of values of ε.

Tangling estimates how non-smooth a flow-field would have to be to have produced the observed trajectories. While there are many potential measures one could use, tangling is simple to compute directly from the data, without any need to attempt to estimate the underlying flow-field. The simplicity of the tangling measure is desirable not only from a data analysis standpoint, but also from the standpoint of the optimizations in **Figure** 2.7 and **Figure 2.S7**. A more complicated measure would have resulted in a cost function that was difficult or impossible to minimize. The ability to compute tangling without fitting a flow-field is desirable because even with many conditions and temporally extended trajectories, the data leave many large 'gaps' in high-

dimensional state space, making it difficult to fit an overall flow-field with any confidence. That said, one would still hope that tangling would correlate with how well the flow-field can be fit by a dynamical model with smoothness constraints (*e.g.*, a linear model). This was indeed the case. Muscle trajectories (which were highly tangled) were less well fit by a linear dynamical model (R^2 = 0.51 and 0.37 for monkey D and C) than were the empirical neural trajectories (R^2 = 0.79 and 0.73). Despite this agreement, we avoided using the above R^2 as our primary measure, because there exist trajectories that could be readily produced by a dynamical system with smooth dynamics but are poorly described by a linear model – *e.g.*, the trajectory in Figure 7A (*right subpanel*). We also found that the quality of a linear dynamical fit was somewhat sensitive to both the span of time and the number of dimensions considered. In contrast, tangling gave consistent results regardless of such choices.

Standard Recurrent Neural Networks

We used two very different approaches to train recurrent neural networks (RNNs). In the first approach, we trained RNNs to produce a target output (**Figure** 2.5) as is conventionally done. We used a network with dynamics:

$$\boldsymbol{x}(t+1,c) = f(A\boldsymbol{x}(t,c) + B\boldsymbol{u}(c) + \boldsymbol{w}(t,c))$$

where $\boldsymbol{x}$ is the network state (the 'firing rate' of every unit) for time t and condition c. The function $f \coloneqq \tanh$ is an element-wise transfer function linking a unit's input to its firing rate, $A\boldsymbol{x}$ captures the influence of network activity on itself via the connection weights in A, $B\boldsymbol{u}$ captures external inputs, and the random vector $\boldsymbol{w} \sim N(\boldsymbol{0}, \sigma_w I)$ adds modest noise. Network output is then a linear readout of its firing rates:

$$\boldsymbol{y}(t,c) = C\boldsymbol{x}(t,c)$$

The parameters $A, B, C,$ and $\boldsymbol{x}(0, c)$ were optimized to minimize the difference between the network output, $\boldsymbol{y}$ and a target, $\boldsymbol{y}_{\text{targ}}$. That target output was the pattern of activity, across all muscles, during the middle five cycles of a seven-cycle movement. We used two conditions with different target outputs: $\boldsymbol{y}_{\text{targ}}(:,1)$ and $\boldsymbol{y}_{\text{targ}}(:,2)$ contained muscle activity during forward and backward cycling respectively. The input provided the network with the condition identity: $\boldsymbol{u}(1) = [1; 0]$ and $\boldsymbol{u}(2) = [0; 1]$.

The loss function optimized during training contained both error and regularization terms:

$$L = \sum_{t,c} \left[\frac{1}{2} \left\| \boldsymbol{y}_{\text{targ}}(t, c) - \boldsymbol{y}(t, c) \right\|_2^2 \right] + \frac{\lambda_A}{2} \|A\|_F^2 + \frac{\lambda_C}{2} \|C\|_F^2 + \sum_{t,c} \left[\frac{\lambda_x}{2} \|\boldsymbol{x}(t, c)\|_2^2 \right]$$

where the first term is the error between the network output and the target, the second and third terms penalize large recurrent and output weights respectively, and the last term penalizes large firing rates. By varying the hyper-parameters λ_A, λ_C, $\lambda_x \sigma_w$, and the initial weight values, we simulated a family of networks that found different solutions for producing the same output. This allowed us to ask whether low network-trajectory tangling was a common feature of those solutions.

We trained 1000 such networks. Hyper-parameters were drawn randomly from log uniform distributions, $\lambda_A \in [10^{-4}, 10^{-1}], \lambda_C \in [10^{-6}, 10^{1}]$, $\lambda_x \in [10^{-4}, 10^{1}]$, and $\sigma_w \in [10^{-4}, 10^{1}]$. Each RNN included $n = 100$ units. Each matrix of the RNN was initialized to a random orthonormal matrix. RNNs were trained using TensorFlow's Adam optimizer. We discarded RNNs that were not successful ($R^2 < 0.5$ between target and actual outputs). Because of the broad range of hyper-parameters, only a subset of networks (463) were successful.

As a technical point, we were concerned that, despite regularization, networks might find overly specific solutions. Each cycle of the empirical muscle activity had different small idiosyncrasies,

and optimization might promote overfitting of these small differences. We therefore added 'new' conditions to $\boldsymbol{y}_{\text{targ}}(t, c)$. Each new condition involved a target output that was almost identical to that for one of the original two conditions, but was modified such that the small idiosyncrasies occurred on different cycles. This ensured that networks produced a consistent output very close to the empirical muscle activity, but did not attempt to perfectly match small cycle-specific idiosyncrasies. The inclusion of noise via $\boldsymbol{w}$ also encouraged optimization to find robust, rather than overfit, solutions. Noise magnitude, σ_w, was a hyper-parameter that was varied across networks, to encourage varied solutions. However, σ_w was always set to zero when measuring network tangling.

Trajectory-constrained Neural Networks

To examine how tangling relates to noise-robustness (**Figure** 2.7**B**) we trained RNNs to follow a set of target internal trajectories. This involved the unconventional approach of employing both a target output, $\boldsymbol{y}_{\text{targ}}$, and a target internal network trajectory, $\boldsymbol{s}_{\text{targ}}$. Networks consisted of 100 units. Network dynamics were governed by

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) + \Delta t/\tau \left(-\boldsymbol{v}(t) + A\, f\big(\boldsymbol{v}(t)\big) + \boldsymbol{w}(t)\right)$$
$$\boldsymbol{y}(t) = C f\big(\boldsymbol{v}(t)\big)$$

where $f \coloneqq \tanh$, and $\boldsymbol{w} \sim N(\boldsymbol{0}, \sigma_w I)$ adds noise. $\boldsymbol{v}$ can be thought of as the membrane voltage and $f\big(\boldsymbol{v}(t)\big)$ as the firing rate. $Af\big(\boldsymbol{v}(t)\big)$ is then the network input to each unit: the firing rates weighted by the connection strengths. $Cf\big(\boldsymbol{v}(t)\big)$ is a linear readout of firing rates.

During training, A was adjusted using recursive least squares (Sussillo & Abbott, 2009) so that $Af\big(\boldsymbol{v}(t)\big) \approx \boldsymbol{s}_{\text{targ}}$. Training thus insured that the synaptic inputs to each unit closely followed the pre-determined trajectory defined by $\boldsymbol{s}_{\text{targ}}$. Firing rates therefore also followed a pre-determined

trajectory. C was adjusted so that $\boldsymbol{y} \approx \boldsymbol{y}_{\text{targ}}$. Training was deemed successful if the R^2 between $\boldsymbol{y}$ and $\boldsymbol{y}_{\text{targ}}$ was > 0.9. Noise tolerance was assessed as the largest value of σ_w for which the network could be trained to accurately produce the target output for five consecutive cycles ($R^2 > 0.9$ between $\boldsymbol{y}$ and $\boldsymbol{y}_{\text{targ}}$, averaged across 100 iterations) despite the constraint of following the target internal trajectory, $\boldsymbol{s}_{\text{targ}}$.

We set $\boldsymbol{y}_{\text{targ}} = [\cos t\,;\ \sin 2t]$. To construct $\boldsymbol{s}_{\text{targ}}$, we began with an idealized low-dimensional target, $\boldsymbol{s}(t)'_{\text{targ}} = [\cos t\,;\ \sin 2t\,;\ \beta \sin t]$. To give each unit a target, we set $\boldsymbol{s}_{\text{targ}} = G\boldsymbol{s}'_{\text{targ}}$ where G is a random matrix of size 100×3 with entries drawn independently from a uniform distribution from -1 to 1. Noise tolerance was tested for a range of values of β. That range produced target trajectories that varied greatly in their tangling, allowing us to examine how tangling related to noise tolerance. Noise tolerance was the largest magnitude of state noise for which the network still produced the desired output. For each target trajectory, and each of the 20 random initializations of A, C, and G, we doubled σ_w starting at 0.005 until we found the noise tolerance. We then computed the average (and SEM) noise tolerance across the 20 parameter initializations.

Predicting neural population activity

The optimization described by **Equation** 2.2 was performed using the Theano Python module. Optimization was initialized either with $\hat{X}_{init} = Z$, or with $\hat{X}_{init} = Z + noise$ where the noise was smooth with time but independent for each dimension. Both $\hat{X}$ and Z were $10 \times T$; they contained the projection onto the top ten PCs. T is the total number of timepoints across the conditions being considered. Specifically, we predicted neural activity for three middle cycles of forward cycling and three middle cycles of backward cycling (both taken from seven-cycle movements). Because dimensionality is equal for $\hat{X}$ and Z, the ability to decode Z from $\hat{X}$ will suffer as optimization

modifies $\hat{X}$. However, because some dimensions of Z contain more variance than others, $\hat{X}$ can gain considerable new structure while compromising the decode only modestly. This tradeoff can be determined by the choice of λ. However, for scientific reasons, we employed a modified approach to better control that tradeoff. We wished to ensure that the predictions made by different cost functions all encoded muscle activity equally well. This aids interpretation when comparing the results of the optimization in **Figure** 2.7**C,D** with optimizations using different cost functions in **Figure 2.S7.** By matching encoding accuracy, any differences in similarity must be due to other structure that differs due to the cost function being optimized. Thus, instead of minimizing the first term of **Equation** 2.2 (which attempts to create a perfect decode) we minimized the squared difference between the decode R^2 and 0.95. We only considered optimizations that achieved this with a tolerance of 0.01. This approach insures that muscle encoding is equally good for the predicted populations responses yielded by different cost functions. Optimizations employed gradient descent using an inexact line search for the Wolfe conditions $c_1 = 0.05$ and $c_2 = 0.1$. As a technical point, the derivative used to compute $Q(t_{\text{end}})$ was based on the assumption that the three-cycle pattern would repeat.

Similarity between empirical and predicted data

We assessed similarity using a modified version of canonical correlation (Cunningham & Ghahramani, 2015). This method finds a pair of orthogonal transformations, one for each dataset, that maximizes the correlation between the transformed datasets. Specifically, for mean-centered datasets $X_a \in \mathbb{R}^{KxT}$ and $X_b \in \mathbb{R}^{KxT}$, similarity is:

$$S(X_a, X_b) = \underset{M_a, M_b}{\text{argmax}} \frac{tr(M_a^\intercal X_a X_b^\intercal M_b)}{\sqrt{tr(M_a^\intercal X_a X_a^\intercal M_a) tr(M_b^\intercal X_b X_b^\intercal M_b)}} .$$

Subject to the constraint that M_a and M_b are orthonormal matrices. Similarity will thus be unity if two datasets are the same but for an orthonormal transformation. Note also that an overall shift of one dataset relative to the other does not impact similarity because the data are mean-centered before computing similarity. Due to the normalization in the denominator of the above cost function, similarity is also not impacted by an isotropic scaling of one dataset relative to the other.

Supplementary Materials

Supplemental Note

Here we show that, given limits on how rapidly a flow-field can change, when two trajectories (or two portions of the same trajectory) come close and then diverge, a potential instability is inevitable. We define a potential instability as a direction along which an error will grow with time in the local vicinity. The argument below is a simple proof by contradiction. Avoiding a potential instability requires that, for all directions, local errors shrink with time. For a linearized system, this implies that all eigenvalues are less than zero. Yet if two trajectories diverge, there must be at least one positive eigenvalue.

Assume two time-evolving trajectories, $\boldsymbol{x}_1(t)$, and $\boldsymbol{x}_2(t')$. These could be two portions of a larger trajectory or could correspond to two different conditions. We consider the moment where they become closest: *i.e.*, when $\|\boldsymbol{x}_1(t) - \boldsymbol{x}_2(t')\|$ is smallest. Without loss of generality, we assume this happens at $t = 0$ and $t' = 0$. We also consider the state, $\overline{\boldsymbol{x}}$ halfway between $\boldsymbol{x}_1(0)$ and $\boldsymbol{x}_2(0)$. Without loss of generality, we define $\overline{\boldsymbol{x}}$ as the origin. Thus $\boldsymbol{x}_1(0) = -\boldsymbol{x}_2(0)$. As in Supplemental Figure 1, we assume that tangling between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is high because $\|\dot{\boldsymbol{x}}_1(0) - \dot{\boldsymbol{x}}_2(0)\|$ is large while $\|\boldsymbol{x}_1(0) - \boldsymbol{x}_2(0)\|$ is small. We can therefore use the Taylor series to approximate the flow-field at state $\boldsymbol{x}$ in the vicinity of $\overline{\boldsymbol{x}}$. We ignore higher-order terms:

$$\dot{\boldsymbol{x}} = \boldsymbol{a} + B\boldsymbol{x}$$

where the matrix B is the Jacobian evaluated at $\boldsymbol{x} = \boldsymbol{0}$.

Because both $\boldsymbol{x}_1(0)$ and $\boldsymbol{x}_2(0)$ are near $\overline{\boldsymbol{x}}$, we have:

$$\dot{\boldsymbol{x}}_1(0) = \boldsymbol{a} + B\boldsymbol{x}_1(0)$$

and

$$\dot{\boldsymbol{x}}_2(0) = \boldsymbol{a} + B\boldsymbol{x}_2(0) = \boldsymbol{a} - B\boldsymbol{x}_1(0).$$

We now consider some perturbation of the $\boldsymbol{x}_1$ trajectory, such that $\boldsymbol{x}_1'(0) = \boldsymbol{x}_1(0) + \boldsymbol{\varepsilon}$. Stability requires, $\forall\ \boldsymbol{\varepsilon}$:

$$\|\boldsymbol{x}_1'(\Delta t) - \boldsymbol{x}_1(\Delta t)\|^2 < \|\boldsymbol{x}_1'(0) - \boldsymbol{x}_1(0)\|^2$$

$$\implies \left\|\left(\boldsymbol{x}_1'(0) + \Delta t\left(\boldsymbol{a} + B\boldsymbol{x}_1'(0)\right)\right) - \left(\boldsymbol{x}_1(0) + \Delta t\left(\boldsymbol{a} + B\boldsymbol{x}_1(0)\right)\right)\right\|^2 < \|\boldsymbol{x}_1(0) + \boldsymbol{\varepsilon} - \boldsymbol{x}_1(0)\|^2$$

$$\implies \|\boldsymbol{\varepsilon} + \Delta t B\boldsymbol{\varepsilon}\|^2 < \|\boldsymbol{\varepsilon}\|^2$$

$$\implies \|\boldsymbol{\varepsilon}\|^2 + 2\Delta t\boldsymbol{\varepsilon}^{\mathrm{T}}B\boldsymbol{\varepsilon} + \Delta t^2\boldsymbol{\varepsilon}^{\mathrm{T}}B^{\mathrm{T}}B\boldsymbol{\varepsilon} < \|\boldsymbol{\varepsilon}\|^2$$

$$\implies \|\boldsymbol{\varepsilon}\|^2 + 2\Delta t\boldsymbol{\varepsilon}^{\mathrm{T}}B\boldsymbol{\varepsilon} < \|\boldsymbol{\varepsilon}\|^2 \text{ , as } \Delta t^2 \text{ is very small.}$$

$$\implies \boldsymbol{\varepsilon}^{\mathrm{T}}B\boldsymbol{\varepsilon} < 0$$

Because this must be true for all $\boldsymbol{\varepsilon}$, this is equivalent to stating that all eigenvalues of B must be negative. However, because $\boldsymbol{x}_1(t)$, and $\boldsymbol{x}_2(t)$ are closest at $t = 0$, we have:

$$\|\boldsymbol{x}_1(\Delta t) - \boldsymbol{x}_2(\Delta t)\|^2 > \|\boldsymbol{x}_1(0) - \boldsymbol{x}_2(0)\|^2$$

$$\implies \left\|\left(\boldsymbol{x}_1(0) + \Delta t\left(\boldsymbol{a} + B\boldsymbol{x}_1(0)\right)\right) - \left(\boldsymbol{x}_2(0) + \Delta t\left(\boldsymbol{a} + B\boldsymbol{x}_2(0)\right)\right)\right\|^2 > \|\boldsymbol{x}_1(0) - \boldsymbol{x}_2(0)\|^2$$

$$\implies \|\ 2\boldsymbol{x}_1(0) + 2\Delta t B\boldsymbol{x}_1(0)\|^2 > \|2\boldsymbol{x}_1(0)\|^2$$

$$\implies \|\boldsymbol{x}_1(0)\|^2 + 2\Delta t\boldsymbol{x}_1(0)^{\mathrm{T}}B\boldsymbol{x}_1(0) + \Delta t^2\boldsymbol{x}_1(0)^{\mathrm{T}}B^{\mathrm{T}}\boldsymbol{x}_1(0) > \|\boldsymbol{x}_1(0)\|^2$$

$$\implies \|\boldsymbol{x}_1(0)\|^2 + 2\Delta t\boldsymbol{x}_1(0)^{\mathrm{T}}B\boldsymbol{x}_1(0) > \|\boldsymbol{x}_1(0)\|^2 \quad \text{, as } \Delta t^2 \text{ is very small.}$$

$$\implies \boldsymbol{x}_1(0)^{\mathrm{T}}B\boldsymbol{x}_1(0) > 0$$

This is in contradiction to the claim above that $\boldsymbol{\varepsilon}^{\mathrm{T}}B^{\mathrm{T}}\boldsymbol{\varepsilon} < 0$ for $\forall\ \boldsymbol{\varepsilon}$. Equivalently, it implies that at least one eigenvalue of B must be positive, in contrast to the claim above that all eigenvalues must be negative.

Thus, local stability is inconsistent with the fact that trajectories are close but diverging. The above argument does not strictly depend on $\|\dot{\boldsymbol{x}}_1(0) - \dot{\boldsymbol{x}}_2(0)\|$ being large. However, a larger $\|\dot{\boldsymbol{x}}_1(0) - \dot{\boldsymbol{x}}_2(0)\|$ implies larger positive eigenvalue(s) of B. All other things being equal, this will result in a larger potential instability due to greater local divergence.

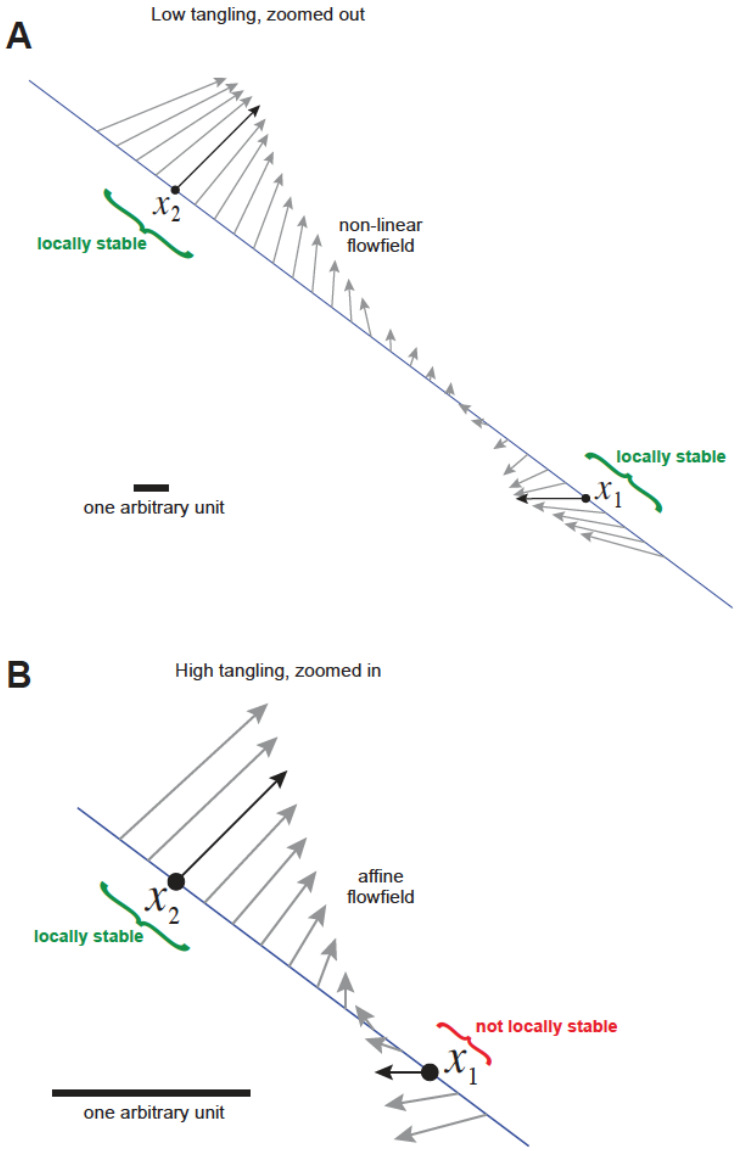

Figure 2.S1 Illustration of how low tangling allows stable flow-fields, while high tangling leads to potential instabilities.

Illustration of how low tangling allows stable flow-fields, while high tangling leads to potential instabilities. **A**. Illustrated are two states, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, that are weakly tangled with one another; they have very different derivatives but are well separated in state space. Due to that distance, it is possible for the flow-field to be locally stable in the vicinity of both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Gray arrows plot one potential flow-field for points along the line between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. In the example shown, if the neural state in the vicinity of $\boldsymbol{x}_1$ is perturbed slightly along the blue line towards $\boldsymbol{x}_2$, then that error will be reduced by the self-correcting structure of the flow-field (arrows converge locally). Note that this requires a non-linear flow-field. **B**. Illustration of potential instabilities when tangling is high. We assume that high tangling between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ occurs because $\|\dot{\boldsymbol{x}}_2 - \dot{\boldsymbol{x}}_1\|$ is large while $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|$ is small. We can express the flow-field using the Taylor series expansion around $\boldsymbol{x}_1$: $\dot{\boldsymbol{x}} = \boldsymbol{a} + B(\boldsymbol{x} - \boldsymbol{x}_1) +$ higher order terms. We

assume some limit on smoothness, such that in the vicinity of $\boldsymbol{x}_1$, higher order terms are small. Conversely, because $\|\dot{\boldsymbol{x}}_2 - \dot{\boldsymbol{x}}_1\|$ is large, B must be large. Thus, in the vicinity of $\boldsymbol{x}_1$, dynamics are dominated by the first two terms of the expansion. Therefore, if we consider a point $\boldsymbol{x}'$ that is a distance d along the line intersecting $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, then $\dot{\boldsymbol{x}}' = \dot{\boldsymbol{x}}_1 + \frac{d}{\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|}(\dot{\boldsymbol{x}}_2 - \dot{\boldsymbol{x}}_1)$. In the present example, given $\dot{\boldsymbol{x}}_1$ and $\dot{\boldsymbol{x}}_2$ illustrated by the black arrows, the resulting flow-field is shown in gray. This flow-field is locally unstable near $\boldsymbol{x}_1$; the gray arrows diverge from that point. This cannot be avoided if the local flow-field is locally linear. Thus, when the local approximation is limited to being linear (or affine), errors introduced by noise cannot be consistently corrected. This 'potential instability' is compounded by the fact that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are close, even small amounts of noise may move the state a relatively large distance. Whether this actually renders the system unstable depends on the level of noise, and on the structure of the rest of the flow-field. Thus, high tangling does not necessarily produce global instabilities, but does introduce potential instabilities. In particular, a potential instability necessarily occurs whenever two highly tangled states are diverging. This is demonstrated formally in the **Supplemental Note**.

