Interplay between external inputs and recurrent dynamics during movement preparation and execution in a network model of motor cortex.

Ludovica Bachschmid-Romano[1], Nicholas G. Hatsopoulos[2,3] and Nicolas Brunel[1,4,5,6⋆]

1 Department of Neurobiology, Duke University
2 Department of Organismal Biology and Anatomy, The University of Chicago
3 Committee on Computational Neuroscience, The University of Chicago
4 Department of Physics, Duke University
5 Duke Institute for Brain Sciences, Duke University
6 Center for Cognitive Neuroscience, Duke University ⋆ Corresponding author

Abstract

The primary motor cortex has been shown to coordinate movement preparation and execution through computations in approximately orthogonal subspaces. The underlying network mechanisms, and in particular the roles played by external and recurrent connectivity, are central open questions that need to be answered to understand the neural substrates of motor control. We develop a recurrent neural network model that recapitulates the temporal evolution of single-unit activity recorded from M1 of a macaque monkey during an instructed delayed-reach task. We explore the hypothesis that the observed dynamics of neural covariation with the direction of motion emerges from a synaptic connectivity structure that depends on the preferred directions of neurons in both preparatory and movement-related epochs. We constrain the strength both of synaptic connectivity and of external input parameters by using the data as well as an external input minimization cost. Our analysis suggests that the observed patterns of covariance are shaped by external inputs that are tuned to neurons' preferred directions during movement preparation, and they are dominated by strong direction-specific recurrent connectivity during movement execution, in agreement with recent experimental findings on the relationship between motor–cortical and motor–thalamic activity, both before and during movement execution. We also demonstrate that the manner in which single-neuron tuning properties rearrange over time can explain the level of orthogonality of preparatory and movement-related subspaces. We predict that the level of orthogonality is small enough to prevent premature movement initiation during movement preparation; however, it is not zero, which allows the network to encode a stable direction of motion at the population level without direction-specific external inputs during movement execution.

Introduction

The activity of the primary motor cortex (M1) during movement preparation and execution plays a key role in the control of voluntary limb movement [1, 2, 3, 4, 5, 6, 7]. Classic studies of motor preparation were performed in a delayed-reaching task setting, showing that firing rates correlate with task-relevant parameters during the delay period, despite no movement occurring [8, 9, 10, 11, 12, 13, 14, 14, 15, 16, 17]. More recent works have shown that preparatory activity is also displayed before non-delayed movements [18], that it is involved in reach correction [19], and that when multiple reaches are executed rapidly and continuously, each upcoming reach is prepared by the motor cortical activity while the current reach is in action [20]. Preparation and execution of different reaches are thought to be processed simultaneously without interference in the motor cortex through computation along orthogonal dimensions [20]. Indeed, the preparatory and movement-related subspaces identified by linear dimensionality reduction methods are almost orthogonal [21] so that

simple linear readouts that transform motor cortical activity into movement commands will not produce premature movement during the planning stage [22]. However, response patterns in these two epochs of motion are nevertheless linked, as demonstrated by the fact that a linear transformation can explain the flow of activity from the preparatory subspace to the movement subspace [21]. How this population-level strategy is implemented at the circuit level is still under investigation [23]. A related open question [24] is whether inputs from areas upstream to the primary motor cortex (such as from the thalamus and other cortical regions, here referred to as *external inputs*) that have been shown to be necessary to sustain movement generation [25] are specific to the type of movement being generated throughout the whole course of the motor action, or if they serve to set the initial conditions for the dynamics of the motor cortical network to evolve as shaped by recurrent connections [26, 27, 28, 29].