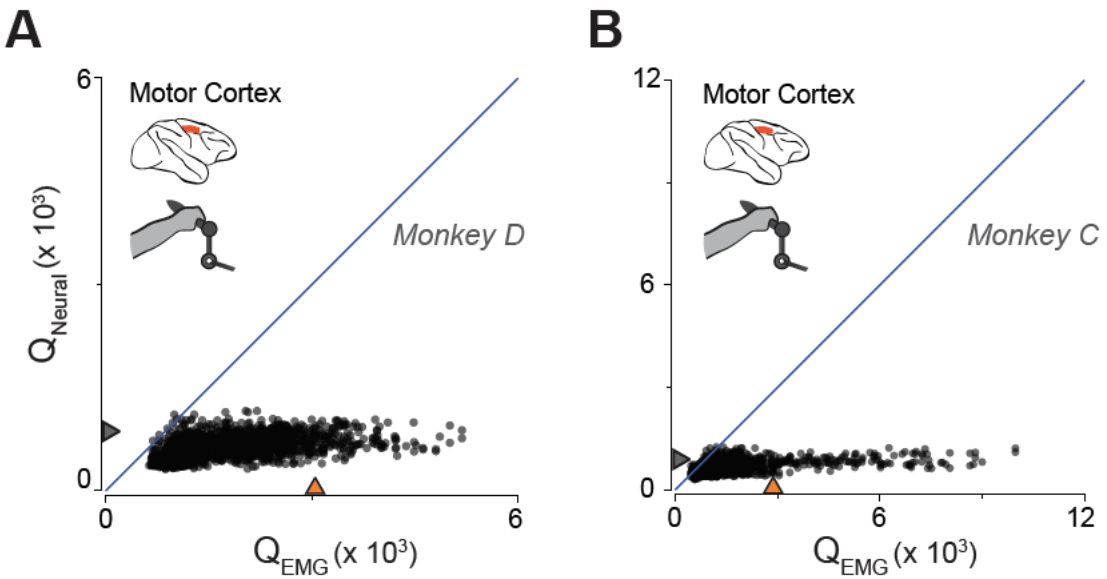

Figure 2.S2 Trajectory tangling without dimensionality reduction.

A,B. Analysis was as in Figure 6A,B, except no dimensionality reduction was employed. Tangling was instead based on vectors that included the activity of every neuron / muscle.

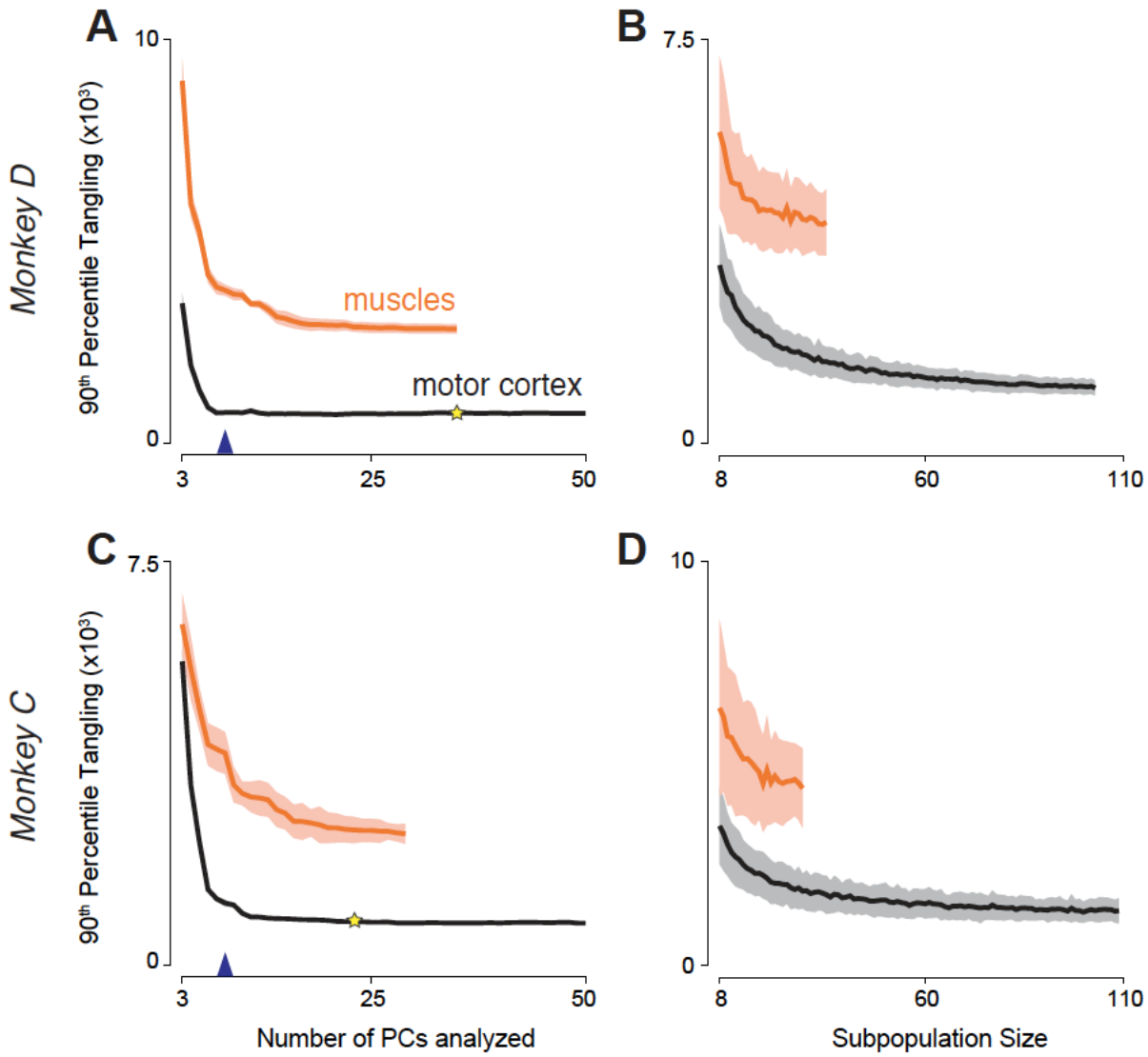

Figure 2.S3 The difference between neural- and muscle-trajectory tangling is not due to differences in dimensionality or population size

A. Neural-trajectory tangling (*black*) and muscle-trajectory tangling (*orange*) as a function of the number of PCs used when computing tangling. Tangling was quantified as the 90th percentile of the distribution. The *triangle* on the horizontal axis indicates eight PCs, which were used for the analyses in Figure 6. Flanking traces show the standard error, computed via bootstrap (see Figure 6 legend). *Star* indicates the number of neural PCs necessary such that the percentage of variance captured equaled that captured by eight muscle PCs. Neural-trajectory tangling changes little over the range from eight dimensions to the dimensionality indicated by the star. Thus, the difference in neural versus muscle tangling would be essentially identical if we had matched the variance accounted for rather than the number of PCs. Data are for monkey D. **B.** Neural-trajectory tangling (*black*) and muscle-trajectory tangling (*orange*) as a function of the number of recordings considered when computing tangling. For a given number of recordings, we drew that many neurons (or muscles) from the full population and computed tangling. Flanking traces show the standard error, computed via bootstrap across 200 such repetitions. Data are for monkey D. **C,D.** Same as A,B but for monkey C.

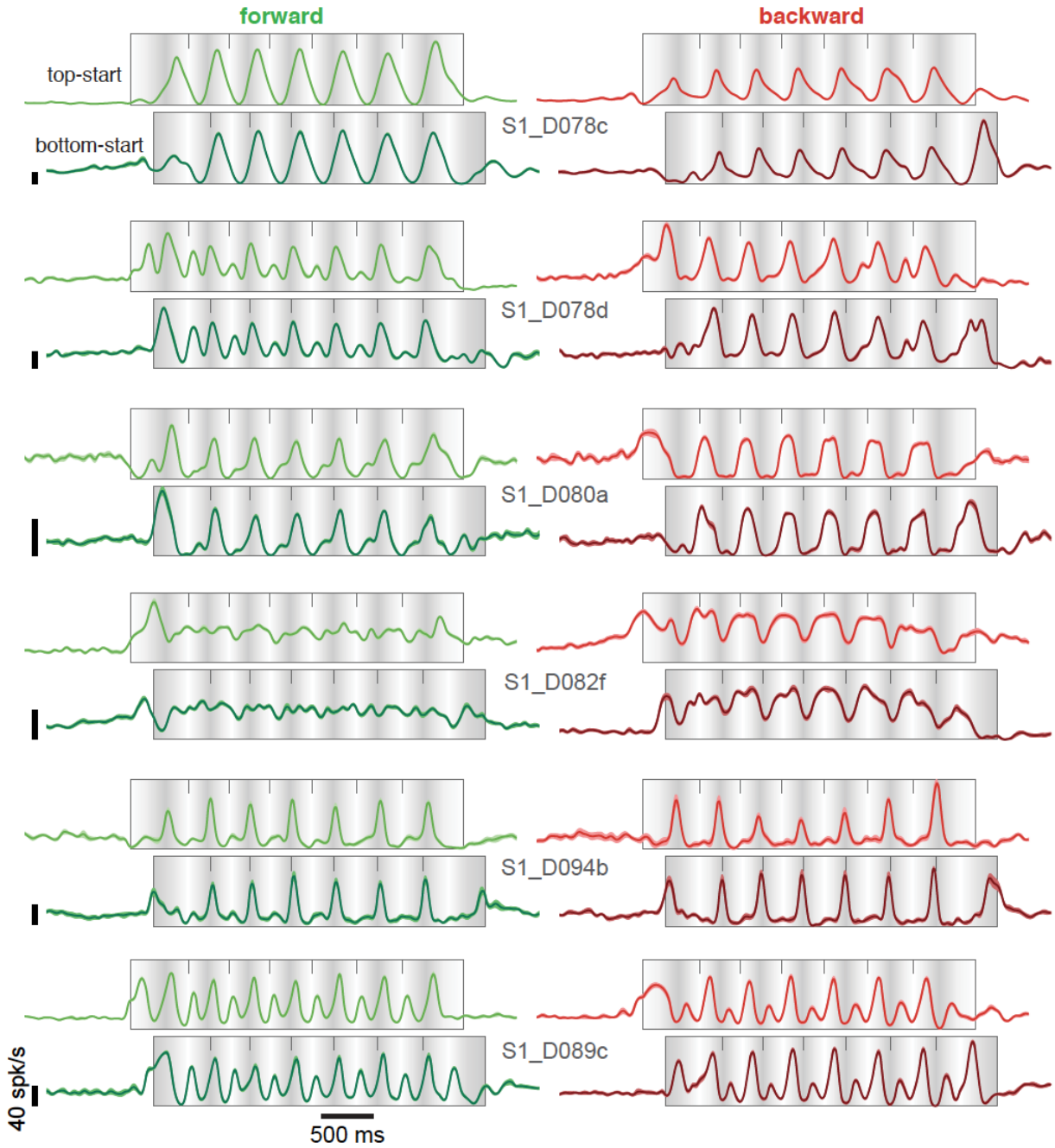

Figure 2.S4 Firing rates of six example neurons recorded from primary somatosensory cortex.

Same format as Figure 2 and 3.

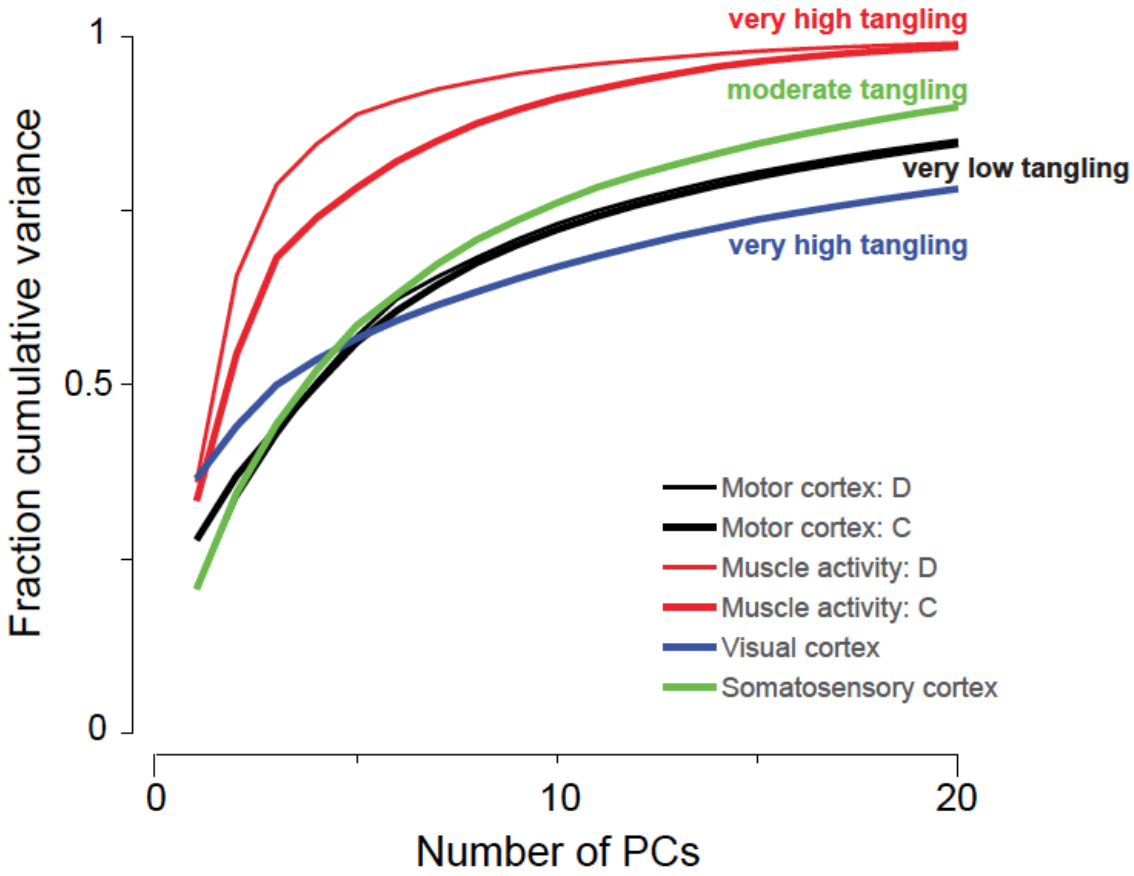

Figure 2.S5 Tangling cannot be predicted from the dimensionality of a dataset.

The fraction of cumulative variance accounted for is plotted as a function of number of PCs used for reconstruction. *Red traces* corresponding to muscle activity climb quickly, indicating that these datasets are relatively low-dimensional: most of the variance is captured by a few dimensions. *Blue and green traces* (corresponding to visual and somatosensory cortex data respectively) climb more slowly, indicating higher dimensionality. In spite of these differences in dimensionality, muscle activity, visual cortex activity and somatosensory cortex data all possess moderate to high tangling. Motor cortex data (*black traces*) is intermediate in dimensionality relative to visual and somatosensory cortex yet has strikingly low tangling.

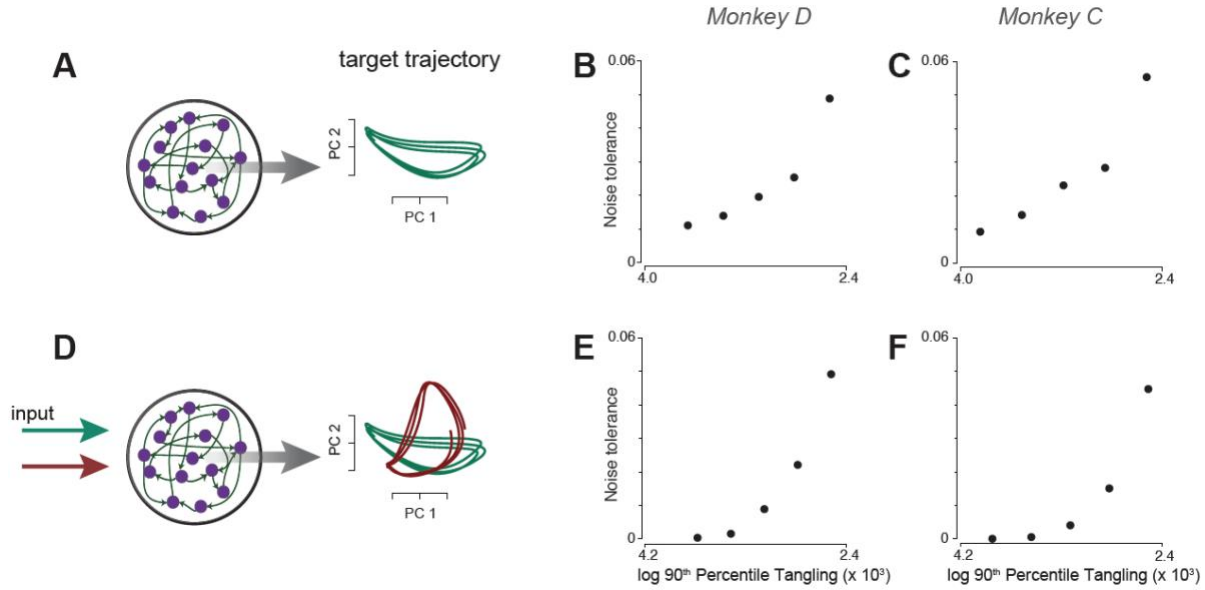

Figure 2.S6 Relationship between low tangling and noise robustness in networks trained to follow specified internal trajectories.

These trajectories encoded muscle activity with varying degrees of tangling. **A.** Schematic of network architecture and internal trajectory for networks trained to produce trajectories corresponding to forward cycling only. Networks (50 fully connected units) were trained to produce ten-dimensional target trajectories that encode muscle activity with varying degrees of trajectory tangling. To create target trajectories, we used an optimization that was the same as that described in the main text (and that produced the data in Figure 7C-F) but was applied to a single cycle of muscle data for forward cycling only. Optimization was repeated 10 times with smooth noise added during initialization to produce a family of solutions. As optimization ran, we kept the solution for different iterations: 0, 1, 2, 3, 4, 5, 10 ,100, and the final iteration. This yielded 90 trajectories: one for each optimization and iteration. These trajectories were all ten-dimensional and had a wide variety of tangling values. For each such trajectory, 20 networks (each with a different set of initial weights) were trained to autonomously and repeatedly follow that trajectory. As for Figure 7B, networks were not trained to produce the trajectory as an output but rather to internally follow that trajectory. **B,C**. Analysis of the noise robustness of the networks described in A. Noise tolerance was assessed by training networks in the presence of different levels of additive Gaussian noise. Noise tolerance was defined as the maximum noise level at which the network still followed the target trajectory. Each black circle plots the mean noise tolerance across many networks whose tangling fell within a given bin. Standard errors are within the symbol size. **D.** Schematic for networks trained to produce trajectories corresponding to either forward or backward cycling depending on an input. The input was two-dimensional. The command to produce forward / backward cycling involved one dimension being high and the other low. Each input dimension was connected to all network units with random weights. All other details are as in A. **E,F.** Same as B,C, but for the networks described in D.

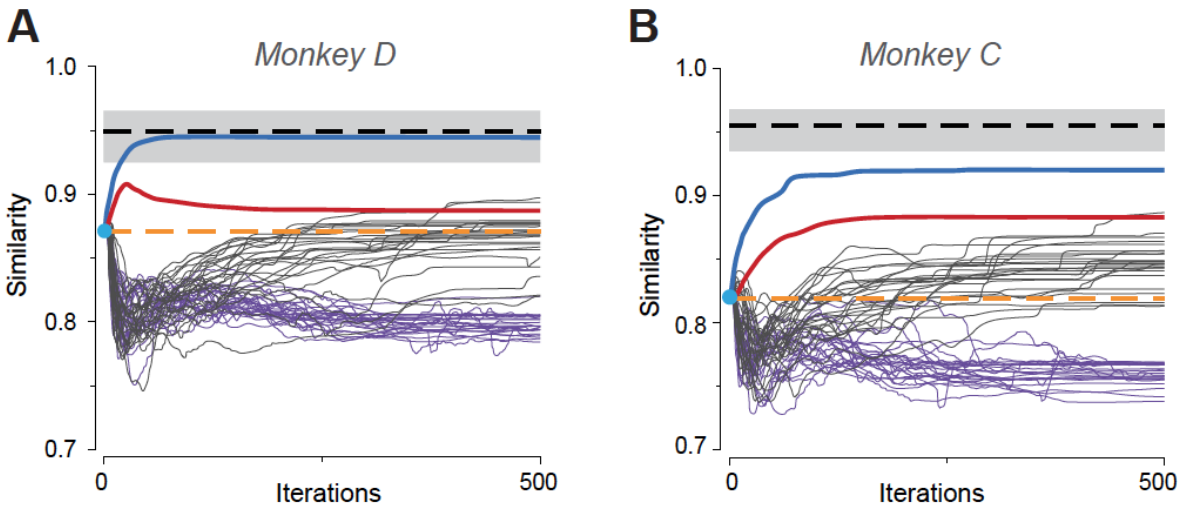

Figure 2.S7 Elaboration of analyses in Figure 7C,D

A,B. Same as Figure 7C,D but using additional cost functions. These cost functions are described below, and formalized subsequently. Each cost function embodies a hypothesis regarding the relationship between neural and muscle activity. The similarity metric thus indicates how well that hypothesis predicts the data. *Blue traces* (reproduced from Figure 7) show similarity between empirical and predicted population responses when prediction employed the cost function in Equation 2. That cost function included linear-decode error and trajectory tangling. Optimization thus embodies the hypothesis that neural activity seeks to encode muscle activity fairly directly while maintaining low tangling. *Purple traces*: predictions yielded by minimizing non-linear decode error and the L2-norm of population activity. Optimization thus embodies the hypothesis that neural activity may wish to be as modest as possible while still allowing muscle activity to be decoded. Each muscle was allowed its own non-linearity, the parameters of which were optimized. This potentially allowed neural activity to be lower-dimensional and/or simpler than muscle activity, with different patterns of activity across muscles accounted for via different non-linearities. In principle, this might have explained why the dominant neural signals are 'simpler' and different from the dominant muscle signals. In fact, similarity between the empirical and predicted populations typically declined. (There were many local minima so the algorithm was run from many different initializations.) *Gray traces*: predictions yielded by minimizing both non-linear decode error and trajectory tangling. This cost function embodies the same hypothesis as in Equation 2, but allows each muscle's activity to be decoded nonlinearly as above. Across multiple initializations, similarity occasionally increased, especially when compared to the purple traces. However, similarity did not increase to the same degree as for the simpler cost function in Equation 2. This might mean that the 'true' readout is already close to linear (such that the constraint of linearity is beneficial). More likely, the space of non-linear readouts is sufficiently large that we did not find an instance where the non-linear model improved upon the linear approximation. *Red trace*: prediction yielded by minimizing linear-decode error and trajectory curvature within each condition. Trajectory curvature is effectively a local measure of tangling. Similarity increased, but not as much as if tangling was minimized directly. *Not shown*: prediction yielded by minimizing linear-decode error and sparseness. Similarity declined dramatically and immediately, with traces falling off the bottom of the plot.

Cost functions

All cost functions were of the form:

$$\hat{X} = \underset{X}{\operatorname{argmin}} \sum_{k=1}^{K} \lambda_k f_k(X, Z)$$

where $f_k$is some function of the input data and λ_k are scaling coefficients used to ensure that one term of the cost function did not dominate at the expense of the others. The arguments of $f_k()$ are the optimization variable, X and the empirical muscle activity, Z. All cost functions examined in Supplementary Figure 7 are described below in terms of different definitions of $f_k()$.

Muscle encoding and low tangling (same as **Equation** 2.2)

$$f_1(X, Z) = f_{\text{decode}}(X, Z) = \left\| Z - ZX^{\dagger}X \right\|_F^2$$

$$f_2(X) = f_{\text{tangling}}(X) = \sum_t Q_X(t)$$

Nonlinear mapping with L-2 minimization

$$f_1(X, \bar{Z}) = f_{\text{decode-nonlin}}(X, \bar{Z}) = \left\| \bar{Z} - \hat{Z} \right\|_F^2$$

$\bar{Z}$ contains individual muscle activity. Here we consider the activity of all muscles individually (rather than the top ten PCs as above) because this matters in the non-linear case. The hypothesis being considered is that motor cortex may use a simplified set of muscle 'synergies' that becomes, via a set of non-linear transformations, the activity of each muscle. $\hat{Z} = \boldsymbol{\alpha} + \tanh(BX + \boldsymbol{\gamma})$ with the parameters $\boldsymbol{\alpha}$, B, and $\boldsymbol{\gamma}$ optimized to minimize $f_{\text{decode-nonlin}}(X, \bar{Z})$.

$$f_2(X) = f_{\text{norm}}(X) = \|X\|_F^2$$

where F denotes the Frobenius norm.

Nonlinear mapping with tangling minimization:

$$f_1(X, \bar{Z}) = f_{\text{decode}-\text{nonlin}}(X, \bar{Z})$$

$$f_2(X) = f_{\text{tangling}}(X)$$

where $f_{\text{decode}-\text{nonlin}}$ and f_{tangling} are as described above.

Low curvature:

$$f_1(X, Z) = f_{\text{decode}}(X, Z)$$

$$f_2(X) = f_{curvature}(X) = \sum_t \frac{\|\dot{\boldsymbol{x}}_t^{\text{norm}} - \dot{\boldsymbol{x}}_{t-1}^{\text{norm}}\|}{s_t}$$

where,

$$\dot{\boldsymbol{x}}_t^{\text{norm}} = \frac{\dot{\boldsymbol{x}}_t}{\|\dot{\boldsymbol{x}}_t\|}$$

and s_t is the normalized 'speed' of the neural trajectory,

$$s_t = \frac{\|\dot{\boldsymbol{x}}_t\|}{\sum_{t'} \|\dot{\boldsymbol{x}}_{t'}\|}$$

As a technical point, we wished to ensure that the predictions made by different cost functions all encoded muscle activity equally well. By matching the accuracy of muscle encoding, any differences in similarity must be due to other structure introduced during optimization. We therefore modified $f_{\text{decode}}(X, Z)$ and $f_{\text{decode}-\text{nonlin}}(X, \bar{Z})$ so that they were minimized when decode accuracy had an R^2 of 0.95, rather than 1.0. We only considered optimizations that achieved this with a tolerance of 0.01.

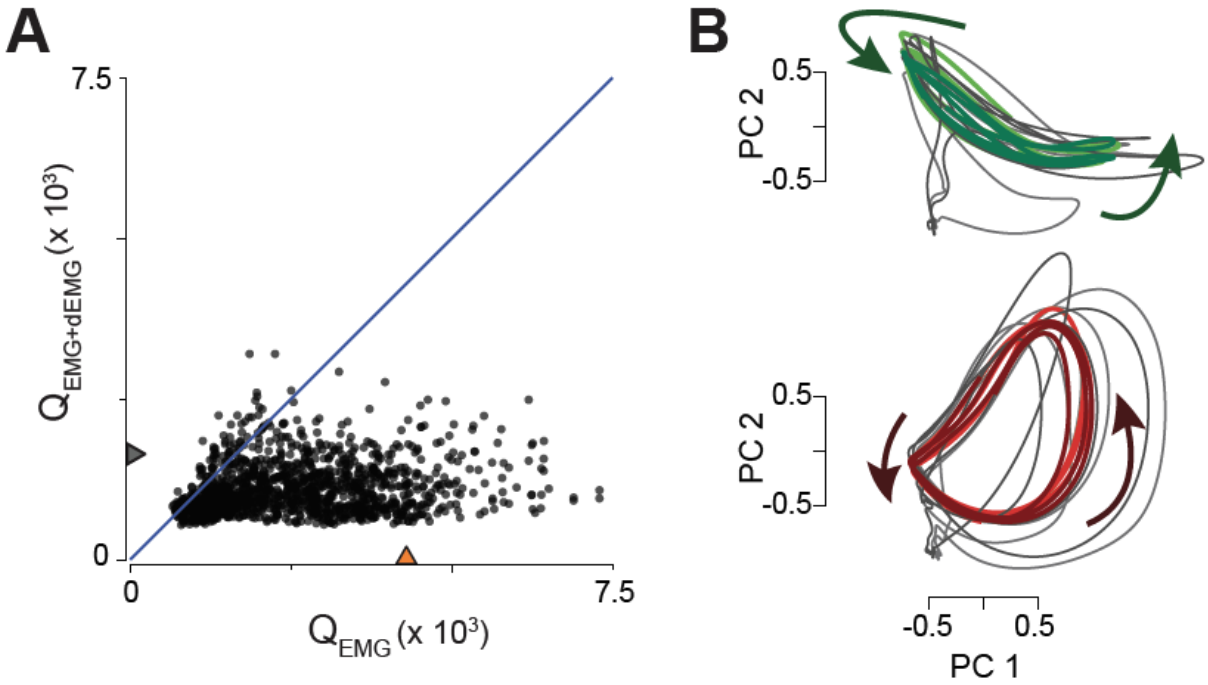

Figure 2.S8 Examination of tangling for a simulated dataset based on the hypothesis that neural activity might encode muscle activity and its derivatives

Each unit in this population had a response that was either the response of a given muscle or the derivative of that response. All units were normalized to have a response range of one. **A.** Tangling for a simulated dataset based on the muscle activity of monkey D. As expected, the simulated dataset has fairly low tangling. This is essentially insured by the addition of derivatives. Thus, introduction of derivatives is one potential way of reducing tangling. **B.** Projection of simulated data onto the top two PCs for forward (*top*) and backward (*bottom*) cycling. Compare with Figure 4D. Although this simulated dataset had fairly low tangling, the dominant signals did not qualitatively resemble the dominant signals in the neural population. For example, trajectories were often elongated and rather than circular. Further, this simulated population did not result in a consistent increase in quantitative similarity to the empirical data. Compared with the improvement in similarity produced by the optimization for low tangling directly (Figure 7C,D) the improvement in similarity that resulted from including derivatives of muscle activity was modest (43.5% as large for monkey D) or non-existent (-4.3% for monkey C).

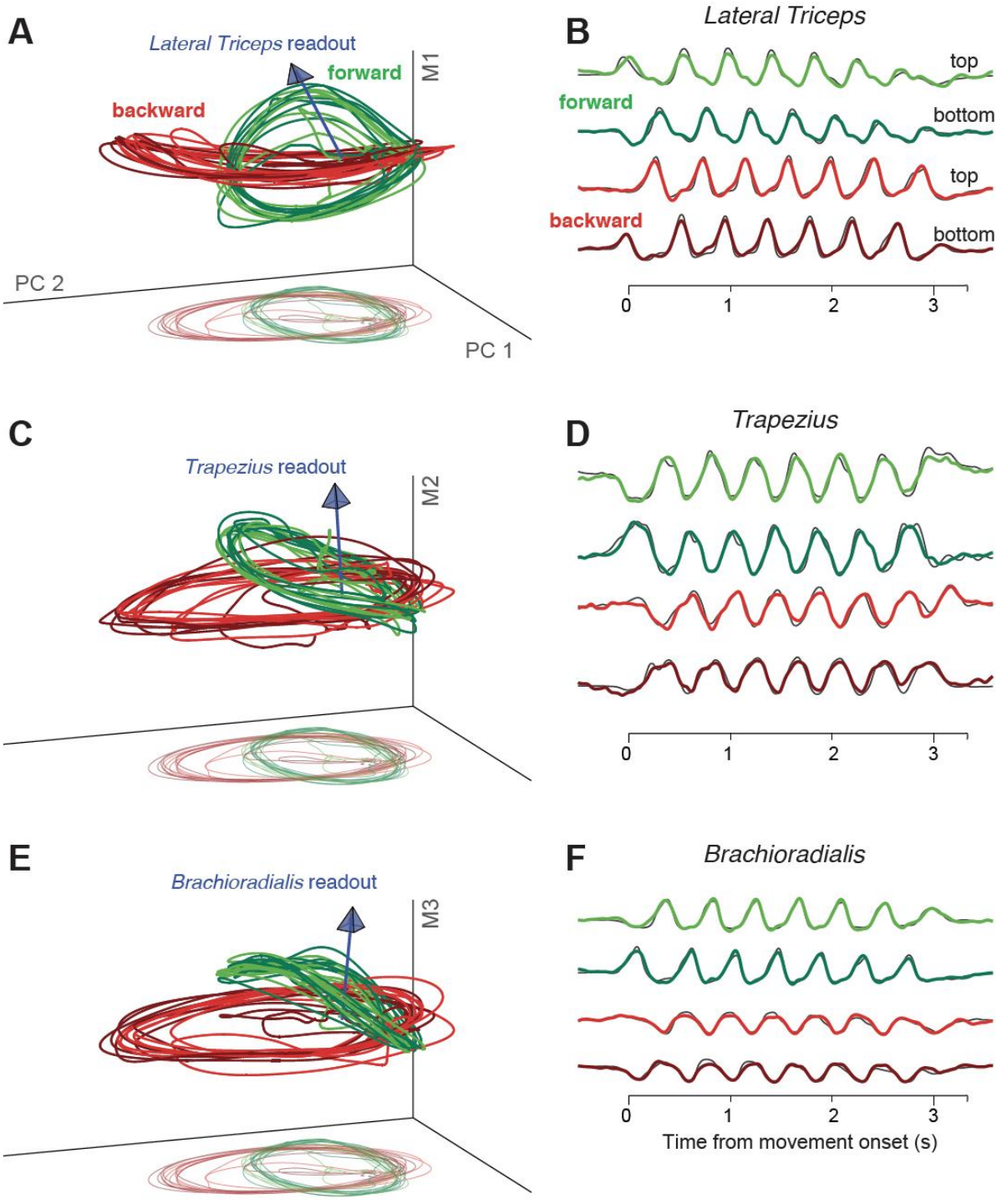

Figure 2.S9 Muscle-like signals coexist with signals that contribute to low tangling

Same format as Figure 8 but for monkey C.

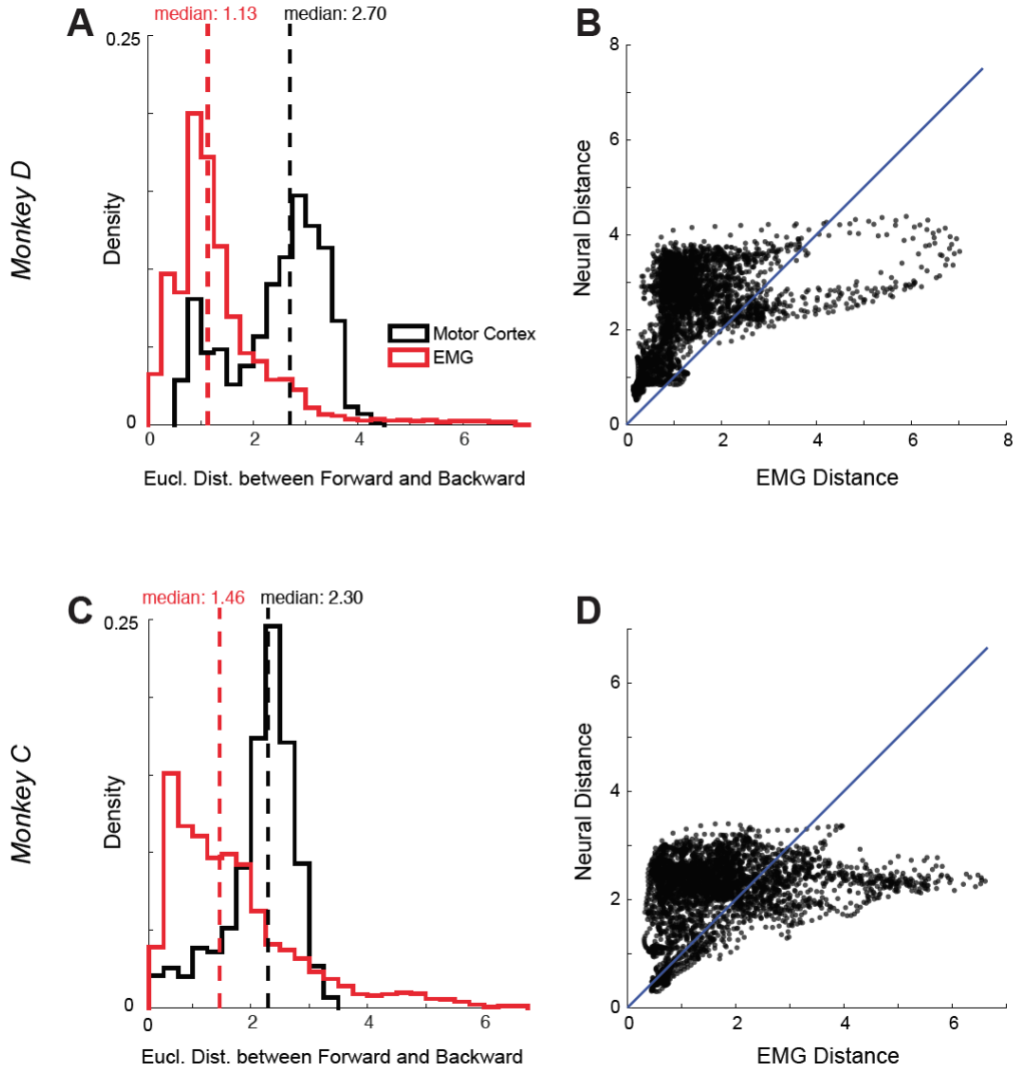

Figure 2.S10 Examination of an alternative metric related to tangling: the distance between trajectories corresponding to forward and backward cycling.

This analysis examines the possibility that low neural- versus muscle-trajectory tangling is due in part to greater separation between forward / backward trajectories for the neural population relative to the muscle population. This was indeed the case. Datasets were first reduced to 8-dimensions and normalized to have unit variance (so that distances are comparable between datasets). For each time point for a given cycling direction, we computed the closest distance between that state and all states corresponding to the opposite cycling direction. **A.** Histograms of that distance for all time points for monkey D. *Red distributions* corresponding to muscle activity are shifted left relative to *black distributions* corresponding to neural data. *Dashed lines* show distribution medians. This analysis reveals that trajectories for forward cycling and trajectories for backward cycling tend to be better separated for neural versus muscle populations. Other analyses (not shown) indicate that this effect is largely due to the fact that the subspaces occupied during forward and backward cycling overlap less than the corresponding subspaces for muscle trajectories. **B.** The same data as in A presented as a scatter plot. Most dots lie above the line with unity slope (*blue line*) indicating greater separation for neural versus muscle trajectories. Most cases where separation is greater for the muscle data involve cases where separation was high for both. **C,D.** Same as A,B for monkey C.

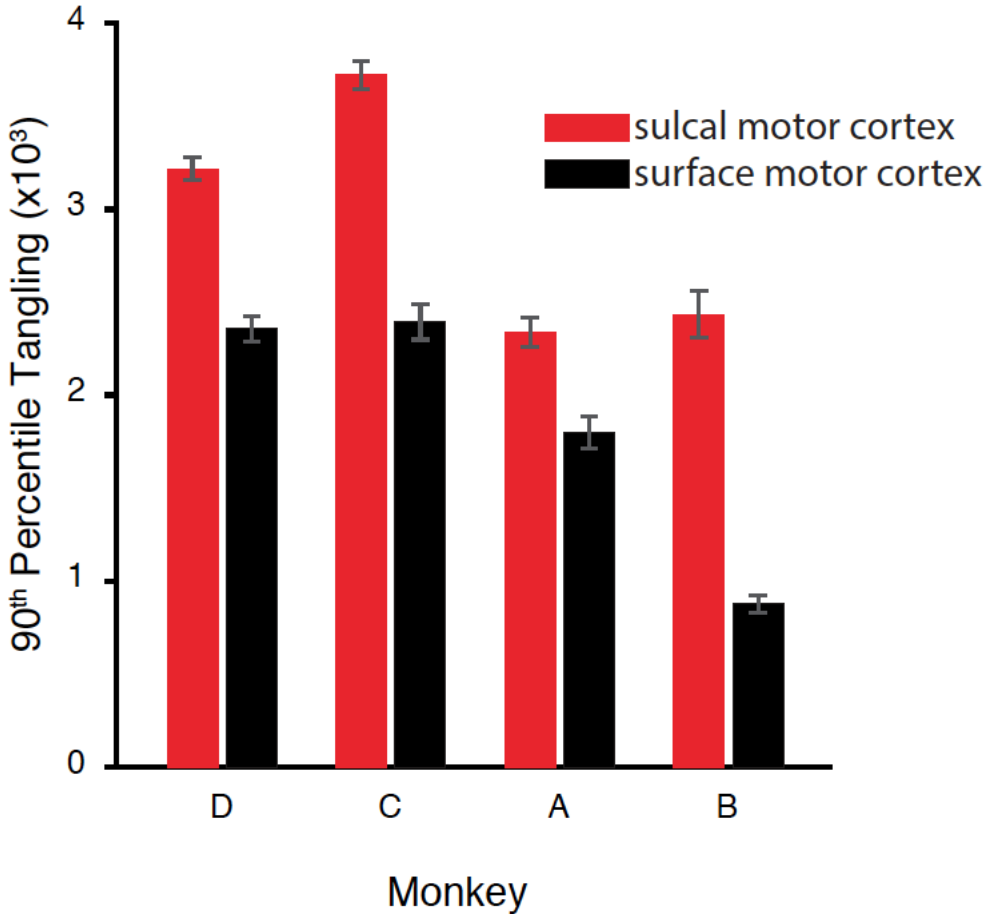

Figure 2.S11 Tangling is modestly but consistently higher in sulcal verses surface motor cortex

Red bars: 90th percentile tangling in a subpopulation of the most sulcal 10-15 neurons for each dataset. *Black bars:* Same as red but for surface motor cortex. Flanking standard errors were computed via bootstrap (see Fig. 6 legend).

Acknowledgements

We thank C. Hussar for task development, and Y. Pavlova for animal care. Support provided by the Grossman Center for the Statistics of Mind, Burroughs Wellcome Fund (MMC), Searle Scholars Program (MMC), Sloan Foundation (MMC and JPC), Simons Foundation (MMC, JPC, LFA, TMJ, AK), McKnight Foundation (MMC, JPC), Helen Hay Whitney Foundation (AM), NIH Director's New Innovator Award DP2 NS083037 (MMC), NIH NS033245 (TMJ), NIH EY016774 (AK), NIH CRCNS R01NS100066 (MMC and JPC), NIH 1U19NS104649 (MMC, LFA, TMJ), NIH R01MH93338 (LFA), NIH F32NS092350 (AHL), NIH 5T32NS064929 (AAR), National Science Foundation (NJM, JSS, SRB), Kavli Foundation (MMC and TMJ), Klingenstein Foundation (MMC), Project ALS (TMJ), Mathers Foundation (TMJ) and the Howard Hughes Medical Institute (TMJ).

Author contributions

M.M.C. conceived the experiments; A.A.R., B.M.L. and S.M.P. collected main datasets. N.J.M. contributed some EMG recordings; A.A.R. and M.M.C. designed data analyses, aided by J.P.C. and L.F.A.; A.A.R. performed analyses. S.R.B. and J.P.C. performed predictive optimizations. S.R.B. and J.S.S. trained network models, supervised by L.F.A.; A.H.L. collected reaching datasets; A.M. collected rodent dataset, with T.M.J. and A.M.; A.K. collected V1 dataset; A.A.R. and M.M.C. wrote the paper. All authors contributed to editing.

Chapter 3 **Neural trajectories in the supplementary motor area and primary motor cortex exhibit distinct geometries, compatible with different classes of computation**

Comparing neural population trajectories with network-model predictions can link empirical observations with hypothesized computations. We applied this approach to the Supplementary Motor Area (SMA), a region implicated in higher-order motor control. We hypothesized that a computationally important feature of SMA activity is avoidance of 'divergence': neural trajectories that follow the same path before separating. We reasoned that low divergence is necessary if network dynamics guide behavior over long timescales. We compared activity in SMA and primary motor cortex (M1) as monkeys turned a pedal to progress through a virtual environment. Population trajectories in SMA, but not M1, avoided divergence. Network models replicated both this difference in divergence, and the basic features of trajectory geometry: cyclical in M1 and helix-like in SMA. The low-divergence SMA population trajectory accounts for a constellation of diverse single-neuron response properties, and indicates a class of computation that could be performed by SMA but not M1.

Introduction

The supplementary motor area (SMA) is implicated in higher-order aspects of motor control (Eccles, 1982; Penfield & Welch, 1951; Roland, Larsen, Lassen, & Skinhoj, 1980). SMA lesions cause motor neglect (Krainik et al., 2001; Laplane, Talairach, Meininger, Bancaud, & Orgogozo, 1977), unintended utilization (Boccardi, Della Sala, Motto, & Spinnler, 2002), and difficulty performing temporal sequences (Nakamura, Sakai, & Hikosaka, 1998; Shima & Tanji, 1998). Relative to primary motor cortex (M1), SMA activity is less coupled to actions of a specific body part (Tanji & Kurata, 1982; Tanji & Mushiake, 1996). Instead, SMA computations appear related to learned sensory-motor associations (Tanji & Kurata, 1982), reward anticipation (Sohn & Lee, 2007), internal initiation and guidance of movement (Eccles, 1982; Thaler, Chen, Nixon, Stern, & Passingham, 1995), movement timing (Remington, Narain, et al., 2018; Wang et al., 2018), and movement sequencing (Nakamura et al., 1998; Tanji & Shima, 1994). SMA single-neuron responses reflect a variety of task-specific contingencies. For example, in a sequence of three movements, an SMA neuron may burst only when pulling precedes pushing. Another neuron might reliably burst before the third movement regardless of the sequence (Shima & Tanji, 2000). Different response features are then observed in different tasks. For example, during an interval timing task, single-SMA neurons exhibit a mixture of ramping and rhythmic activity (Cadena-Valencia, Garcia-Garibay, Merchant, Jazayeri, & de Lafuente, 2018).

A common thread linking prior studies is that SMA computations are hypothesized to be critical when pending action depends upon internal, abstract, and/or contextual factors. An important challenge is linking these high-level ideas, and accompanying conceptual models (Shima & Tanji, 2000), with network-level computations. What are the natural strategies that a network might use to track contextual information and guide motor output? How would those strategies shape the population response?

Characterizations of population trajectory geometry – the shape traced by activity in state-space – have emerged as one way of linking hypotheses regarding network-level computation with the details of empirical data. We recently characterized M1 activity using a metric of population geometry, 'trajectory tangling', that assesses whether activity could be generated by noise-robust network dynamics. The prediction that trajectory tangling should be low was confirmed across multiple tasks, allowed prediction of neural activity from muscle activity, and explained otherwise-confusing aspects of neural activity. Population trajectory geometry was also explicitly assessed in a recent study of activity in dorsomedial frontal cortex (including part of SMA) during a movement-timing task (Remington, Narain, et al., 2018). Again, trajectory geometry was employed to link the properties of empirical data and hypotheses regarding how networks might perform the proposed computations. In a similar vein, recent studies have linked the shape of neural trajectories to hypotheses regarding underlying neural dynamics (Foster et al., 2014; Remington, Egger, Narain, Wang, & Jazayeri, 2018; Remington, Narain, et al., 2018; Stopfer & Laurent, 1999; Sussillo & Barak, 2013; Sussillo et al., 2015).

Here, our goal is to take existing ideas regarding SMA computations and distill them into a hypothesis regarding the trajectory geometry appropriate for such computations. Our strategy is to test whether that geometry is present in a novel task, and ask whether the hypothesized population-level properties can help understand single-neuron response properties. We employed a recently developed cycling task which adopts some features from sequence / timing tasks, but involves continuous motor output and thus provides a novel perspective on SMA response properties.

A simple metric of trajectory geometry, 'trajectory divergence', distinguished between the population response in M1 and SMA. Simulations confirmed low divergence was necessary for a network to robustly guide action based on internal / contextual information. Furthermore, artificial networks naturally adopted SMA-like or M1-like population geometries when performing

computations that did, or did not, require internally tracking contextual factors. The major features of SMA responses, both at the population and single-neuron levels, could be understood as serving to maintain low divergence. These results show that classes of computation can be linked to abstract properties of trajectory geometry. Doing so can allow one to consider properties that may be conserved across tasks, while also accounting for response features during a specific task.

Results

Task and behavior

We trained two rhesus macaque monkeys to grasp a hand-pedal and cycle through a virtual landscape (Russo et al., 2018) (**Figure** 3.1**A**). Each trial required the monkey to cycle between a pair of targets. The trial began with the monkey stationary on the first target, with the pedal orientation either straight up ('top-start') or straight down ('bottom-start'). After a 1000 ms hold period, the second target appeared. Second-target distance determined the number of revolutions that had to be performed: 1, 2, 4, or 7 cycles. Following a 500-1000 ms randomized delay period, a go-cue (brightening of the second target) was delivered. The monkey then cycled to that target and remained stationary to receive a juice reward. Because targets were separated by an integer number of cycles, the second target was acquired with the same orientation (straight up or down) as for the first target. Landscape color indicated whether forward virtual motion required 'forward' cycling (the hand moved away from the body at the top of the cycle) or 'backward' cycling (the hand moved toward the body at the top of the cycle). Using a block-randomized design, monkeys performed all combinations of two cycling directions, two starting orientations, and four distances. Averages of hand kinematics, muscle activity and neural activity were computed after temporal alignment to account for small trial-by-trial differences in cycling speed (Russo et al., 2018).

Vertical and horizontal hand velocity displayed nearly sinusoidal temporal profiles (**Figure** 3.1**B**). Muscle activity patterns (**Figure** 3.1**C**) were often non-sinusoidal, and initial-cycle and/or terminal-cycle patterns often departed from the middle-cycle pattern (e.g., the initial-cycle response is larger for the example shown). This is an expected consequence of the need to accelerate the arm when starting, and to decelerate the arm when stopping.

Muscle activity and hand kinematics differed in many ways, yet shared the following property: the response when cycling a given distance was a concatenation of an initial-cycle response, some number of middle cycles with a repeating response, and a terminal-cycle response. We refer to the middle cycles as 'steady-state' cycling, reflecting the fact that kinematics and muscle activity repeated across such cycles, both within a distance and across distances. Seven-cycle movements had ~5 steady-state cycles and four-cycle movements had ~2 steady-state cycles. Two- and one-cycle movements involved little or no steady-state cycling. Such structure is reminiscent of a sequence task (e.g., a four-cycle movement follows an ABBC pattern). However, both movement and accompanying muscle activity were continuous; cycle divisions are employed simply for presentation and analysis.