Here, we use a network modeling approach to explain the relationship between network recurrent connectivity, external inputs, and computations in orthogonal dimensions. Our analysis is based on recordings from M1 of a macaque monkey performing a delayed center-out reach task. One easily measurable feature of motor cortical activity during straight limb reaches is its covariance with the direction of motion (see [30, 31, 32, 33] but also [34, 35]). Recorded neurons are tuned to the direction of motion both during movement preparation and execution, but their tuning properties change over time ([36, 37, 38]). Interestingly, major changes in single neuron tuning happen when the activity flows from the preparatory to the movement-related subspaces. Yet, the direction of motion that we can decode at the population level is stable throughout the course of the motor action. To model these observations, we considered a recurrent neural network of rate-based neurons and we reduced the network dynamics to a few latent variables, or *order parameters*, that recapitulate the temporal evolution of neurons tuning properties. We then fitted the model to the data, by imposing that the model reproduce the observed dynamics of the order parameters, and that the energy cost associated with large external inputs is minimized. Our analysis suggests that during the delay period, information encoding the direction of movement is inherited by the network through external inputs that are tuned to the preferred directions of the neurons. During movement execution, external inputs are untuned and information regarding the movement direction is maintained via strong direction-specific recurrent connections. We will discuss how this prediction is in line with recent findings [39] showing that the co-firing patterns of pairs of motor-thalamic and motor-cortical neurons significantly correlate with the difference in the neurons' preferred directions before, but not during, movement execution. Finally, we show how the specific way in which neurons tuning properties rearrange over time produces the observed level of orthogonality between the preparatory- and movement-related subspaces.

Results

Subjects and Task

We analyzed multi-electrode recordings from the primary motor cortex (M1) of two macaque monkeys performing a previously reported [40] instructed-delay, center-out reaching task. The monkey's arm was on a two-link exoskeletal robotic arm, so that the position of the monkey's hand controlled the location of a cursor projected onto a horizontal screen. The task consisted of three periods (Fig. 1): a hold period, during which the monkey was trained to hold the cursor on a center target and wait 500 ms for the instruction cue; an instruction period, where the monkey was presented with one of eight evenly spaced peripheral targets and continued to hold at the center for an additional 1,000–1,500 ms; a movement period, signalled by a go cue, were the monkey initiated the reach to the peripheral target. Successful trials where the monkeys reached the target were rewarded with a juice or water reward. The peripheral target was present on the screen throughout