Our motivating hypothesis, derived from prior studies, is that SMA contributes to guidance of action based on internal and/or contextual factors. If so, SMA activity should consistently differentiate between situations that involve different future actions, even if the present motor output is identical. The cycling task produced multiple instances of this scenario, both within and between conditions. Consider the second and fifth cycles of a seven-cycle movement. Present motor output is essentially identical, but in two more cycles the output will differ. A similar situation occurs when comparing the second cycle of seven-cycle and four-cycle movements. A key question is whether these moments of behavioral 'divergence' are paralleled or avoided in the neural response, and whether this differs between M1 and SMA. While this is fundamentally a population-level question, we begin by examining single-neuron responses. Some key features are clear at the single-neuron level, providing a useful foundation for approaching population-level structure.

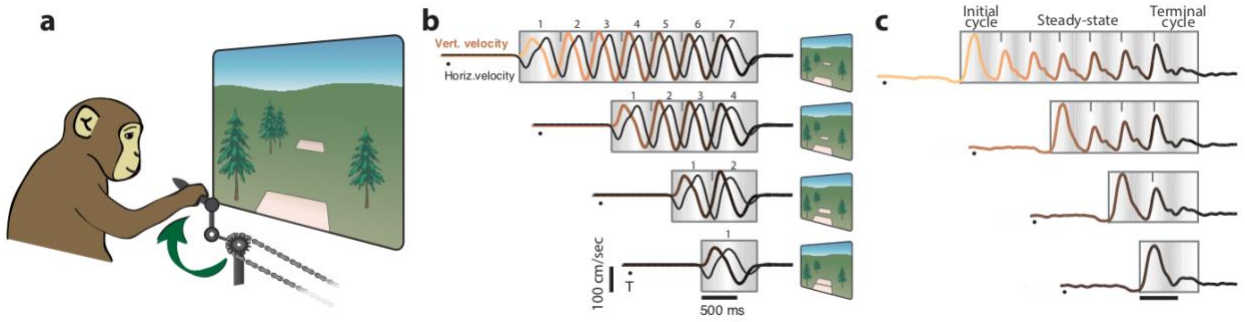

Figure 3.1 Task schematic and behavioral response during cycling

a) Schematic of the task during forward cycling. Monkeys grasped a hand pedal and cycled through a virtual environment for a number of cycles prescribed by target distance.
b) Trial-averaged vertical (colored lines) and horizontal (black lines) hand velocity corresponding to forward, bottom-start conditions: 7-cycle (top row), 4-cycle (second row), 2-cycle (third row) and 1-cycle (bottom row). Coloring from tan to black indicates time with respect to the end of movement. Black dots indicate the time of target appearance onset. Gray box indicates movement period. Shading indicates vertical hand position with light shading indicating the cycle apex. Task schematic panels (right) indicate how target distance is indicated in the virtual environment.
c) Example EMG recording (triceps, monkey D) corresponding to backward, top-starting conditions.

Single-neuron responses

Well-isolated single neurons were recorded sequentially from SMA (77 and 70 recordings for monkeys C and D) and M1 (109 and 103 recordings). Recording locations were guided via MRI landmarks, microstimulation, light touch, and muscle palpation to confirm the trademark properties of each region. M1 recordings included not only sulcal and surface primary motor cortex (M1 proper) but also recordings from the immediately adjacent aspect of dorsal premotor cortex (Russo et al., 2018). Neurons in both SMA and M1 were robustly modulated during cycling. Firing rate modulations (maximum minus minimum rate) averaged 52 and 57 spikes/s for SMA (monkey C and D) and 73 and 64 spikes/s for M1.

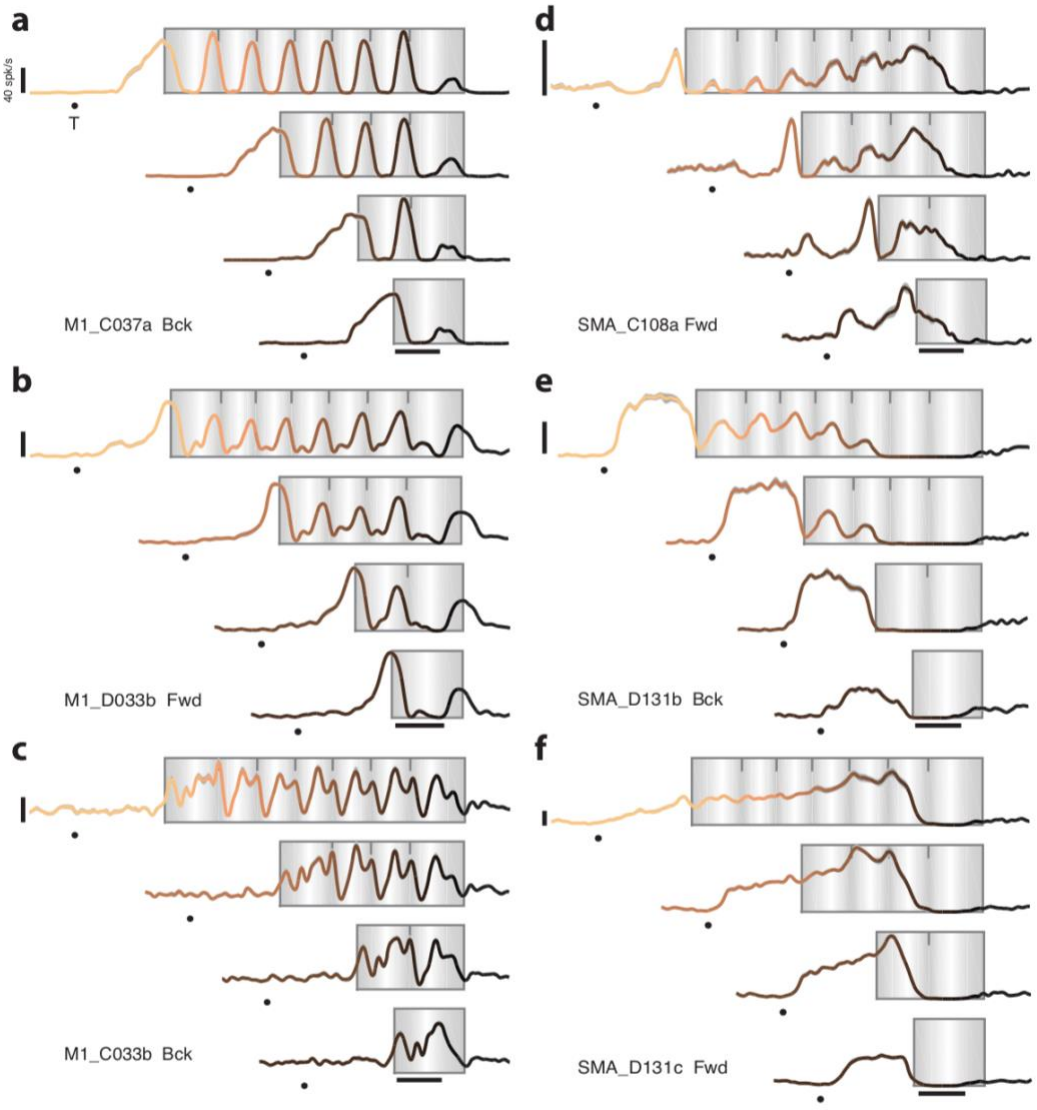

Figure 3.2 Responses of example M1 and SMA neurons

Format as for Figure 1
a-c) Trial-averaged PSTHs from example neurons recorded in M1. Average firing rate was computed across a median of 15 trials/condition per neuron. Neuron names indicate cortical region (M1 or SMA) and monkey (C or D). Data correspond to forward, bottom-starting conditions (Fwd) or backward bottom-starting conditions (Bck). Calibrations are 40 spikes/s.
d-f) Trial-averaged PSTHs from example neurons recorded in SMA. Same format as (a-c)

In M1, single-neuron responses (**Figure** 3.2**A-C**) were typically complex, yet showed two consistent features. First, for a given distance, responses repeated across steady-state cycles. For example, for a seven-cycle movement, the firing rate profile was very similar across cycles 2-6

(Russo et al., 2018). Second, response elements – initial-cycle, steady-state, and terminal-cycle responses – were conserved across distances. Thus, while M1 responses rarely matched patterns of muscle activity or kinematics, they shared the same general structure. Across all distances, responses were essentially a concatenation of an initial-cycle response, a steady-state response, and a terminal-cycle response. Even complex responses that might be mistaken as 'noise' displayed this structure (**Figure** 3.2**C**).

Neurons in SMA (**Figure** 3.2**D-F**) displayed a different set of properties. Responses were typically a mixture of rhythmic and ramp-like features (**Figure** 3.2**D**). As a result, a clear 'steady-state' response was rarely reached. Unlike for M1, the initial-cycle response in SMA often differed across distances (e.g., compare seven-cycle with two-cycle responses). Yet terminal-cycle responses were largely preserved across distances. For example, the response during a four-cycle movement frequently resembled the response during the last four cycles of a seven-cycle movement, but did not match the response during the first four cycles.

Individual-cycle responses are more distinct in SMA

The examples in **Figure** 3.2 illustrate that responses in SMA, but not M1, are distinct when compared across steady-state cycles. Furthermore, when comparing across distances, initial-cycle and steady-state responses tended to be conserved only in M1. To provide a quantitative summary, we compared the response during each cycle with that for every other cycle. We did so both within 7-cycle movements (**Figure** 3.3**A,C**), and between 7-cycle and 4-cycle movements (**Figure** 3.3**B,D**). For each comparison, we computed 'response distance': the root-mean-squared difference in firing rates. Rather than take the mean across all neurons, we used PCA to reduce the dimensionality of the data to twelve. While dimensionality reduction has only a modest impact on

measurements of distance, it provides a useful denoising step. Results were not sensitive to the choice of dimensionality so long as it was high enough to capture a majority of the data variance. All response distances were normalized by the typical intra-cycle distance, then averaged across cycling directions and starting locations. This analysis thus assesses the degree to which responses differ across cycles, relative to the response magnitude of a single cycle.

For M1, responses were similar among all steady-state cycles, resulting in a central dark block. This block is square when comparing within seven-cycle movements and block is rectangular when comparing between seven- and four-cycle movements. Outer rows and columns are lighter; initial- and terminal-cycle responses differed both from one another and from steady-state responses. This analysis confirms that M1 responses involve a distinct initial-cycle response, a repeating steady-state response, and a distinct terminal-cycle response. Essentially identical structure was observed for the muscle populations (top row).

For SMA, the central block of high similarity was largely absent. Instead, distance grew steadily with temporal separation. For example, within a seven-cycle movement, the second-cycle response was modestly different from the third-cycle response, fairly different from the fifth-cycle response, and very different from the seventh-cycle response. As a result, the average normalized distance between steady-state responses was 3.1 times larger for SMA than for M1 for monkey C ($p<0.0001$ via bootstrap), and 6.1 times larger for monkey D ($p<0.0001$). Thus, SMA, unlike M1, showed dissimilar responses across steady-state cycles. This was true ($p<0.0001$ in all cases) both within a distance, and when comparing between distances.

Intriguingly, the 'distance specificity' of SMA responses was reduced when comparing responses aligned to movement's end. In particular, responses were more similar when comparing terminal cycles versus initial cycles (dark entry in lower-right corner versus lighter entry in upper-right

corner). This tendency developed over multiple cycles leading up to movement end. As a result, response distance in SMA was significantly smaller when comparing the last three cycles versus the first three cycles ($p<0.001$, for each monkey, bootstrap). This asymmetry was greater for SMA than for M1 ($p<0.05$ for monkey C and $p<0.0001$ for monkey D).

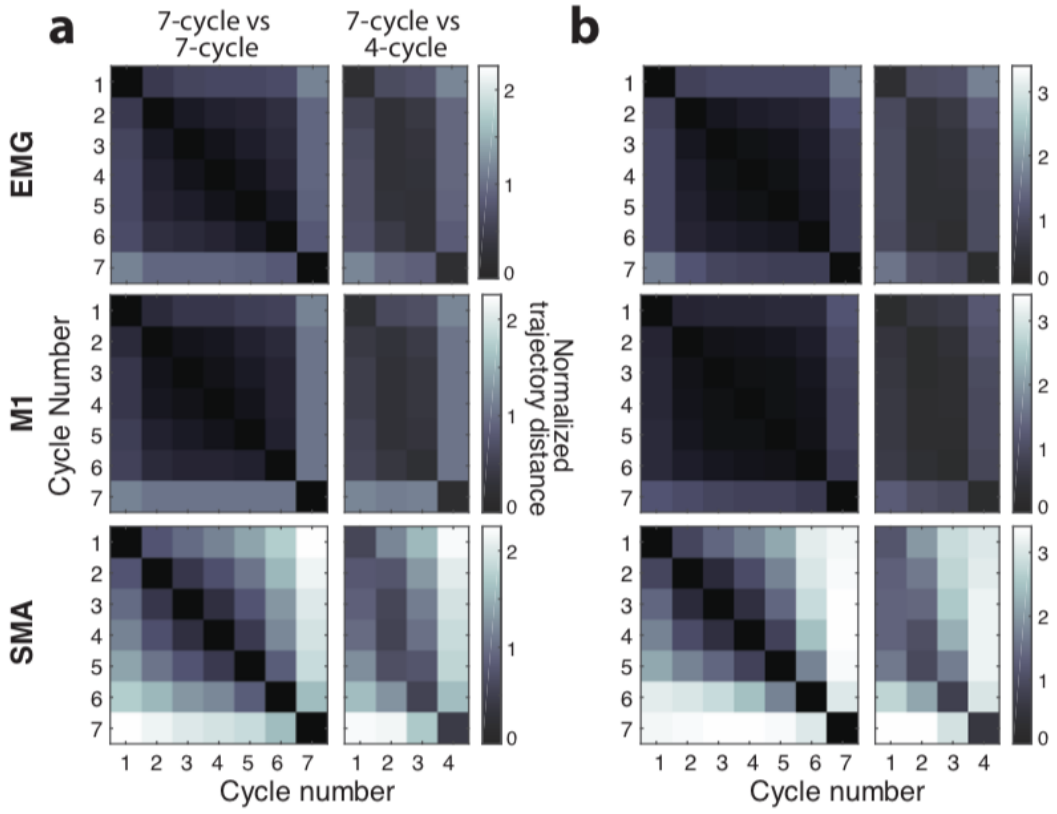

Figure 3.3 Cycle-to-cycle analysis of trajectory distance

a) Normalized trajectory distance was computed in 12 dimensions for muscle activity (top row), M1 (middle row) and SMA (bottom row). Population responses for each cycle were compared within the 7-cycle condition (left column) and between 7-cycle and 4-cycle conditions (right column) and averaged across the four condition types (all combinations of pedaling direction and starting position). Data correspond to monkey C.
b) Same for monkey D.

The cycle- and distance-specificity of SMA responses resembles, in some ways, contingency-specific activity during a movement sequence (Shima & Tanji, 2000) Yet specificity during cycling is manifested rather differently: by responses that evolve continuously, rather than burst at

a key moment. The ramping activity we observed was more reminiscent of pre-movement responses in a timing task (Cadena-Valencia et al., 2018). To further explore the continuous unfolding of activity during cycling, we consider the evolution of the population trajectories.

SMA and M1 display different population trajectories

Using PCA, we projected each population response onto a three-dimensional state-space. Projections are shown for one seven-cycle condition for M1 (**Figure** 3.4**A,B**) and SMA (**Figure** 3.4**C,D**). Traces are shaded light to dark with the passage of time. For the M1 populations, trajectories exited a baseline state just before movement onset, entered a periodic orbit during steady-state cycling, and remained there until settling back to baseline as movement ended. To examine within-cycle structure, we also applied PCA separately for each cycle (bottom of each panel). For M1, this revealed little new; the dominant structure on each cycle was an ellipse, in agreement with what was seen in the projection of the full response.

In SMA, the dominant geometry was quite different, and also more difficult to summarize in three dimensions. We first consider the response for monkey C (**Figure** 3.4**C**). Just before movement onset, the trajectory moved sharply away from baseline (from left to right in the plot). The trajectory then returned to baseline in a rough spiral, with each cycle separated from the last. The population response for monkey D was different in some details (**Figure** 3.4**D**) but it was again the case that a translation separated cycle-specific features.

SMA population trajectories appear to have a 'messier' geometry than M1 trajectories. In particular, cycle-specific loops appear non-elliptical and kinked. Yet it should be stressed that a three-dimensional projection is necessarily a compromise. The view is optimized to capture the largest features in the data; smaller features can be missed or partially captured and distorted.

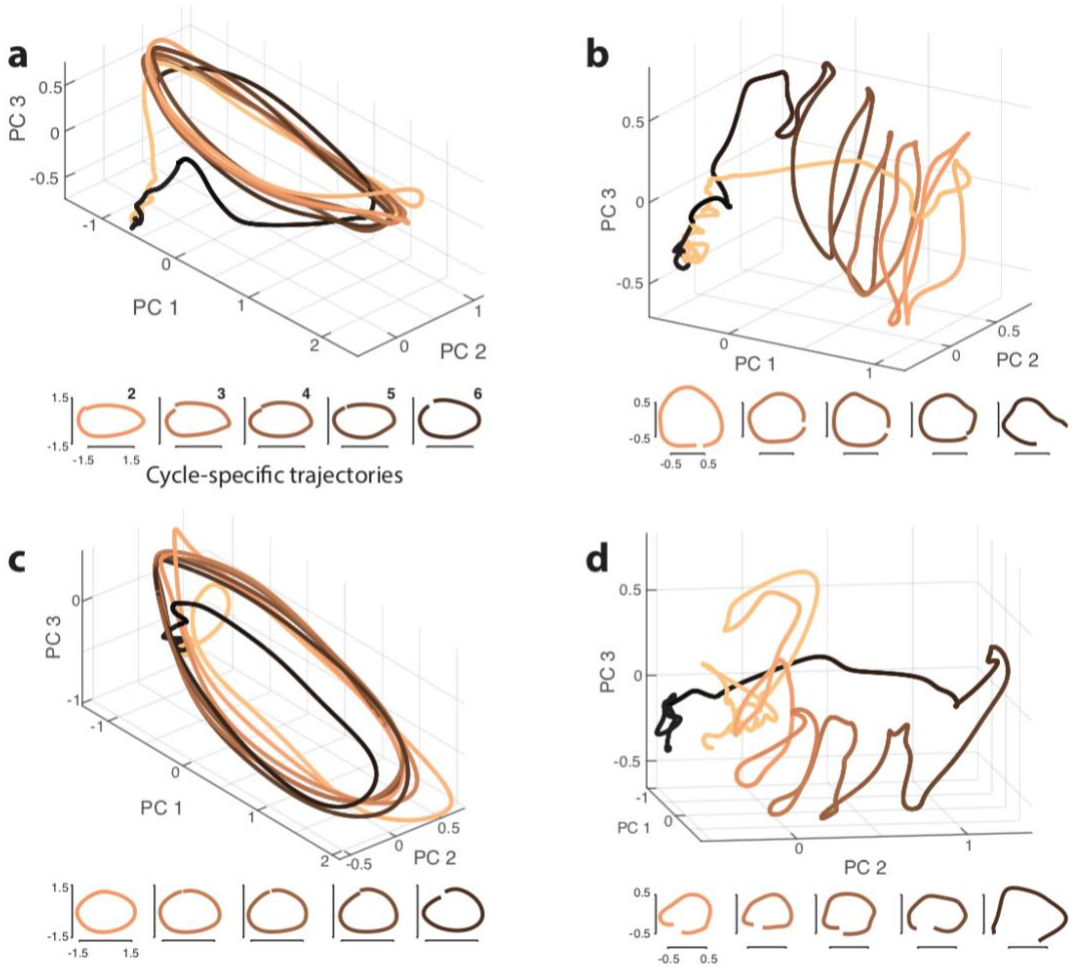

Figure 3.4 Visualization of population structure via PCA

a) M1 population trajectory corresponding to the 7-cycle, forward, bottom-start condition (monkey C). PCs for three-dimensional projection (top) were found using data from all four 7-cycle conditions (all combinations of pedaling direction and starting position) from 200ms before movement began to 200ms after movement ended. All times from this condition were then projected onto the top three PCs. Color from tan to black indicates distance to movement end. Individual cycles 2-6 are visualized by applying PCA separately to each cycle (bottom row). Horizontal axis corresponds to PC 1 for each cycle and vertical axes correspond to PC 2.
b) SMA population trajectory corresponding to the same condition as for (a), (monkey C).

c-d) Same for monkey D, data corresponds to 7-cycle forward, top-start condition.

We thus employed cycle-specific PCs to visualize the shape of the trajectory on each cycle separately. Doing so revealed near-circular trajectories, much as in M1. Thus, individual-cycle orbits are present in SMA, but are a smaller feature relative to the large translation.

In summary, M1 trajectories are dominated by a repeating elliptical orbit while SMA trajectories are better described as helical. Each cycle involves an orbit, but these are separated by a translation. Also, unlike an ideal helix, individual-cycle orbits in SMA occur in somewhat different subspaces, as will be documented below.

The SMA population response occupies different dimensions across cycles

We noted above that elliptical path of individual-cycle SMA trajectories is distorted when projecting all cycles into the same three dimensions, suggesting that trajectories occupy different dimensions on different cycles. To investigate further, we applied PCA separately for each cycle and computed 'subspace overlap': how well PCs derived from one cycle capture trajectories for the other cycles. For example, we found PCs from the responses during cycle one, projected the response during cycle two onto those PCs, and computed the percent variance explained. This was repeated across all combinations. We employed six PCs, which captured most of the variance for a given cycle. Essentially identical results were obtained using more dimensions (**Figure 3.S1**). Variance was normalized so that unity indicates that two cycles occupy the same subspace. For comparison, we also analyzed muscle and M1 trajectories. As in **Figure** 3.3, we compared within seven-cycle movements and between seven- and four-cycle movements.

For the muscles, subspace overlap was high for all comparisons (top row of **Figure** 3.5). Subspace overlap was somewhat lower for M1 (middle row of panels) yet still high. In particular, overlap was high among steady-state cycles, resulting in a central block structure similar to that observed in **Figure** 3.3. The block structure reveals that the subspace found for any of the steady-state cycles overlaps heavily with that for all the other steady-state cycles. For SMA, the central block was

largely absent. Comparing SMA versus M1, the average subspace overlap among steady-state cycles was 0.56 versus 0.83 (monkey C, $p<0.0001$ via bootstrap) and 0.51 versus 0.84 (monkey D, $p<0.0001$). Note that the changing subspace in SMA is not a consequence of the translating trajectory (**Figure** 3.4); a translation changes only where activity is centered, not the subspace in which it resides.

The finding that SMA activity occupies different subspaces across steady-state cycles, both within and between distances, can be thought of as an additional form of selectivity. A possibility explored below is that such selectivity is important when future action depends upon contextual factors. For example, M1 activity and muscle activity are similar on the first three cycles of seven- and four-cycle movements, even though activity will soon be very different in those two cases. Yet SMA responses – including the subspace they occupy – discriminates between those scenarios, potentially eliminating ambiguity regarding what action should come next. For SMA, both response distance (**Figure** 3.3) and subspace overlap (**Figure** 3.5) were quite different when comparing cycles 1-3 of a seven- versus four-cycle movement.

Intriguingly, SMA activity was less selective when comparing situations where there was no need to resolve any ambiguity. For example, the entirety of the remaining movement is identical whether one is on the fifth cycle of a seven-cycle movement or the second cycle of a four-cycle movement. Correspondingly, SMA activity was much less different than when comparing the first three cycles. This can be appreciated by comparing the three-element diagonal starting in the top-left corner with that ending in the bottom-right corner (in both **Figure** 3.3 and **Figure** 3.5). This asymmetry was significantly greater in SMA versus M1 ($p<0.05$ for each monkey, via bootstrap).

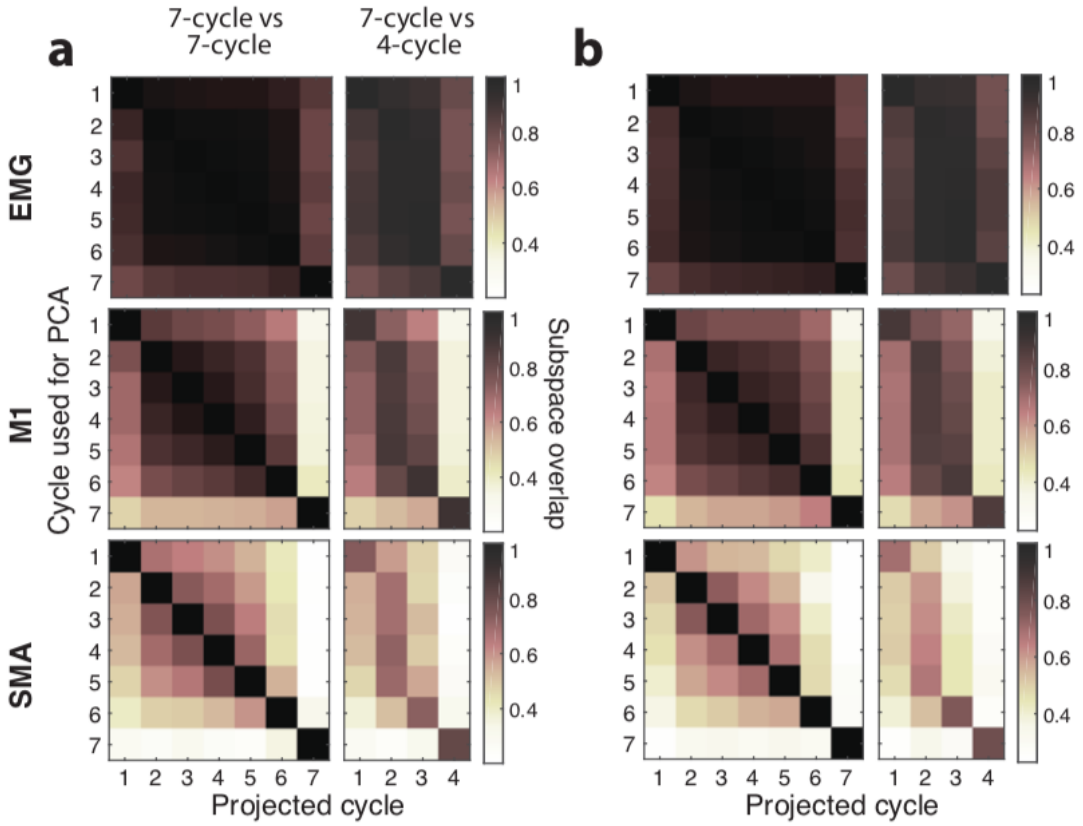

Figure 3.5 Cycle-to-cycle analysis of subspace overlap

c) Subspace overlap was computed in 12 dimensions for muscle activity (top row), M1 (middle row) and SMA (bottom row). Population responses for each cycle of the 7-cycle condition were used to find PCs corresponding to that cycle. Then, population responses for each cycle of the 7-cycle condition (left column) and the 4-cycle condition (right column) were projected onto those PCs and the percent variance was calculated and normalized. Data are averaged across the four condition types (all combinations of pedaling direction and starting position). Data correspond to monkey C.
d) Same for monkey D.

Population trajectories adopted by artificial networks

SMA is hypothesized to guide action based on internal / contextual considerations. For practical purposes, we define 'motor context' as information that is important for guiding future movement, but may not impact present motor output. Contextual information may be remembered (e.g., "I am performing a particular sequence"), internally estimated ("it has been 800 ms since the last button press"), or derived from abstract cues ("this fixation-point color means I must reach quickly when the target appears").

In the cycling task, salient contextual information arrives when the target appears, specifying the number of cycles to be produced. The current motor context (how many cycles remain) can then be updated throughout the movement, based on both visual cues and internal knowledge of the number of cycles already produced. To ask how contextual information might be reflected in population trajectories, we trained artificial recurrent networks that did, or did not, need to internally track motor context.

We considered highly simplified inputs (pulses at specific times) and outputs (pure sinusoids lasting four or seven cycles). We trained two families of recurrent networks. A family of 'context-naïve' networks received one input pulse, indicating that output generation should begin, and a different input pulse, indicating that output should be terminated. Initiating and terminating inputs were separated by four or seven cycles, corresponding to the desired output. Thus, context-naïve networks had no information regarding context until the arrival of the second input. Similarly, such networks had no need to track context as the key information was provided at the critical moment. A family of 'context-tracking' networks, received only an initiating input. For context-tracking networks only, this input pulse differed depending on whether a four- or seven-cycle output should be produced. These networks then had to generate a sinusoid with the appropriate number of

cycles, and terminate appropriately with no further external guidance. For each family, we trained 500 networks that differed in their initial connection weights (*Methods*).

The two network families learned qualitatively different solutions involving population trajectories with different geometries (**Figure** 3.6**A,B**). Context-naïve networks employed a limit cycle. The initiating input caused the network trajectory to enter an orbit, and the terminating input prompted the trajectory to return to baseline. This solution was not enforced but emerged naturally. There was network-to-network variation in how quickly activity settled into the limit cycle (**Figure 3.S2**) but essentially all networks that succeeded in performing the task employed a version of this strategy.

Context-tracking networks utilized population trajectories that were more helical, with the trajectory on each cycle being separated from the others by an overall translation. While there was network-to-network variability in the exact learned trajectory (**Figure 3.S3**), all successful context-tracking networks employed some form of helical or spiral trajectory. This solution is intuitive: context-tracking networks do not have the luxury of following a repeating orbit. If they did, information regarding context would be lost, and the network would have no way of 'knowing' when to cease producing the output.

For context-tracking networks, trajectories could also occupy somewhat different subspaces on different cycles. When plotting in three dimensions, this geometry resulted in individual-cycle trajectories of seemingly different magnitude (first and third examples in **Figure** 3.6**B**). As with the helical structure, this geometry creates separation between individual-cycle trajectories. There was considerable variation in the degree to which this strategy was employed. Some context-tracking networks used nearly identical subspaces for every cycle while other context-tracking

networks used quite different subspaces for each cycle. In contrast, context-naïve networks never employed this strategy; the same limit cycle was followed across middle cycles.

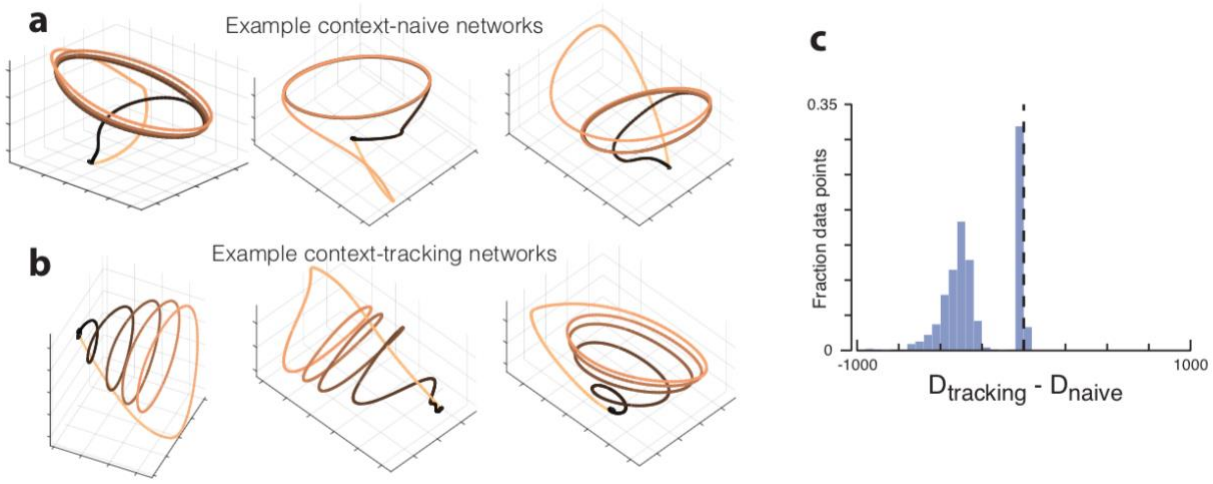

Figure 3.6 Analysis of trajectory geometry in context-naïve and context-tracking networks

a) Network trajectories for three example context-naïve networks during the 4-cycle condition. For all examples, lower left axes correspond to PC 1, lower right axes correspond to PC2 and vertical axes correspond to PC3.
b) Same for three example context-tracking networks.
c) 5000 random pairs of context-tracking networks and context-naïve networks were compared.

For each pair of networks, the difference between trajectory divergence in context-tracking ($D_{tracking}$) and in context-naïve (D_{naive}) was computed for each time point. The resulting distribution is plotted cumulatively across network pairs as a histogram. Vertical dashed line indicates zero. For almost all time points, trajectory divergence was lower in the context-tracking than in the context-naïve networks as indicated by the leftward shift of the distribution.

The population geometry adopted by context-naïve and context-tracking networks bears obvious similarities to the empirical population geometry in M1 and SMA, respectively. That said, we stress that neither family is intended to faithfully model the corresponding area. Furthermore, a number of reasonable alternative modeling choices exist. For example, rather than asking context-tracking networks to track progress using internal dynamics alone, one can provide a ramping input

that does so. Interestingly, context-tracking networks trained in the presence / absence of ramps employed very similar population trajectories (**Figure 3.S4**). The slow translation that produces helical structure is a useful computational tool – one that networks produced on their own if needed but were also content to inherit from upstream sources. For these reasons, we focus not on the details of the network trajectories, but rather on the geometric features that differentiate context-tracking from context-naïve network trajectories, and that might similarly differentiate M1 and SMA population trajectories.

Trajectory divergence

We developed a metric of trajectory geometry that assesses whether population activity reflects motor context (as defined above) in a way that could guide future action. We define 'trajectory divergence' as two trajectories (or portions of the same trajectory) passing through a similar neural state but eventually separating to follow different future trajectories. High divergence indicates an absence of contextual information, because two situations that are different (in the long term) are not distinguished by the neural state. Trajectory divergence differs from trajectory tangling (Russo et al., 2018), which was very low in both SMA and M1 (**Figure 3.S5**). Trajectory tangling assesses whether trajectories are consistent with a locally smooth flow-field. Trajectory divergence assesses whether similar paths eventually separate, smoothly or otherwise. A trajectory can have low tangling but high divergence, or vice versa (**Figure 3.S6**).

The results of Figures 3.3-3.5 suggest that trajectory divergence may be low only in SMA. Because M1 trajectories repeat, they pass through similar states multiple times both within a movement and between distances. The more helix-like SMA trajectories may eliminate such points, although this is difficult to discern in three dimensions where trajectories often cross. Furthermore, it is critical

to assess whether trajectory divergence remains low when comparing across distances (one, two, four and seven).

To construct a quantitative metric that can summarize the geometry of multiple high-dimensional trajectories, we consider times t and t', associated population states X_t and $X_{t'}$, and future population states $X_{t+\Delta}$ and $X_{t'+\Delta}$. We consider all possible pairings of t and t'. For example, t and t' might occur during different cycles of the same movement, or during different movement distances. We compute the ratio $\frac{\|X_{t+\Delta}-X_{t'+\Delta}\|^2}{\|X_t-X_{t'}\|^2+\propto}$, which becomes large if $X_{t+\Delta}$ differs from $X_{t'+\Delta}$ despite X_t and $X_{t'}$ being similar. The constant $\propto$ is small and proportional to the variance of X, and functions to prevent hyperbolic growth. For a given time t, this ratio will be small for most values of t', simply because the typical difference between two random states is sizeable.

Given that the difference between two random states is typically sizeable, the above ratio will be small for most values of t'. As we are interested in whether the ratio ever becomes large, we take the maximum, and define divergence for time t as:

$$\mathrm{D}(t) = \max_{t',\Delta} \frac{\|X_{t+\Delta} - X_{t'+\Delta}\|^2}{\|X_t - X_{t'}\|^2+\propto}$$

Equation 3.1

We consider only positive values of Δ. Thus, $\mathrm{D}(t)$ becomes large if similar trajectories diverge but not if dissimilar trajectories converge. Divergence was assessed using 12 dimensions. Results were similar for all reasonable choices of dimensionality.

Application to simulated data confirmed that $\mathrm{D}(t)$ differentiated between context-tracking and context-naïve networks. To provide a quantitative summary, we considered pairs of networks, one context-tracking and one context-naïve, and at each time computed the difference in the

corresponding values of $D(t)$ (**Figure 3.S6C**). Both context-tracking and context-naïve trajectories contained many moments where divergence was low, resulting in a narrow peak near zero. However, context-naïve trajectories (but not context-tracking trajectories) also contained moments where divergence was high, yielding a large set of negative values. The distribution of differences in **Figure 3.S6C** consider all times for 5000 network pairs. We also asked, for every pair, whether the context-naïve network had lower average trajectory divergence. This was true for all pairs, despite the variety of trajectories adopted by individual networks (**Figure 3.S2** and **Figure 3.S3**). This underscores a key advantage of the divergence metric: it assesses a computationally relevant aspect of trajectory geometry in a manner that abstracts away from the details of particular trajectories.

Trajectory divergence is lowest for SMA

For each time t, we plotted SMA versus M1 divergence (**Figure** 3.7**A,B**). Divergence was almost always lower for SMA trajectories. We computed distributions of the difference in divergence, at matched times, between SMA and M1 (**Figure** 3.7**C,D**). There was a narrow peak at zero (times where divergence was low for both) and a set of negative values (indicating multiple times with lower divergence for SMA). Strongly positive values (lower divergence for M1) were absent (monkey C) or very rare (monkey D). It was also the case that divergence was much lower in SMA than in the muscle populations (**Figure 3.S7**).

Thus, trajectory divergence for SMA and M1 differed in much the same way as it had for context-tracking and context-naïve networks (compare **Figure** 3.7**C,D** with **Figure 3.S6C**). The overall scale of divergence values was greater for the networks; this is expected as simulated trajectories can repeat almost perfectly, yielding very small values of the denominator of equation 1. The low

divergence of SMA trajectories relates to population-level features documented in Figure 3.3-3.5. However, consistently low divergence could not have been confidently inferred from those analyses, for three reasons. First, the trajectories in **Figure** 3.4 show multiple instances where the neural state appears similar at different moments. If this were true in all dimensions, it would lead to high trajectory divergence. This highlights that it is critical to assess divergence across enough dimensions to capture most of the structure of the responses. Second, the fact that individual-cycle trajectories trace different paths (**Figure** 3.3) does not imply that those paths don't come near one another. Paths can be different but still cross. Finally, we wished to infer whether divergence was low when considering not only all pairs of times (*t* and *t'*) within a condition, but all possible pairings between distances.

The above underscores a useful property of trajectory divergence as a metric: it summarizes a property that is expressed via a variety of response features, some of which might otherwise seem unrelated. Because trajectory divergence abstracts away from the details of the specific trajectories, it is readily applied in new situations. For example, the present task involved not just different distances, but also different cycling directions and different starting positions. The latter is particularly relevant, because movements ended at the same position (top versus bottom) that they started. Thus, how a movement will end depends on information present at the movement's beginning. One could ask whether SMA responses keep track of such information by assessing 'starting-position-tuning' in a variety of ways, following the model of **Figure** 3.3 and **Figure** 3.5. However, it is simpler, and more relevant to the hypothesis being considered, to ask whether divergence remains low when comparisons are made across all conditions, including starting positions. This was indeed the case (**Figure 3.S8**). This reveals the utility of a metric that focuses

on a computationally relevant property, regardless of how that property is realized in a particular task.

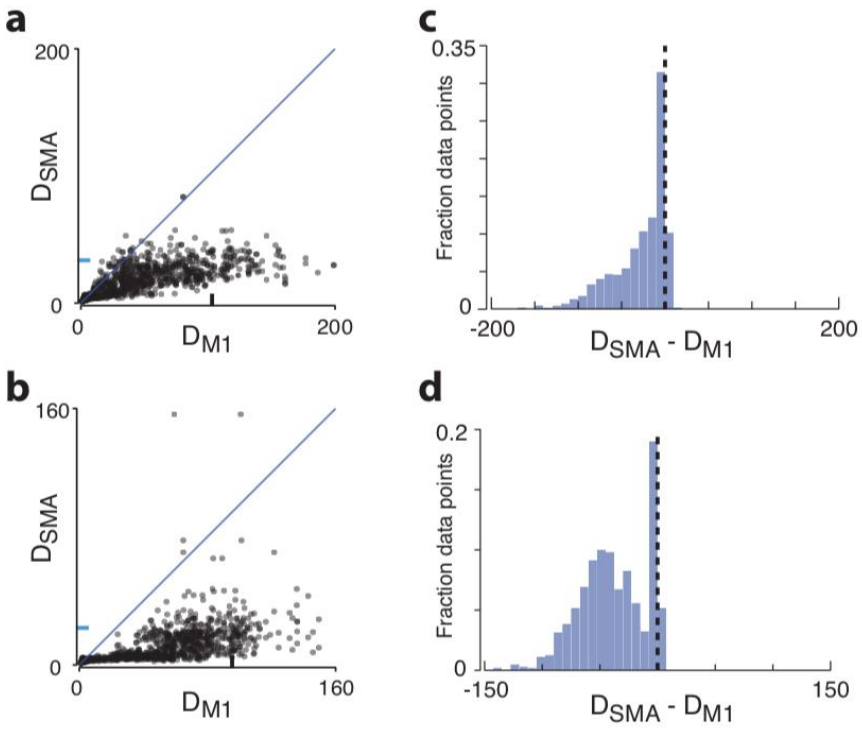

Figure 3.7 Trajectory divergence in M1 and SMA

a) SMA versus M1 trajectory divergence (monkey C) is plotted for all time points (black dots). Blue tick mark along the vertical axis denotes the 90th percentile trajectory divergence for SMA. Black tick mark along the horizontal axis denotes 90th percentile trajectory divergence for M1.
b) Same for monkey D
c) For each time point, the difference between the trajectory divergence in SMA and in M1 was computed (monkey C). Trajectory divergence is almost always lower in SMA than in M1 as indicated by very little mass of the distribution to the right of zero (vertical dashed line).
d) Same for monkey D

Computational implications of trajectory divergence

We assessed trajectory divergence because of its expected computational implications. A network with a high-divergence trajectory may accurately and robustly generate its output on short timescales. Yet unless guided by external inputs at key moments, such a network may be susceptible to errors on longer timescales. For example, if a trajectory approximately repeats, a likely error would be the generation of extra cycles, or the inappropriate skipping of a cycle.

The simulations above support this idea: when networks could not depend on a second stopping pulse, they adopted low-divergence trajectories. However, on its own this does not necessarily imply that a high-divergence solution would fail. To test this, we used an atypical training approach that enforced an internal network trajectory, as opposed to the usual approach of training a target output. We trained networks to exactly follow the M1 trajectory recorded during a four-cycle movement, without any input indicating when to stop (**Figure** 3.8**A**). To ensure that the solutions found were not overly 'delicate', networks were trained in the presence of additive noise. For each monkey, we trained forty networks: ten for each of the four-cycle conditions. Networks were able to reproduce the cyclic portion of the M1 trajectory. However, without the benefit of a stopping pulse, networks failed to consistently follow the end of the trajectory. For example, networks sometimes erroneously produced extra cycles (**Figure** 3.8**B**) or skipped cycles and stopped early (**C**).

We also trained networks to follow the empirical SMA trajectories. Those trajectories contained both a rhythmic component, and lower-frequency 'ramping' signals (**Figure** 3.8**D**) related to the translation visible in **Figure** 3.4**C,D**. In contrast to the high-divergence M1 trajectories, which were never consistently followed for the full trajectory, the majority of network initializations resulted in good solutions where the low-divergence SMA trajectory was successfully followed

from beginning to end. Thus, in the absence of a stopping pulse, the empirical SMA trajectories could be produced, and could end reliably, in a way that the empirical M1 trajectories could not.

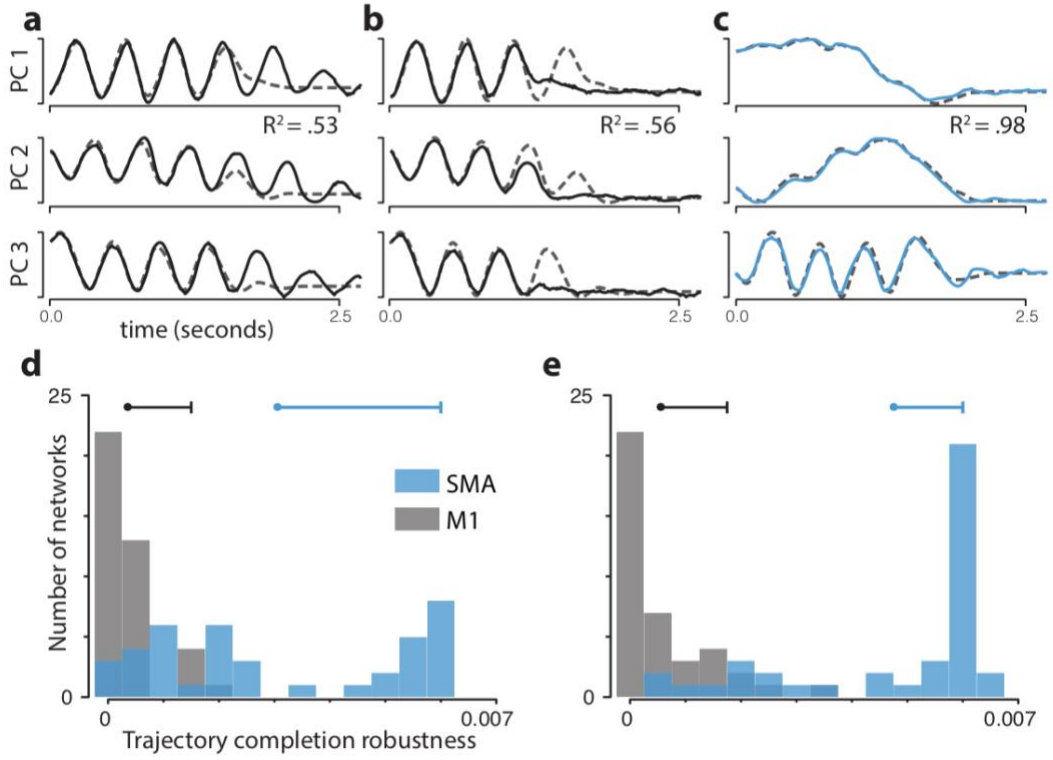

Figure 3.8 Low trajectory divergence allows networks to complete trajectories in the presence of noise

a-b) Two example network trajectories (black lines) constrained to follow M1 target trajectory (dashed gray lines) during a 4-cycle condition. These networks were less noise robust than those following the SMA target trajectory and tended to produce too many cycles (a) or abort early (b).
c) An example network trajectory (blue lines) constrained to follow SMA target trajectory (dashed gray lines).
d) Trajectory completion robustness of networks constrained to follow either the M1 (gray) or SMA (blue) population trajectories during the 4-cycle conditions (monkey C). 10 networks were trained for each of the four 4-cycle conditions (all combinations of starting position and pedaling direction) for each region. Dots correspond the mean of each distribution and rightward-going hash corresponds to the 90th percentiles.
e) Same for monkey D.

Discussion

Prior studies argue that SMA computations are critical when future action would be ambiguous without the contribution of internal or abstract factors. Our goal was to translate this conceptual hypothesis into a hypothesis regarding the geometry of population activity. The hope was that such a hypothesis, while derived from prior work, would generalize well and describe population activity in a novel task. Furthermore, we hoped that population geometry might provide a link between network models and features of the empirical data. Specifically, the need for a particular population geometry, given a hypothesized class of computation, might provide a cohesive explanation for diverse features of neural responses, at both single-neuron and population levels.

We employed our recently developed cycling task both because it has proved useful in characterizing population geometry in M1, and because it produces multiple instances of behavioral divergence: situations with the same present motor output but different future motor outputs. The cycling task is neither a sequence task nor a timing task, yet shares commonalities with both paradigms. Consistent with this, there were both differences and commonalities between single-neuron responses during cycling and during other tasks. The ramping firing rates we observed resemble those seen in timing tasks(Cadena-Valencia et al., 2018), although the activity ramps we observed were often non-monotonic. We also observed cycle-specific responses – e.g., different firing rates across steady-state cycles – which may be thought of as a form of sequence selectivity. However, cycle-selectivity was produced not by response bursts tied to a particular contingency (Shima & Tanji, 2000), but by a combination of ramping and cyclic activity, with different subspaces being occupied on different cycles. Distance-selectivity (e.g., different responses when starting a four- versus seven-cycle movement) can also be seen as a form of sequence selectivity. Yet distance-selectivity was not equally present across all comparisons; it was pronounced when comparing situations where future motor output would be different.

These diverse properties can be understood given a simple hypothesis regarding population trajectories: that they should avoid trajectory divergence. That hypothesis embodies an essential component of prior ideas: the ability to guide action depending on internal / contextual factors implies that activity, somewhere in the brain, differentiates between situations that are the same now but will soon become different. In SMA, population trajectories traced out a roughly helical geometry, which naturally avoids divergence while still reflecting the rhythmic nature of the task. The rhythmic features of the SMA responses had a similar shape on every cycle but occupied different subspaces. Again, this can be seen as a neural 'strategy' for avoiding trajectory divergence by differentiating among situations that have the same present motor output but different future outputs.

Simulations confirmed that divergence was naturally high in networks that did not have to internally track context. Context-naïve networks displayed elliptical population trajectories that resembled the dominant structure in motor cortex (but of course lacked the finer-grained structure related to encoding of muscle activity). Conversely, divergence was low in networks that had to track context. Context-tracking networks displayed helical population trajectories that resembled a simplified version of the SMA trajectories. Although the helical structure was universal across such networks, there was also variability in the exact solution. For some networks, low-divergence was achieved solely through the translation that separated cycles along the long axis, while in other networks different cycles occupied somewhat different subspaces (as in the neural data, but typically to a lesser degree). This underscores the value of a metric such as trajectory divergence, which can abstract away from solution-specific features and indicate whether a trajectory is appropriate for a particular computation.

Thus, population geometry provides a bridge between conceptual ideas regarding the class of computation being performed, and the solutions adopted by networks (real or simulated) that may be performing those computations. We recently employed a different metric of trajectory geometry, trajectory tangling, when examining the population response in motor cortex. Trajectory tangling revealed a large difference between

M1 trajectories and the downstream muscle population trajectories. That difference – much lower tangling in M1 – was apparent across task and species, and helped explain seemingly paradoxical features of M1 activity. We also found that trajectory tangling was much lower in M1 than in sensory cortical areas. Low tangling is necessary for a network to robustly generate an output via internal dynamics. The presence of low tangling in M1, but not sensory areas, argues that M1 activity is structured to allow robust pattern generation. In the present study, we found that trajectory tangling was similarly low in both SMA and M1, consistent with activity both areas being strongly shaped by internal dynamics. However, the nature of the computation performed by those internal dynamics is likely very different, given the finding that trajectory divergence was low only in SMA. Only in SMA is the population trajectory consistent with guidance of movement based on contextual information. While the M1 population trajectory is sufficient for robust pattern generation –due to trajectory tangling being low – this is true only if M1 receives occasional guiding inputs (which could of course come from SMA). This is underscored by context-naïve networks, which employed strong internal dynamics to generate their output, but still depended on an input to terminate the cycling pattern by moving the population state from a limit cycle to a stable baseline.

A growing number of studies have used examinations of the shape and nature of population activity to evaluate hypotheses regarding computation. Most commonly, such studies quantify specific features, or 'motifs' that relate to how a network might perform the task of interest [many refs]. This will almost certainly remain an essential strategy. Yet one may often wish to supplement this strategy with metrics of population geometry, such as trajectory tangling and divergence, that can abstract away properties that may be preserved across a class of computation, regardless of the particular instantiation. The present work argues that the property of low-divergence may help provide a unifying understanding of the diversity of SMA response properties both within and between tasks. That said, an important caveat is that low divergence in SMA still needs to be confirmed for sequence and timing tasks. The known response properties during these tasks – e.g. the various contingent-specific responses during movement sequence – strongly suggest low trajectory divergence (indeed, that was part of the motivation for the present

experiments). Yet confirming this directly remains an important goal for future research. If trajectory divergence is indeed consistently low in SMA across the relevant set of tasks, this could provide a unifying way for classifying the type of computation where SMA makes an essential contribution.

Methods

Main experimental datasets

Subjects were two adult male rhesus macaques (monkeys C and D). Animal protocols were approved by the Columbia University Institutional Animal Care and Use Committee. Experiments were controlled and data collected under computer control (Speedgoat Real-time Target Machine). During experiments, monkeys sat in a customized chair with the head restrained via a surgical implant. Stimuli were displayed on a monitor in front of the monkey. A tube dispensed juice rewards. The left arm was loosely restrained using a tube and a cloth sling. With their right arm, monkeys manipulated a pedal-like device. The device consisted of a cylindrical rotating grip (the pedal), attached to a crank-arm, which rotated upon a main axel. That axel was connected to a motor and a rotary encoder that reported angular position with 1/8000 cycle precision. In real time, information about angular position and its derivatives was used to provide virtual mass and viscosity, with the desired forces delivered by the motor. The delay between encoder measurement and force production was 1 ms.

Horizontal and vertical hand position were computed based on angular position and the length of the crank-arm (64 mm). To minimize extraneous movement, the right wrist rested in a brace attached to the hand pedal. The motion of the pedal was thus almost entirely driven by the shoulder and elbow, with the wrist moving only slightly to maintain a comfortable posture. Wrist movements were monitored via two reflective spheres attached to the brace, which were tracked optically (Polaris system; Northern Digital, Waterloo, Ontario, Canada) and used to calculate wrist angle. The small wrist movements were highly stereotyped across cycles. Visual monitoring (via infrared camera) confirmed the same was true of the arm as a whole (*e.g.*, the lateral position of

the elbow was quite stereotyped across revolutions). Eye position and pupil dilation were monitored but are not analyzed here.

Task

Monkeys performed the 'cycling task' as described previously (Russo et al., 2018). The monitor displayed a virtual landscape, generated by the Unity engine (Unity Technologies, San Francisco). Surface texture and landmarks provided visual cues regarding movement through the landscape along a linear 'track'. One rotation of the pedal produced one arbitrary unit of movement. Targets on the track indicated where the monkey should stop for juice reward.

Each trial of the task began with the appearance of an initial target. To begin the trial, the monkey had to cycle to and to acquire the initial target (*i.e.*, stop on it and remain stationary) within 5 seconds. Acquisition of the initial target yielded a small reward. After a 1000 ms hold period, the final target appeared at a prescribed distance. Following a randomized (500-1000 ms) delay period, a go-cue (brightening of the final target) was given. The monkey then had to cycle to acquire the final target. After remaining stationary in the final target for 1500 ms, the monkey received a large reward.

The full task included 20 conditions distinguishable by final target distance (half-, one-, two-, four-, and seven-cycles), initial starting position (top or bottom of the cycle), and cycling direction. For all analyses here, we excluded half-cycle conditions which were brief and more similar to reaching than to the sequence-like movements studied here. Salient visual cues (landscape color) indicated whether cycling must be 'forward' (the hand moved away from the body at the top of the cycle) or 'backward' (the hand moved toward the body at the top of the cycle) to produce forward virtual progress. Trials were blocked into forward and backward cycling. Other trials types were

interleaved using a block-randomized design. For each neural / muscle recording, we collected a median of 15 trials / condition for both monkeys.

Neural recordings during cycling

After initial training, we performed a sterile surgery during which monkeys were implanted with a head restraint and recording cylinders. Initial cylinders (Crist Instruments, Hagerstown, MD) were placed surface normal to the cortex and centered over the border between caudal PMd and primary motor cortex (for M1 recordings). After recording in M1, we performed a second sterile surgery to move the cylinders in order to record from the SMA. Cylinders were angled ~20 degrees to avoid the central sulcus vein and centered over the SMA as determined from a previous magnetic resonance imaging scan. To perform recordings, the skull within the cylinder was left intact and covered with a thin layer of dental acrylic. Electrodes were introduced through small (3.5 mm diameter) burr holes drilled by hand through the acrylic and skull, under ketamine / xylazine anesthesia. Neural recordings were made using conventional single electrodes (Frederick Haer Company, Bowdoinham, ME) driven by a hydraulic microdrive (David Kopf Instruments, Tujunga, CA).

Recording locations were guided via microstimulation, light touch, and muscle palpation protocols to confirm the trademark properties of each region. For motor cortex, recordings were made from primary motor cortex (both surface and sulcal) and the adjacent (caudal) aspect of dorsal premotor cortex. These recordings are analyzed together as a single motor cortex population. All recordings were restricted to regions where microstimulation elicited responses in shoulder, upper arm, chest and forearm.

Neural signals were amplified, filtered, and manually sorted using Blackrock Microsystems hardware (Digital Hub and 128-channel Neural Signal Processor). A total of 380 isolations were made across the two monkeys. On each trial, the spikes of the recorded neuron were filtered with a Gaussian (25 ms standard deviation; SD) to produce an estimate of firing rate versus time. These were then averaged across trials and aligned as described previously (Russo et al., 2018).

EMG recordings

Intra-muscular EMG was recorded from the major muscles of the arm, shoulder, and chest using percutaneous pairs of hook-wire electrodes (30mm x 27 gauge, Natus Neurology) inserted ~1 cm into the belly of the muscle for the duration of single recording sessions. Electrode voltages were amplified, bandpass filtered (10-500 Hz) and digitized at 1000 Hz. To ensure that recordings were of high quality, signals were visualized on an oscilloscope throughout the duration of the recording session. Recordings were aborted if they contained significant movement artifact or weak signal. That muscle was then re-recorded later. Offline, EMG records were high-pass filtered at 40 Hz and rectified. Finally, EMG records were smoothed with a Gaussian (25 ms SD, same as neural data) and trial averaged (see below). Recordings were made from the following muscles: the three heads of the *deltoid*, the two heads of the *biceps brachii*, the three heads of the *triceps brachii*, *trapezius*, *latissimus dorsi*, *pectoralis*, *brachioradialis*, *extensor carpi ulnaris*, *extensor carpi radialis*, *flexor carpi ulnaris*, *flexor carpi radialis*, and *pronator*. Recordings were made from 1-8 muscles at a time, on separate days from neural recordings. We often made multiple recordings for a given muscle, especially those that we have previously noted can display responses that vary with recording location (*e.g.*, the *deltoid*).

Preprocessing and PCA

Because PCA seeks to capture variance, it can be disproportionately influenced by differences in firing rate range (*e.g.*, a neuron with a range of 100 spikes/s has 25 times the variance of a similar neuron with a range of 20 spikes/s). This concern is larger still for EMG, where the scale is arbitrary and can differ greatly between recordings. The response of each neuron / muscle was thus normalized prior to application of PCA. EMG data were fully normalized: $response \coloneqq response/range(response)$, where the range is taken across all recorded times and conditions. Neural data were 'soft' normalized: $response \coloneqq response/(range(response) + 5)$. We standardly (Churchland et al., 2012; Russo et al., 2018; Seely et al., 2016) use soft normalization to balance the desire for PCA to explain the responses of all neurons with the desire that weak responses not contribute on an equal footing with robust responses. In practice, nearly all neurons had high firing rate ranges during cycling, making soft normalization nearly identical to full normalization.

Following preprocessing, neural data were formatted as a 'full-dimensional' matrix, X^{full}, of size $n \times t$, where n is the number of neurons (or muscles) and t indexes across all analyzed times. Unless otherwise specified, analyzed times were from 100 ms before movement onset to 100 ms after movement offset, for all conditions. Because PCA operates on mean-centered data, we mean-centered X^{full} so that every row had a mean value of zero.

PCA was used to find X, a reduced-dimensional version of X^{full} with the property that $X^{full} \approx VX$, where V are the PCs ('dimensions' upon which the data are projected). For most analyses, we employed twelve PCs, such that X was of size $12 \times t$. Twelve PCs captured 77% and 78%

(monkey C and D) of the M1 neural data variance, 71% and 77% of the SMA neural data variance, and 94% and 97% of the muscle data variance.

Cycle-to-cycle trajectory distance and subspace overlap

We began quantifying the population structure of neural data by comparing the difference in trajectory responses between pairs of cycles. First, neural data were reduced to 12 dimensions as described above. We employed position-dependent temporal alignment (Russo et al., 2018) on each cycle to ensure differences were not simply due to small variations in hand position or cycle duration. We then computed the root-mean squared difference between trajectories corresponding to pairs of cycles. For each of the four condition types (both cycling directions and starting positions), differences were normalized by response variance within the fourth cycle of the seven-cycle movement corresponding to the same condition type. Difference matrices (Figure 3) were averaged across condition types: both cycling directions and starting positions.

Neural population structure was also quantified by measuring cycle-to-cycle subspace overlap. Here, PCA was applied separately to each cycle in each 7-cycle condition. Then, data from each cycle of the 7-cycle and 4-cycle condition of the same condition type were projected onto those twelve PCs. The amount of variance in this projected data was then normalized by the amount of variance in twelve dimensions of this data when projected into its native space (i.e. the space found when PCA is applied to that data) to yield subspace overlap. This normalization ensures that subspace overlap ranges from 0 to 1, where 0 indicates that two cycles utilize fully orthogonal spaces and 1 indicates that two cycles occupy the identical space.

Bootstrap analyses were performed by resampling all neurons with replacement before the dimensionality reduction step. Resampling was performed 1000 times and analyses were then

performed on these bootstrapped datasets. For analyses that compared SMA and M1, comparison was performed across all pairs of SMA and M1 bootstrapped datasets resulting in 1 million comparisons.

Trajectory Divergence

Visualization and quantification of the population geometry indicate that SMA is characterized by low trajectory divergence. High divergence is defined as two trajectories passing through the same (or nearly the same) neural state, but eventually diverging to follow very different future trajectories. Trajectory divergence was measured on X, the PCA reduced data matrix (described above). Importantly, trajectory divergence was measured on times well after the target stimuli appeared (which occurred at least 500ms before movement onset). If divergence were measured on times that included pre-movement baseline activity, divergence would trivially become high when the context-distinguishing input arrived to the system.

To compute a general metric of trajectory divergence, we considered times t and t', which could occur within the same condition or in different conditions of the same condition type (e.g. seven-cycle and four-cycle for the forward, top-start condition type). Divergence, D, for each pair of times was defined as:

$$D(t, t', \Delta) = \frac{\|X_{t+\Delta} - X_{t'+\Delta}\|^2}{\|X_t - X_{t'}\|^2 + \propto}$$

Where X_t and $X_{t'}$ are population states associated with each time and $X_{t+\Delta}$ and $X_{t'+\Delta}$ are population states associated with $t + \Delta$ and $t' + \Delta$. $\|\cdot\|$ indicates the L-2 norm. The constant $\propto$ was set to 0.01 times the variance of X. Results were essentially identical across a range of values of $\propto$.

Because we are interested in whether the ratio ever becomes large, we take the maximum across all values of t'. We thus define divergence for time t as:

$$\mathrm{D}(t) = \max_{t',\Delta} \frac{\|X_{t+\Delta} - X_{t'+\Delta}\|^2}{\|X_t - X_{t'}\|^2 + \propto}$$

Δ could be as large as $\min(T - t,\ T' - t')$ where T is the duration of the condition associated with time t and T' is the duration of the condition associated with time t'.

Recurrent Neural Networks

We trained recurrent neural networks (RNNs) to perform a variation of the cycling task, specifically to produce 4 and 7 cycles of a sinusoid in response to external inputs. The RNN consists of N = 50 firing rate units with dynamics:

$$\tau \frac{\partial \mathbf{r}}{\partial t} = -\mathbf{r}(t) + \phi(\mathbf{J}\mathbf{r} + \mathbf{I}(t) + \mathbf{b})$$

$$z = \mathbf{w}_{\mathrm{out}}^{\mathrm{T}}\mathbf{r}$$

where τ is a time-constant, $\mathbf{r}$ represents an N-dimensional vector of firing rates, $\phi = \tanh$ is a nonlinear input-output function, $\mathbf{J}$ is an N x N matrix of recurrent weights, $\mathbf{I}(t)$ represents time-varying external input, and $\mathbf{b}$ is a vector of constant biases. The network output z is a linear readout of the rates multiplied by N output weights $\mathbf{w}_{\mathrm{out}}$. Both $\mathbf{J}$ and $\mathbf{w}_{\mathrm{out}}$. are initially drawn from a normal distribution of zero mean and variance 1/N, while $\mathbf{b}$ is initialized to zero. Throughout training, $\mathbf{J}$, $\mathbf{w}_{\mathrm{out}}$, and $\mathbf{b}$ are modified.

We considered two networks trained to perform the same cycling tasks but with different input configurations: context-tracking and context-naive networks. In the context-tracking case, the network is trained to generate 4 or 7 cycles of a sine wave after receiving a short go pulse (a square pulse that lasts for half a cycle duration prior to the start of cycling). Go pulses that elicit 4 or 7

cycles are distinguished by entering the network through different sets of random input weights; $\mathbf{I}(t) = \mathbf{w}_4\mathbf{I}(t)$ or $\mathbf{I}(t) = \mathbf{w}_7\mathbf{I}(t)$, where $\mathbf{I}(t)$ is a square pulse of unit amplitude. Training set consists of 50 trials (batches). Each trial is in random order and at random time with no overlap.

In the context-naive case, networks perform the same task as in the context-tracking case, but they receive both a go pulse and a stop pulse. Go and stop pulses are distinguished by entering the network through different sets of random input weights; $\mathbf{I}(t) = \mathbf{w}_{\text{go}}\mathbf{I}(t)$ or $\mathbf{I}(t) = \mathbf{w}_{\text{stop}}\mathbf{I}(t)$. Thus, go pulses do not carry any information about the desired number of cycles. Instead, the go and stop pulses are separated by an appropriate amount of time to compete the desired number of cycles. Training is done as in context-tracking case, except that the network is trained to cycle continuously in the absence of a stop-pulse.

We also considered a situation in which networks receive a go pulse that does not distinguish trial types (as in the context-naive case) but, rather than a stop pulse, they received a downward ramping input through another set of weights $\mathbf{w}_{ramp}$. The ramping input has a constant slope but different starting values for different numbers of desired cycles. The end of the cycling period in this case is indicated by the ramp signal reaching zero.

In all three cases, networks were trained using back-propagation-through-time (Werbos, 1988) using TensorFlow and an Adam optimizer to adjust, $\mathbf{J}$, $\mathbf{w}_{\text{out}}$, and $\mathbf{b}$ to minimize the squared difference between the network output z and the sinusoidal target function. All the input weights, $\mathbf{w}_4$, $\mathbf{w}_7$, $\mathbf{w}_{\text{go}}$, $\mathbf{w}_{\text{stop}}$ and $\mathbf{w}_{\text{ramp}}$, were drawn from a zero-mean unit-variance normal distribution and remain fixed throughout training. The amplitude of pulses and cycles are set to a value (unit amplitude) that produced a response but avoided saturating the units. The height of the ramp signal is set to the same amplitude as the input pulses for the 7-cycle condition. For each condition, we trained 500 networks each initialized with a different realization of $\mathbf{J}$, $\mathbf{w}_{\text{out}}$, and $\mathbf{b}$.

Trajectory-constrained Neural Networks

We sought to test the computational implications of trajectory divergence. To this end, we trained recurrent neural networks (RNNs) with an atypical approach. Rather than training networks to produce an output, we trained them to autonomously follow a particular internal trajectory. We then asked whether networks were able to follow those trajectories from beginning to end, without the benefit of any inputs indicating when to stop.

The RNN target trajectories were derived from neural recordings (M1, and SMA) during the 4-cycle movements for each of the 4 conditions (forward-bottom-start, forward-top-start, backward-bottom-start, backward-top-start). Target trajectories reflect the time period from movement onset until 250ms after movement offset. To emphasize that the RNN should remain in its final state post-movement, we extended the final sample of the target trajectory for an additional 500ms. Neural data were mean-centered and projected onto their top six principal components. Each target trajectory was normalized by its greatest norm along the time-series. For each target trajectory (two areas, two monkeys, and four conditions) we trained ten networks, each with a different weight initialization.

Network dynamics were governed by:

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) + \Delta t/\tau \left(-\boldsymbol{v}(t) + A\, f\big(\boldsymbol{v}(t)\big) + \boldsymbol{w}(t)\right)$$

With the learning rule for synaptic input trajectories:

$$Af\big(v(t)\big) \approx s_{\text{targ}}(t) = Gy_{\text{targ}}(t)$$

where $f \coloneqq \tanh$, and $\boldsymbol{w}{\sim}N(\boldsymbol{0}, \sigma_w^2 I)$ adds noise. $\boldsymbol{v}$ can be thought of as the membrane voltage and $f\big(\boldsymbol{v}(t)\big)$ as the firing rate. $Af\big(\boldsymbol{v}(t)\big)$ is then the network input to each unit: the firing rates weighted by the connection strengths. A was initialized such that $A_{ij} \sim N(0, \frac{1}{\sqrt{n}})$ and trained using recursive

least squares. $\boldsymbol{y}_{\text{targ}}$ is the idealized low-dimensional trajectory. G is a matrix of random weights, sampled from $U[-.5, .5]$, that maps the target trajectory onto a target input of each model unit. The entries of A were initialized by draws from a centered normal distribution with variance 1/n (where n = 50, the number of network units). Simulation employed 4 ms time steps.

To begin a given training epoch, the initial state was set with $v(0)$ based on $s_{\text{targ}}(0)$ and A. The RNN was simulated, applying recursive least squares (Sussillo & Abbott, 2009) with parameter $\alpha = 1$ to modify A as time unfolds. After 1000 training epochs, stability was assessed by simulating the network 100 times, and computing the mean squared difference between the actual and target trajectory. That error was normalized by the variance of the target trajectory, yielding an R^2 value. An average (across the 100 simulated trials) $R^2 < 0.9$ was considered a failure.

Because population trajectories never perfectly repeated, it was trivially true that networks could follow the full trajectory, for both M1 and SMA, in the complete absence of noise (i.e., for $\sigma_w = 0$). Because it is unclear what level of noise is physiologically relevant, we repeated the analysis at multiple values of σ_w. Results are reported for a value of σ_w where all networks failed to follow the M1 trajectories. At this level, most networks successfully followed the SMA trajectories (though not all, as some weight initializations never resulted in good solutions). We also performed an analysis where we swept the value of σ_w until failure. The level of noise that was tolerated was much greater when networks followed the SMA trajectories. Indeed, some M1 trajectories (i.e., for particular conditions) could never be consistently followed even at the lowest noise level tested. The visualization of example network activity (Figure 3.8 b-d) was produced by 'decoding' network activity, by inverting G, to reconstruct the first three dimensions of the target trajectory.

Supplementary Material

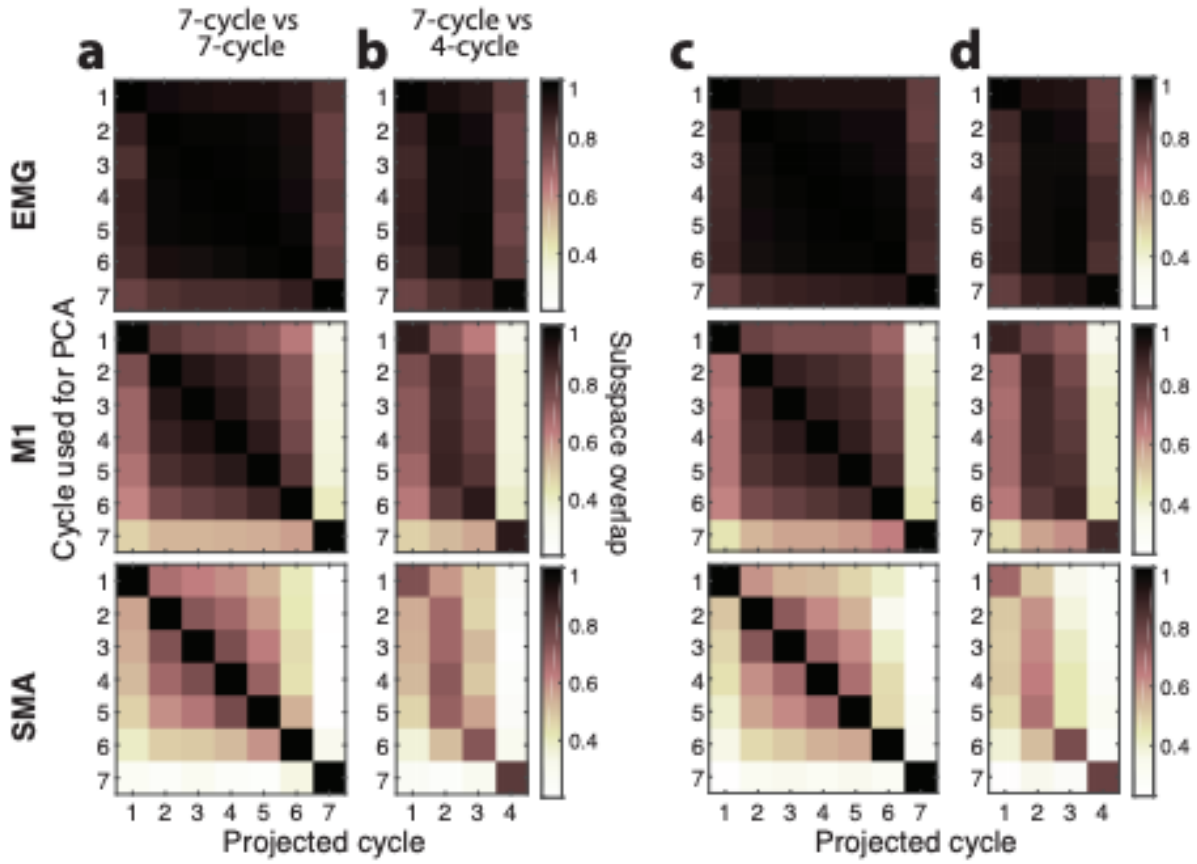

Figure 3.S1: Cycle-to-cycle analysis of subspace overlap in 12 dimensions

Same as Figure 5 but in 12 dimensions.

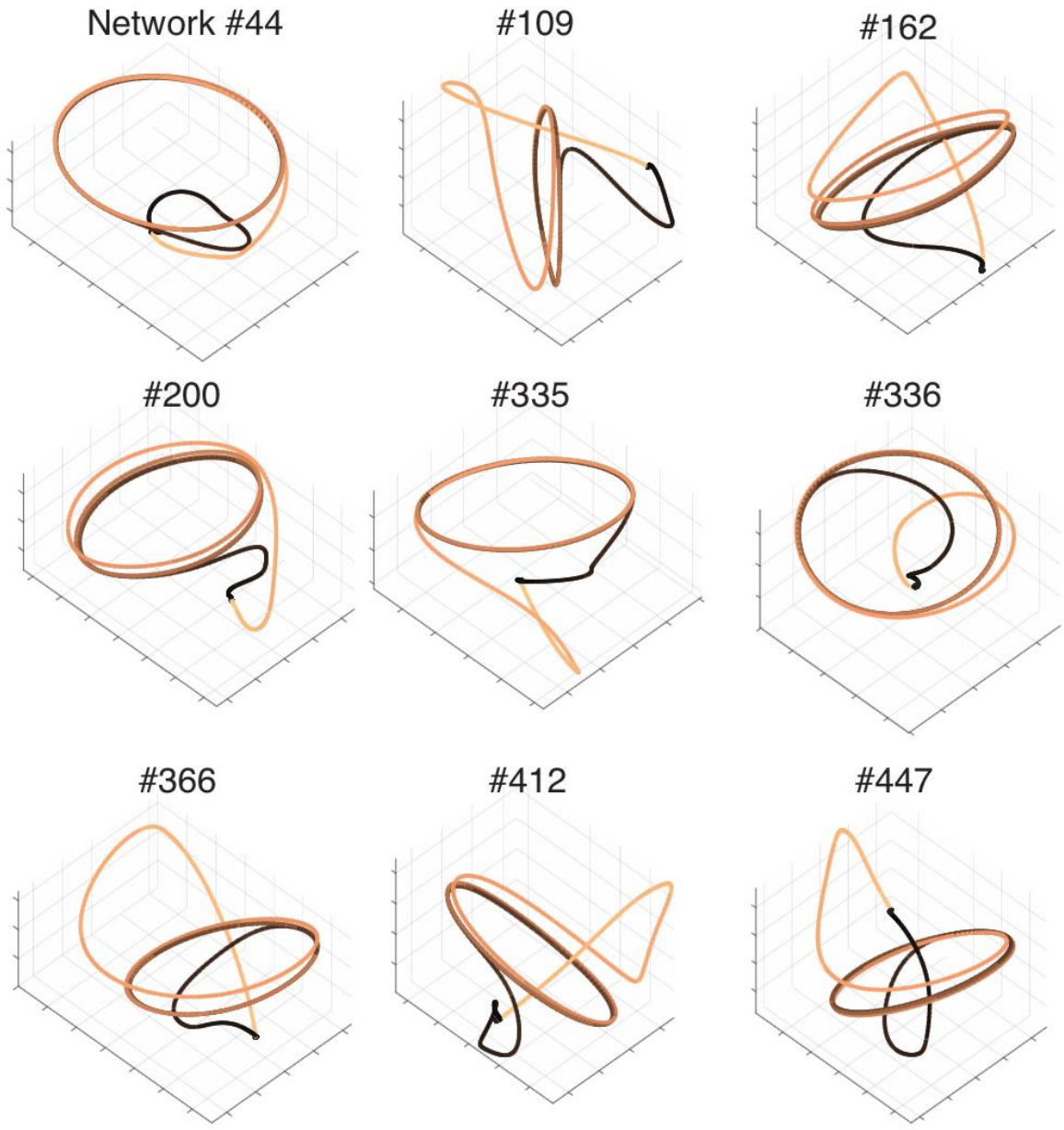

Figure 3.S2 Additional examples of context-naïve networks

Format as for Figure 6a. Nine examples of context-naïve networks trained with different initializations.

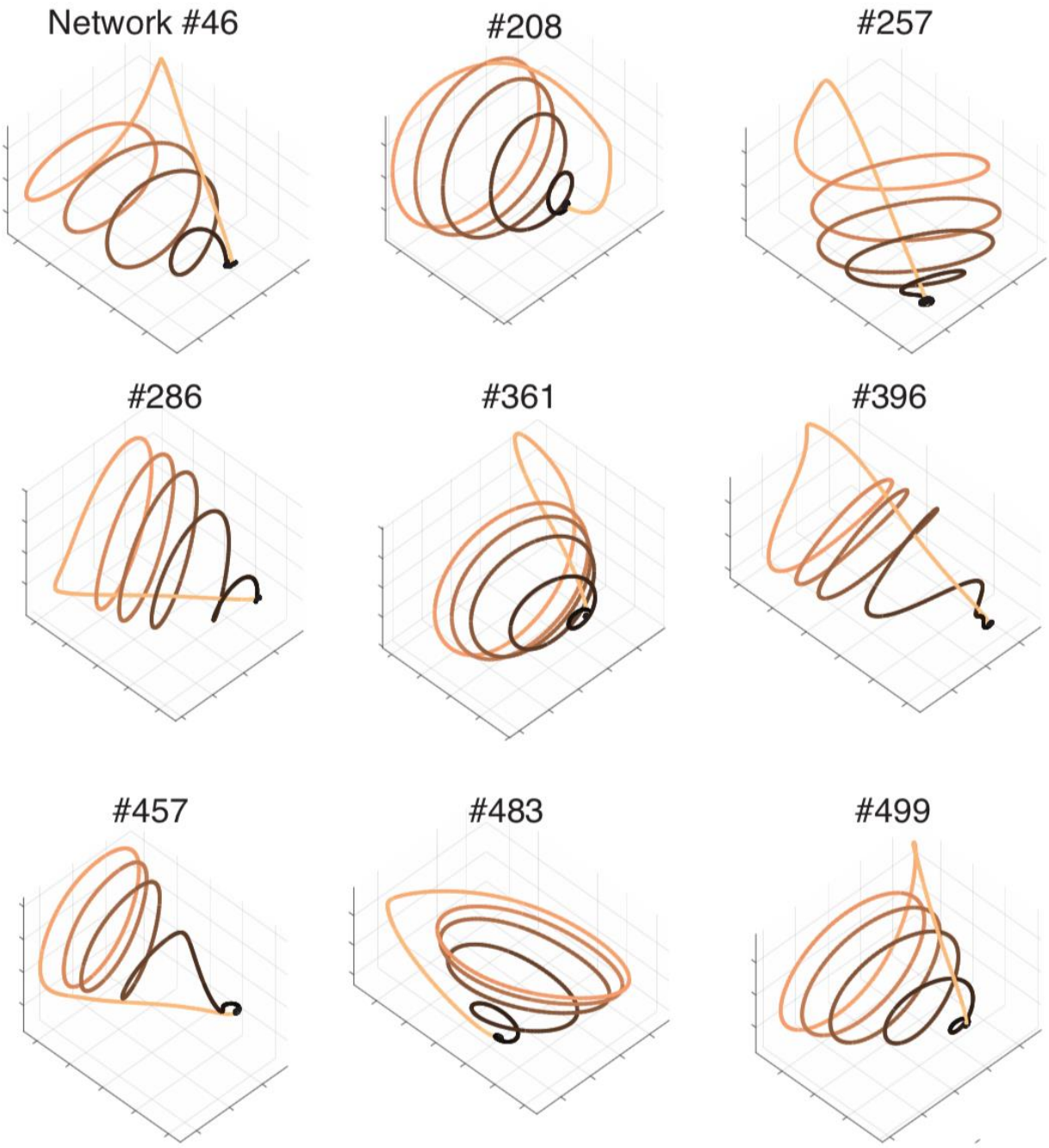

Figure 3.S3 Additional examples of context-tracking

Format as for Figure 6b. Nine examples of context-tracking networks trained with different initializations

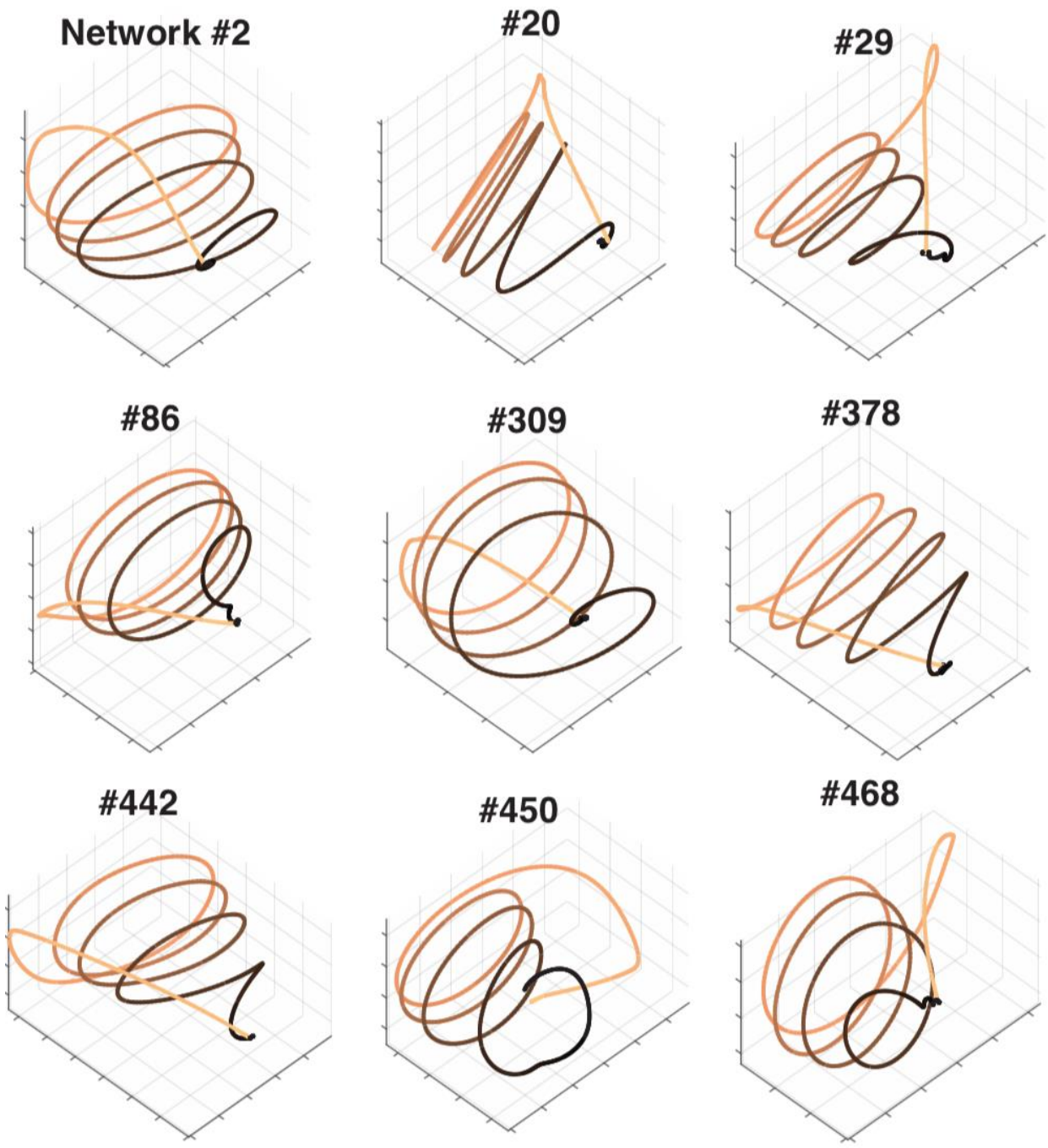

Figure 3.S4 Examples of context-tracking networks trained with a ramping input

Format as for Figure 6b. Nine examples of context-tracking networks trained with different initializations in the presence of a ramping input.

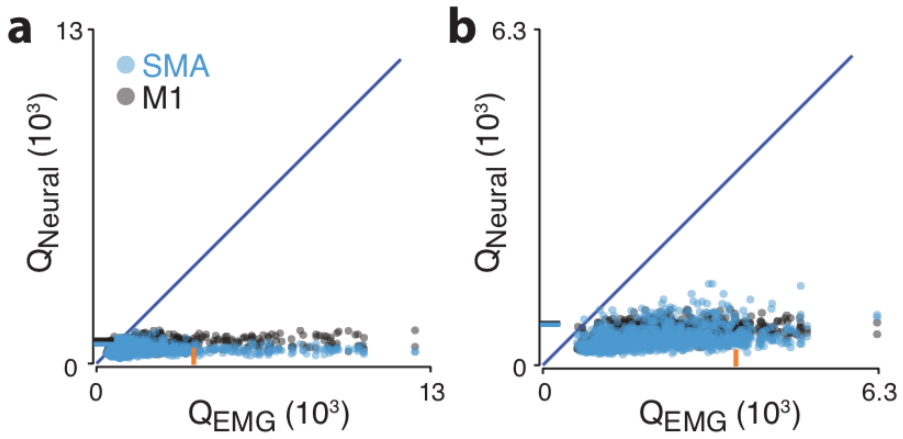

Figure 3.S5 Relationship between trajectory tangling and trajectory divergence

Format as for Figure 7a,b. We employed two metrics that assess different aspects of trajectory structure, yet are conceptually and mathematically related. Trajectory tangling, defined as $Q(t) = \max_{t\prime} \frac{\|\dot{x}(t)-\dot{x}(t\prime)\|^2}{\|x(t)-x(t\prime)\|^2+\varepsilon}$, assesses whether the trajectory could have been produced by a smooth dynamical flow-field. Trajectory divergence, defined as $D(t) = \max_{t',\Delta t} \frac{\|x(t+\Delta)-x(t'+\Delta)\|^2}{\|x(t)-x(t\prime)\|^2+\propto}$, assesses whether two trajectories (or two portions of the same trajectory) are close but eventually diverge. Intuitively, divergence is related to tangling but considers longer timescales and future (but not past) events. For example, if two trajectories track together and then slowly separate, tangling may remain low, yet divergence will be high. Two such examples are shown in the lower-right quadrant. Conversely, if two trajectories rapidly converge, tangling will briefly become high yet divergence will remain low. Two such examples are shown in the upper-left quadrant.

The relationship between tangling and divergence can be appreciated by inspection: the denominators are the same (ignoring constants) but the numerators differ. For tangling, the numerator assesses whether two trajectories (one considered at t and one at t') are headed, at that instant, in different directions. For divergence, the numerator asks whether trajectories eventually separate by time Δ in the future. Thus, the numerator for tangling and divergence are related by integration. This can be appreciated by considering two quantities. First, $\boldsymbol{s}(\tau) = \boldsymbol{x}(t+\tau) - \boldsymbol{x}(t'+\tau)$, the separation between two trajectories at the indicated times. Second, $\boldsymbol{v}(\tau) = \dot{\boldsymbol{x}}(t+\tau) - \dot{\boldsymbol{x}}(t'+\tau)$, the difference in trajectory velocities. Trajectory divergence is based on $\frac{\|\boldsymbol{s}(\Delta)\|}{\|\boldsymbol{s}(0)\|}$. Trajectory tangling is based on $\frac{\|\boldsymbol{v}(0)\|}{\|\boldsymbol{s}(0)\|}$. This latter quantity can be modified to consider differences that accumulate over time: $\frac{\left\|\int_0^\Delta \boldsymbol{v}(\tau)\,\boldsymbol{d\tau}\right\|}{\|\boldsymbol{s}(0)\|} = \frac{\|\boldsymbol{s}(\Delta)-\boldsymbol{s}(0)\|}{\|\boldsymbol{s}(0)\|} \approx \frac{\|\boldsymbol{s}(\Delta)\|}{\|\boldsymbol{s}(0)\|}$ whenever $\|\boldsymbol{s}(\Delta)\| \gg \|\boldsymbol{s}(0)\|$. Thus, $\frac{\left\|\int_0^\Delta \boldsymbol{v}(\tau)\,\boldsymbol{d\tau}\right\|}{\|\boldsymbol{s}(0)\|}$ and the divergence metric are nearly identical whenever either is high (differences among small values are irrelevant to our analyses). This exercise illustrates that divergence can be thought of as a version of tangling that considers the future, rather than just the present. While we could have based analysis on $\frac{\left\|\int_0^\Delta \boldsymbol{v}(\tau)\,\boldsymbol{d\tau}\right\|}{\|\boldsymbol{s}(0)\|}$, with nearly identical results, we prefer the more straightforward definition used in the manuscript.

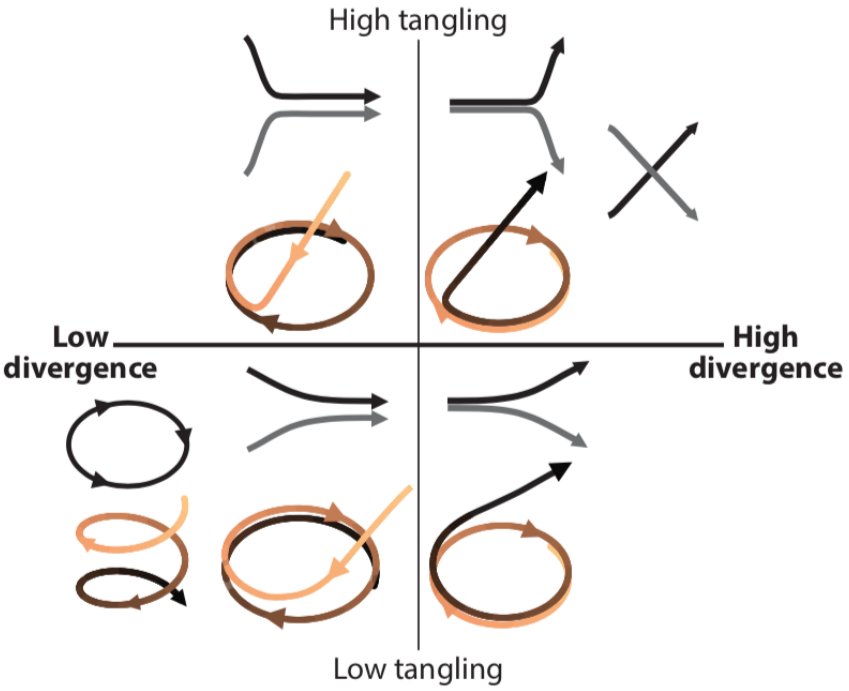

Figure 3.S6 Illustration of trajectories that would yield low or high trajectory divergence and trajectory tangling.

Pairs of lines (black and gray) indicate trajectories that might correspond to two different conditions while circular tan-black lines indicate trajectories that might correspond to a single condition over time. Trajectories that have high tangling (upper two quadrants) may have sharp turns and crossing points. Trajectories that have high divergence (right two quadrants) are similar at some point in time but later separate. Divergence will remain low (left two quadrants) if trajectories start dissimilar and converge (e.g. trajectories in the right column), start similar and stay similar (e.g. black circular trajectory in the bottom left quadrant), or maintain dissimilarity over time (e.g. helical trajectory at the bottom left corner).

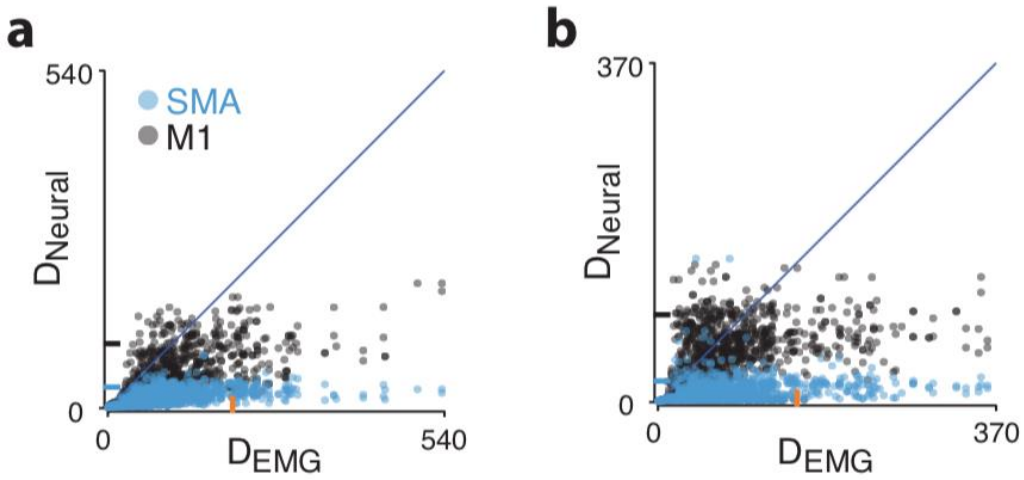

Figure 3.S7 Trajectory divergence is high in muscle activity.

a) Format as for Figure 7a,b. Black dots indicate trajectory divergence for each time point in M1 vs trajectory divergence for corresponding time points in muscle activity. Blue dots are the same but for SMA. Blue (black) tick mark along the vertical axis denotes the 90th percentile trajectory divergence for SMA (M1). Orange tick mark along the horizontal axis denotes 90th percentile trajectory divergence for EMG. Trajectory divergence is lower in SMA than in M1 (blue dots are lower than black dots). Trajectory divergence is much higher in muscle activity than in SMA (blue dots lie along a flat distribution with very few points above the unity line). Data is for monkey C.
b) Same for monkey D.

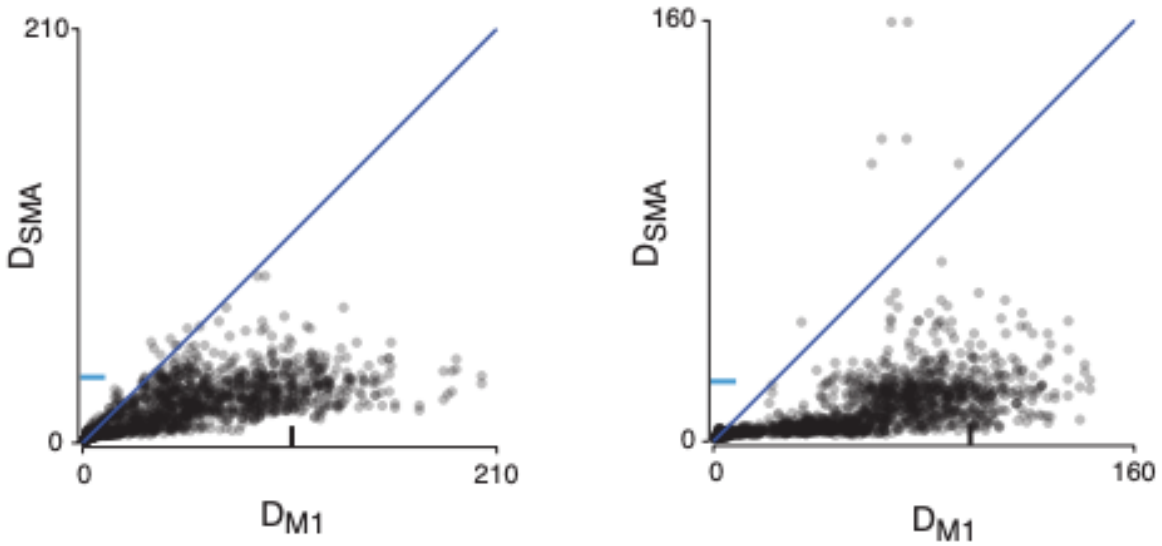

Figure 3.S8 Trajectory divergence in M1 and SMA computed by indexing across all conditions

Same as for Figure 7a,b except divergence for each time point is computed by indexing across all time points from all conditions. In Figure 7, divergence was computed by indexing across times that belonged to either the same condition or a condition with the same starting position and pedaling direction but di-erent distances. Here, divergence was computed by indexing across all times from all conditions regardless of distance, starting position, or pedaling direction.

Chapter 4 **Conclusions**

Characterizing motifs of population structure and geometric properties has proved a fruitful avenue for the study of high-dimensional neural datasets. As novel modeling tools develop, the interplay between data-driven and model-driven approaches promises to yield deeper insights. Here, I offer some concluding remarks.

Remaining caveats

I've discussed the merits of measuring geometric properties of population activity. This strategy is aimed at abstracting away from task-specific structure and toward more fundamental constraints that drive the observed motifs. Is this strategy always revealing or might motifs themselves be the fundamental feature? Regions closer to the periphery such as primary motor cortex may be more likely to have a canonical computation that is task-general and revealed by geometric features. Yet it might be the case that more cognitive regions have very general computational abilities that are shaped by the individual tasks. For these regions, it might be the case that understanding motifs provides a complete understanding of the computation and abstracting away from task-specificity

is unwarranted. Going forward, it will be important to keep these possibilities in mind. When we describe motifs of population structure, do we propose that they speak to a fundamental computation that the region is performing or the specific manifestation of geometry for this task?

It is also essential to consider whether the feature of interest might be a consequence of simpler phenomena (Elsayed & Cunningham, 2017). Results can be validated by comparing across neural regions that are and are not expected to share the feature of interest or across related data that share similar temporal features such as muscle activity (Russo et al., 2018; Seely et al., 2016). Further precision can be gained by generating surrogate data that matches temporal, neural, or condition correlations (Elsayed & Cunningham, 2017). These controls will enable us to determine whether the geometric property of interest is fully or partially a byproduct of such features. Notably, even if this is found to be the case, the property of interest may still be a real and useful property that is important to accomplish the task at hand. Still, the property may be guaranteed given some parameters of the task or statistics of the data.

Finally, if we aim to understand general task-invariant features of a neural region, we must observe the system broadly enough. We need to record enough neurons to ensure the state of the system is accurately measured. Next, we need to record a diverse range of states across time, conditions, and tasks, keeping in mind that experimenter-designed divisions between such parameters may not be reflected by the brain. That is, two conditions of the same task (Russo et al., 2018) or even two temporal periods within the same condition (Ames et al., 2014; Elsayed et al., 2016; Kao, 2018; Kaufman et al., 2014) may be produced by very different neural dynamics. More subtly, observed population structure motifs may be highly suggestive of a particular form of the underlying

dynamics but generally, it is possible for the same structure to be produced by distinctly different dynamic mechanisms (Kao, 2018). This concern will be reduced, although not wholly alleviated, by observing the system across a wide range of states.

Future directions

An increase in the number of neurons we can record simultaneously brings with it an improved ability to asses neural dynamics on single trials (Pandarinath, O'Shea, et al., 2018; Yu et al., 2009). Such tools open the door to a vast array of questions. We will be able to begin to understand computations that occur internally on time-scales that may not be tied to external stimuli (*e.g.* decision making). We will also be able to observe population structure over the course of learning. It would be fascinating to determine whether geometric properties change as a task is learned and to observe whether and how error trials contribute to this change on a single-trial time-scale. Perhaps we will even develop the technology to precisely perturb the network in a manner that is relevant to the population structure and be able to directly test the function of different structural motifs.

This technology will also enable us to study the contribution of different cell populations to neural dynamics. For example, it has long been proposed that cells in cortical layers segregate populations of neurons are primarily dedicated to receiving inputs, producing outputs, or performing an internal computation (Felleman & Van Essen, 1991). Meanwhile, mounting evidence suggests that inter-areal communication occurs in a dedicated subspace (Perich, Gallego, & Miller, 2018; Semedo, Zandvakili, Machens, Yu, & Kohn, 2019) that is null with respect to the receiving region's output

(Kaufman et al., 2014). It would be fascinating to bridge these two levels of thinking by determining whether subpopulations of neurons differentially contribute to these subspaces.

The study of neural dynamics and population geometry is only beginning to reveal its potential. Our growing ability to record many neurons simultaneously and new tools for modeling artificial networks promise to provide a rich source of new insights.

References

Aflalo, T. N., & Graziano, M. S. A. (2007). Relationship between Unconstrained Arm Movements and Single-Neuron Firing in the Macaque Motor Cortex. *Journal of Neuroscience, 27*(11), 2760-2780. Retrieved from http://www.jneurosci.org/content/27/11/2760.full.pdf. doi:10.1523/jneurosci.3147-06.2007

Ajemian, R., Green, A., Bullock, D., Sergio, L., Kalaska, J., & Grossberg, S. (2008). Assessing the function of motor cortex: single-neuron models of how neural response is modulated by limb biomechanics. *Neuron, 58*(3), 414-428. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18466751
http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6WSS-4SFRCHN-G-2&_cdi=7054&_user=145269&_pii=S0896627308002213&_origin=gateway&_coverDate=05%2F08%2F2008&_sk=999419996&view=c&wchp=dGLzVtb-zSkzk&md5=ddbc697fbb1d0d1eb5e1c8f8ed6807f8&ie=/sdarticle.pdf. doi:10.1016/j.neuron.2008.02.033

Ames, K. C., Ryu, S. I., & Shenoy, K. V. (2014). Neural Dynamics of Reaching following Incorrect or Absent Motor Preparation. *Neuron, 81*(2), 438-451. Retrieved from <Go to ISI>://WOS:000330420700020. doi:10.1016/j.neuron.2013.11.003

Ashe, J., & Georgopoulos, A. P. (1994). Movement parameters and neural activity in motor cortex and area 5. *Cerebral cortex, 4*(6), 590-600. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7703686

Boccardi, E., Della Sala, S., Motto, C., & Spinnler, H. (2002). Utilisation behaviour consequent to bilateral SMA softening. *Cortex, 38*(3), 289-308. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12146657.

Burnod, Y., Grandguillaume, P., Otto, I., Ferraina, S., Johnson, P. B., & Caminiti, R. (1992). Visuomotor transformations underlying arm movements toward visual targets: a neural network model of cerebral cortical operations. *J Neurosci, 12*(4), 1435-1453. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1556602

Cadena-Valencia, J., Garcia-Garibay, O., Merchant, H., Jazayeri, M., & de Lafuente, V. (2018). Entrainment and maintenance of an internal metronome in supplementary motor area. *Elife, 7*. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30346275. doi:10.7554/eLife.38983

Cheney, P. D., & Fetz, E. E. (1980). Functional classes of primate corticomotoneuronal cells and their relation to active force. *J Neurophysiol, 44*(4), 773-791. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=6253605

Churchland, M. M., & Cunningham, J. P. (2014). A Dynamical Basis Set for Generating Reaches. *Cold Spring Harb Symp Quant Biol, 79*, 67-80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25851506. doi:10.1101/sqb.2014.79.024703

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature, 487*(7405), 51-+. Retrieved from <Go to ISI>://WOS:000305982900048. doi:10.1038/nature11129

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I., & Shenoy, K. V. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron, 68*(3), 387-400. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21040842. doi:10.1016/j.neuron.2010.09.015

Churchland, M. M., & Shenoy, K. V. (2007). Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J Neurophysiol, 97*(6), 4235-4257. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17376854

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience, 14*(7), 811-819. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21709677. doi:10.1038/nn.2842

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: survey, insights, and generalizations. *Journal of Machine Learning Research*(16), 2859-2900.

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience, 17*(11), 1500-1509. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25151264. doi:10.1038/nn.3776

Driscoll, L. N., Golub, M. D., & Sussillo, D. (2018). Computation through Cortical Dynamics. *Neuron, 98*(5), 873-875. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29879388. doi:10.1016/j.neuron.2018.05.029

Druckmann, S., & Chklovskii, D. B. (2012). Neuronal circuits underlying persistent representations despite time varying activity. *Curr Biol, 22*(22), 2095-2103. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23084992. doi:10.1016/j.cub.2012.08.058

Eccles, J. C. (1982). The Initiation of Voluntary Movements by the Supplementary Motor Area. *Archiv Fur Psychiatrie Und Nervenkrankheiten, 231*(5), 423-441. Retrieved from <Go to ISI>://WOS:A1982NZ90000003. doi:Doi 10.1007/Bf00342722

Elsayed, G. F., & Cunningham, J. P. (2017). Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience, 20*(9), 1310-+. Retrieved from <Go to ISI>://WOS:000408587700020. doi:10.1038/nn.4617

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications, 7*, 13239. Retrieved from http://dx.doi.org/10.1038/ncomms13239. doi:10.1038/ncomms13239

Evarts, E. V. (1968). Relation of pyramidal tract activity to force exerted during voluntary movement. *J Neurophysiol, 31*(1), 14-27. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=4966614

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex, 1*(1), 1-47. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1822724.

Fetz, E. E. (1992). Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and Brain Sciences, 15*(4), 679-690.

Fitzsimmons, N. A., Lebedev, M. A., Peikon, I. D., & Nicolelis, M. A. (2009). Extracting kinematic parameters for monkey bipedal walking from cortical neuronal ensemble activity. *Frontiers in Integrative Neuroscience, 3*, 3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19404411. doi:10.3389/neuro.07.003.2009

Foster, J. D., Nuyujukian, P., Freifeld, O., Gao, H., Walker, R., S, I. R., . . . Shenoy, K. V. (2014). A freely-moving monkey treadmill model. *J Neural Eng, 11*(4), 046020. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24995476. doi:10.1088/1741-2560/11/4/046020

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron, 94*(5), 978-984. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28595054. doi:10.1016/j.neuron.2017.05.025

Gallego, J. A., Perich, M. G., Naufel, S. N., Ethier, C., Solla, S. A., & Miller, L. E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat Commun, 9*(1), 4233. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30315158. doi:10.1038/s41467-018-06560-z

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci, 2*(11), 1527-1537. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7143039

Georgopoulos, A. P., Naselaris, T., Merchant, H., & Amirikian, B. (2007). Reply to kurtzer and herter. *J Neurophysiol, 97*(6), 4391-4392. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17553956

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416-1419. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3749885

Griffin, D. M., Hudson, H. M., Belhaj-Saif, A., McKiernan, B. J., & Cheney, P. D. (2008). Do Corticomotoneuronal Cells Predict Target Muscle EMG Activity? *Journal of Neurophysiology, 99*(3), 1169-1986. Retrieved from http://jn.physiology.org/content/99/3/1169.full.pdf. doi:10.1152/jn.00906.2007

Hall, T. M., de Carvalho, F., & Jackson, A. (2014). A common structure underlies low-frequency cortical dynamics in movement, sleep, and sedation. *Neuron, 83*(5), 1185-1199. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25132467
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4157580/pdf/main.pdf.
doi:10.1016/j.neuron.2014.07.022

Hart, C. B., & Giszter, S. F. (2010). A neural basis for motor primitives in the spinal cord. *J Neurosci*, *30*(4), 1322-1336. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20107059. doi:10.1523/JNEUROSCI.5894-08.2010

Hatsopoulos, N. G., Xu, Q., & Amit, Y. (2007). Encoding of movement fragments in the motor cortex. *J Neurosci*, *27*(19), 5105-5114. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17494696

Heming, E. A., Lillicrap, T. P., Omrani, M., Herter, T. M., Pruszynski, J. A., & Scott, S. H. (2016). Primary motor cortex neurons classified in a postural task predict muscle activation patterns in a reaching task. *J Neurophysiol*, *115*(4), 2021-2032. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26843605. doi:10.1152/jn.00971.2015

Hennequin, G., Vogels, T. P., & Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, *82*(6), 1394-1406. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24945778. doi:10.1016/j.neuron.2014.04.045

Herter, T. M., Korbel, T., & Scott, S. H. (2009). Comparison of neural responses in primary motor cortex to transient and continuous loads during posture. *Journal of Neurophysiology*, *101*(1), 150-163. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19005005 http://jn.physiology.org/content/101/1/150.full.pdf. doi:10.1152/jn.90230.2008

Kakei, S., Hoffman, D. S., & Strick, P. L. (1999). Muscle and movement representations in the primary motor cortex. *Science*, *285*(5436), 2136-2139. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10497133

Kao, J. C. (2018). Considerations in using recurrent neural networks to probe neural dynamics. *bioRxiv*, *364489*.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, *17*(3), 440-448. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24487233. doi:10.1038/nn.3643

Kaufman, M. T., Seely, J. S., Sussillo, D., Ryu, S. I., Shenoy, K. V., & Churchland, M. M. (2016). The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not

Movement Type. *eNeuro*, *3*(4). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27761519. doi:10.1523/ENEURO.0085-16.2016

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., . . . Machens, C. K. (2016). Demixed principal component analysis of neural population data. *Elife*, *5*. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27067378. doi:10.7554/eLife.10989

Krainik, A., Lehericy, S., Duffau, H., Vlaicu, M., Poupon, F., Capelle, L., . . . Marsault, C. (2001). Role of the supplementary motor area in motor deficit following medial frontal lobe surgery. *Neurology*, *57*(5), 871-878. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11552019.

Laplane, D., Talairach, J., Meininger, V., Bancaud, J., & Orgogozo, J. M. (1977). Clinical Consequences of Corticectomies Involving Supplementary Motor Area in Man. *Journal of the Neurological Sciences*, *34*(3), 301-314. Retrieved from <Go to ISI>://WOS:A1977EC52900001. doi:Doi 10.1016/0022-510x(77)90148-4

Lara, A. H., Cunningham, J. P., & Churchland, M. M. (2018). Different population dynamics in the supplementary motor area and motor cortex during reaching. *Nat Commun*, *9*(1), 2754. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30013188. doi:10.1038/s41467-018-05146-z

Lara, A. H., Elsayed, G. F., Cunningham, J. P., & Churchland, M. M. (2017). Conservation of preparatory neural events regardless of how movement is initiated. *bioRrxiv*. doi:10.1101/189035

Li, N., Daie, K., Svoboda, K., & Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, *532*(7600), 459-464. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27074502. doi:10.1038/nature17643

Lillicrap, T. P., & Scott, S. H. (2013). Preference distributions of primary motor cortex neurons reflect control solutions optimized for limb biomechanics. *Neuron*, *77*(1), 168-179. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23312524. doi:10.1016/j.neuron.2012.10.041

Machens, C. K., Romo, R., & Brody, C. D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J Neurosci*, *30*(1), 350-360. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20053916. doi:10.1523/JNEUROSCI.3276-09.2010

Maier, M. A., Shupe, L. E., & Fetz, E. E. (2005). Dynamic neural network models of the premotoneuronal circuitry controlling wrist movements in primates. *J Comput Neurosci, 19*(2), 125-146. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16133816

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature, 503*(7474), 78-+. Retrieved from <Go to ISI>://WOS:000326585600035. doi:10.1038/nature12742

Michaels, J. A., Dann, B., & Scherberger, H. (2016). Neural Population Dynamics during Reaching Are Better Explained by a Dynamical System than Representational Tuning. *PLoS Comput Biol, 12*(11), e1005175. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27814352 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096671/pdf/pcbi.1005175.pdf. doi:10.1371/journal.pcbi.1005175

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia output and cognition: evidence from anatomical, behavioral, and clinical studies. *Brain Cogn, 42*(2), 183-200. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10744919. doi:10.1006/brcg.1999.1099

Miri, A., Warriner, C. L., Seely, J. S., Elsayed, G. F., Cunningham, J. P., Churchland, M. M., & Jessell, T. M. (2017). Behaviorally Selective Engagement of Short-Latency Effector Pathways by Motor Cortex. *Neuron*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28735748. doi:10.1016/j.neuron.2017.06.042

Moran, D. W., & Schwartz, A. B. (1999a). Motor cortical activity during drawing movements: population representation during spiral tracing. *J Neurophysiol, 82*(5), 2693-2704. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10561438

Moran, D. W., & Schwartz, A. B. (1999b). Motor cortical representation of speed and direction during reaching. *J Neurophysiol, 82*(5), 2676-2692. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10561437

Moran, D. W., & Schwartz, A. B. (2000). One motor cortex, two different views. *Nature Neuroscience, 3*(10), 963; author reply 963-965. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11017157. doi:10.1038/79880

Morrow, M. M., Pohlmeyer, E. A., & Miller, L. E. (2009). Control of Muscle Synergies by Cortical Ensembles. *629*, 179-199. doi:10.1007/978-0-387-77064-2_9

Mushiake, H., Inase, M., & Tanji, J. (1991). Neuronal activity in the primate premotor, supplementary, and precentral motor cortex during visually guided and internally determined sequential movements. *J Neurophysiol, 66*(3), 705-718. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1753282. doi:10.1152/jn.1991.66.3.705

Mussa-Ivaldi, F. A. (1988). Do neurons in the motor cortex encode movement direction? An alternative hypothesis. *Neurosci Lett, 91*(1), 106-111. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3173781

Nakamura, K., Sakai, K., & Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. *Journal of Neurophysiology, 80*(5), 2671-2687. Retrieved from <Go to ISI>://WOS:000077069000035.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8637596. doi:10.1038/381607a0

Pandarinath, C., Ames, K. C., Russo, A. A., Farshchian, A., Miller, L. E., Dyer, E. L., & Kao, J. C. (2018). Latent Factors and Dynamics in Motor Cortex and Their Application to Brain-Machine Interfaces. *J Neurosci, 38*(44), 9390-9401. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30381431. doi:10.1523/JNEUROSCI.1669-18.2018

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., . . . Sussillo, D. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat Methods, 15*(10), 805-815. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30224673. doi:10.1038/s41592-018-0109-9

Penfield, W., & Welch, K. (1951). The Supplementary Motor Area of the Cerebral Cortex - a Clinical and Experimental Study. *Ama Archives of Neurology and Psychiatry, 66*(3), 289-317. Retrieved from <Go to ISI>://WOS:A1951UF31100004. doi:DOI 10.1001/archneurpsyc.1951.02320090038004

Perich, M. G., Gallego, J. A., & Miller, L. E. (2018). A Neural Population Mechanism for Rapid Learning. *Neuron, 100*(4), 964-976 e967. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30344047. doi:10.1016/j.neuron.2018.09.030

Prinz, A. A. (2010). Computational approaches to neuronal network analysis. *Philos Trans R Soc Lond B Biol Sci, 365*(1551), 2397-2405. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20603360. doi:10.1098/rstb.2010.0029

Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience, 17*(12), 1784-1792. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25383902. doi:10.1038/nn.3865

Rathelot, J. A., & Strick, P. L. (2006). Muscle representation in the macaque motor cortex: an anatomical perspective. *Proceedings of the National Academy of Sciences of the United States of America, 103*(21), 8257-8262. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16702556. doi:10.1073/pnas.0602933103

Rathelot, J. A., & Strick, P. L. (2009). Subdivisions of primary motor cortex based on cortico-motoneuronal cells. *Proceedings of the National Academy of Sciences of the United States of America, 106*(3), 918-923. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19139417. doi:10.1073/pnas.0808362106

Reimer, J., & Hatsopoulos, N. G. (2009). The problem of parametric neural coding in the motor system. *Advances in experimental medicine and biology, 629*, 243-259. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19227503. doi:10.1007/978-0-387-77064-2_12

Remington, E. D., Egger, S. W., Narain, D., Wang, J., & Jazayeri, M. (2018). A Dynamical Systems Perspective on Flexible Motor Timing. *Trends in Cognitive Sciences, 22*(10), 938-952. Retrieved from <Go to ISI>://WOS:000445534000011. doi:10.1016/j.tics.2018.07.010

Remington, E. D., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron, 98*(5), 1005-+. Retrieved from <Go to ISI>://WOS:000436585500017. doi:10.1016/j.neuron.2018.05.020

Rokni, U., & Sompolinsky, H. (2012). How the brain generates movement. *Neural Computation, 24*(2), 289-331. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22023199. doi:10.1162/NECO_a_00223

Roland, P. E., Larsen, B., Lassen, N. A., & Skinhoj, E. (1980). Supplementary Motor Area and Other Cortical Areas in Organization of Voluntary Movements in Man. *Journal of Neurophysiology, 43*(1), 118-136. Retrieved from <Go to ISI>://WOS:A1980JC97100009.

Russo, A. A., Bittner, S. R., Perkins, S. M., Seely, J. S., London, B. M., Lara, A. H., . . . Churchland, M. M. (2018). Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron, 97*(4), 953-+. Retrieved from <Go to ISI>://WOS:000425713200020. doi:10.1016/j.neuron.2018.01.004

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., . . . Batista, A. P. (2014). Neural constraints on learning. *Nature, 512*(7515), 423-426. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25164754. doi:10.1038/nature13665

Sanger, T. D. (1994). Theoretical considerations for the analysis of population coding in motor cortex. *Neural Computation, 6*(1), 29-37.

Schieber, M. H., & Rivlis, G. (2007). Partial Reconstruction of Muscle Activity From a Pruned Network of Diverse Motor Cortex Neurons. *Journal of Neurophysiology, 97*(1), 70-82. doi:10.1152/jn.00544.2006

Schwartz, A. B. (1994). Direct cortical representation of drawing. *Science, 265*(5171), 540-542. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8036499

Schwartz, A. B. (2007). Useful signals from motor cortex. *J Physiol, 579*(Pt 3), 581-601. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17255162. doi:10.1113/jphysiol.2006.126698

Schwartz, A. B., Moran, D. W., & Reina, G. A. (2004). Differential representation of perception and action in the frontal cortex. *Science, 303*(5656), 380-383. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14726593

Scott, S. H. (1997). Comparison of onset time and magnitude of activity for proximal arm muscles and motor cortical cells before reaching movements. *Journal of Neurophysiology, 77*(2), 1016-1022. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9065865.

Scott, S. H. (2008). Inconvenient truths about neural processing in primary motor cortex. *J Physiol, 586*(5), 1217-1224. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18187462. doi:10.1113/jphysiol.2007.146068

Seely, J. S., Kaufman, M. T., Ryu, S. I., Shenoy, K. V., Cunningham, J. P., & Churchland, M. M. (2016). Tensor Analysis Reveals Distinct Population Structure that Parallels the Different Computational Roles of Areas M1 and V1. *PLoS Comput Biol*, *12*(11), e1005164. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27814353. doi:10.1371/journal.pcbi.1005164

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. *Neuron*. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30770252. doi:10.1016/j.neuron.2019.01.026

Sergio, L. E., Hamel-Paquet, C., & Kalaska, J. F. (2005). Motor cortex neural correlates of output kinematics and kinetics during isometric-force and arm-reaching tasks. *J Neurophysiol*, *94*(4), 2353-2378. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15888522

Shalit, U., Zinger, N., Joshua, M., & Prut, Y. (2012). Descending systems translate transient cortical commands into a sustained muscle activation signal. *Cerebral cortex*, *22*(8), 1904-1914. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21965441
http://cercor.oxfordjournals.org/content/22/8/1904.full.pdf. doi:10.1093/cercor/bhr267

Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013). Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience, Vol 36*, *36*, 337-359. Retrieved from <Go to ISI>://WOS:000323892300015. doi:10.1146/annurev-neuro-062111-150509

Shima, K., & Tanji, J. (1998). Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. *J Neurophysiol*, *80*(6), 3247-3260. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9862919. doi:10.1152/jn.1998.80.6.3247

Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *J Neurophysiol*, *84*(4), 2148-2160. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11024102. doi:10.1152/jn.2000.84.4.2148

Sohn, J. W., & Lee, D. (2007). Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *J Neurosci*, *27*(50), 13655-13666. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18077677. doi:10.1523/JNEUROSCI.2982-07.2007

Stopfer, M., & Laurent, G. (1999). Short-term memory in olfactory network dynamics. *Nature, 402*(6762), 664-668. Retrieved from <Go to ISI>://WOS:000084189800066. doi:Doi 10.1038/45244

Sussillo, D., & Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron, 63*(4), 544-557. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19709635.

Sussillo, D., & Barak, O. (2013). Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation, 25*(3), 626-649. Retrieved from <Go to ISI>://WOS:000314562800002. doi:DOI 10.1162/NECO_a_00409

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience, 18*(7), 1025-+. Retrieved from <Go to ISI>://WOS:000356866200018. doi:10.1038/nn.4042

Tanji, J., & Kurata, K. (1982). Comparison of movement-related activity in two cortical motor areas of primates. *J Neurophysiol, 48*(3), 633-653. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7131050. doi:10.1152/jn.1982.48.3.633

Tanji, J., & Mushiake, H. (1996). Comparison of neuronal activity in the supplementary motor area and primary motor cortex. *Brain Res Cogn Brain Res, 3*(2), 143-150. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8713555.

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature, 371*(6496), 413-416. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8090219. doi:10.1038/371413a0

Thaler, D., Chen, Y. C., Nixon, P. D., Stern, C. E., & Passingham, R. E. (1995). The functions of the medial premotor cortex. I. Simple learned movements. *Exp Brain Res, 102*(3), 445-460. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7737391.

Todorov, E. (2000). Direct cortical control of muscle activation in voluntary arm movements: a model. *Nature Neuroscience, 3*(4), 391-398. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10725930

Wang, J., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience, 21*(1), 102-+. Retrieved from <Go to ISI>://WOS:000423155800018. doi:10.1038/s41593-017-0028-6

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks, 1*(4), 339-356.

Williamson, R. C., Doiron, B., Smith, M. A., & Yu, B. M. (2019). Bridging large-scale neuronal recordings and large-scale network models using dimensionality reduction. *Curr Opin Neurobiol, 55*, 40-47. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30677702. doi:10.1016/j.conb.2018.12.009

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., & Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J Neurophysiol, 102*(1), 614-635. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19357332. doi:10.1152/jn.90941.2008