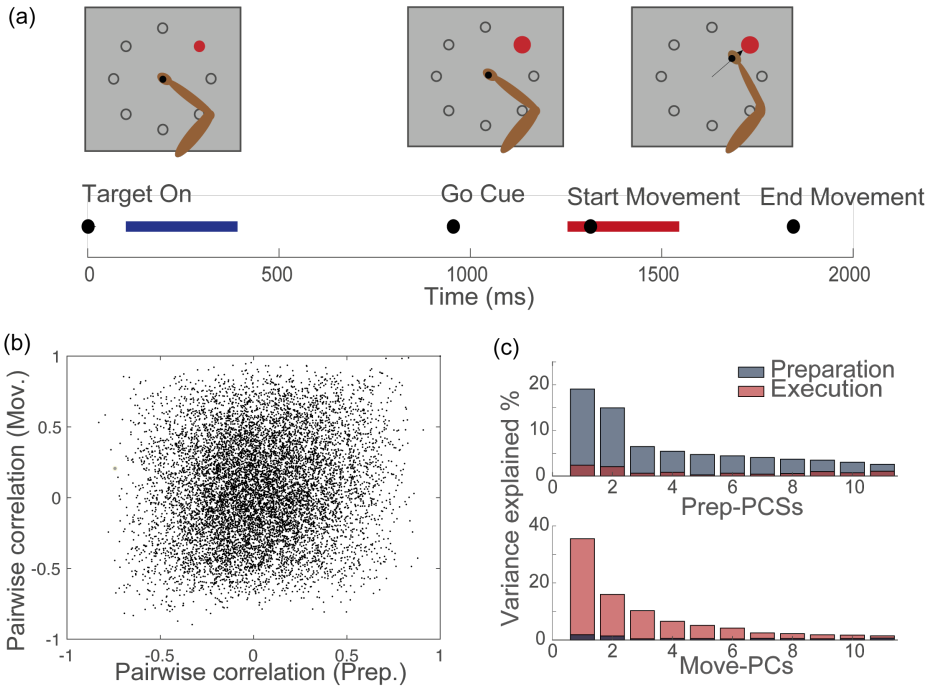


Figure 1: **(a)** Schematic of the center-out delayed reach task and definition of the preparatory (blue) and of the movement-related (red) epochs. Black circles represent the time of: target onset; go cue; start of movement; end of movement, averaged across trials. **(b)** Pairwise correlation of trial averaged preparatory activity plotted against the correlation of trial averaged movement-related activity, for each pair of neurons. The correlation coefficient of the correlations is 0.2 ($P < 10^{-5}$). **(c)** Percentage of variance of the preparatory (blue) and movement-related (red) recorded activity explained by the first 11 principal components calculated from preparatory (top) and movement-related (bottom) trial averaged activity.

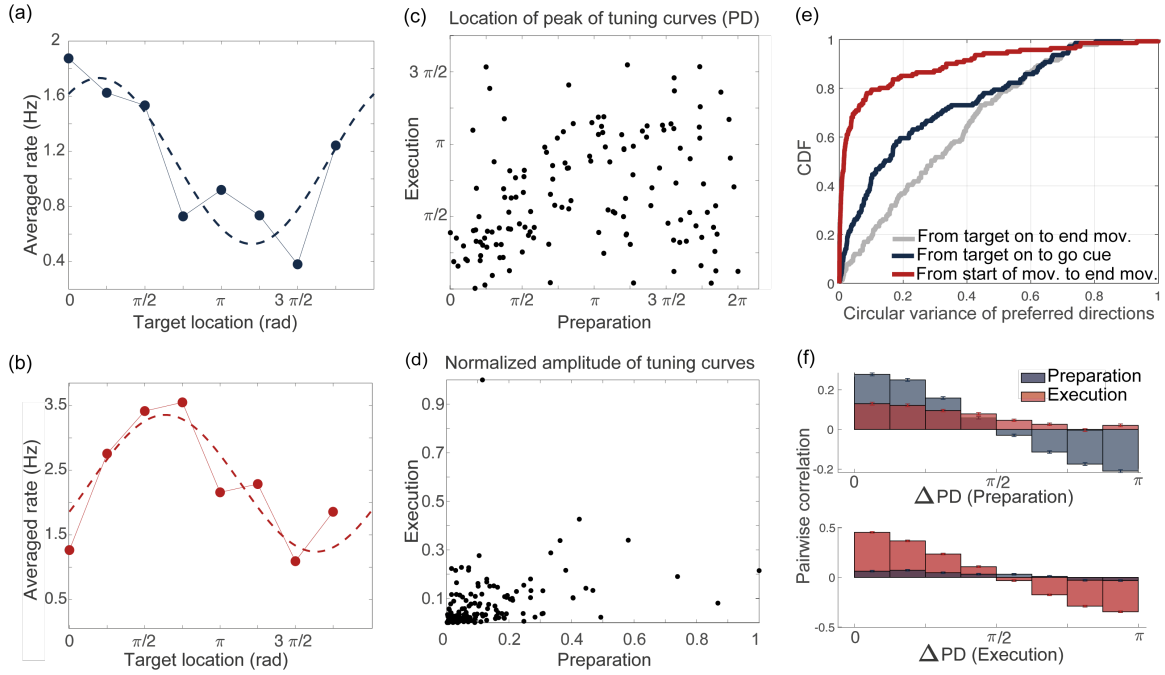


Figure 2: **(a)** Example of the trial averaged firing rate of one neuron during movement preparation (blue dots) as a function of the target location. The dotted line represent the corresponding cosine fit. **(b)** Same as in (a), but for the activity of the same neuron during movement execution. **(c)** Scatter plot between preferred directions for all neurons, defined as the location of the peak of the cosine tuning function; circular correlation coefficient: $r_\theta = 0.4$. **(d)** Scatter plot between the normalized amplitude of the cosine tuning curve for all neurons; correlation coefficient: $r_\eta \approx 0.3$. **(e)** Cumulative density function of the circular variance of preferred directions. For each neuron, preferred directions are computed in 290ms time bins across the whole duration of the task (grey line, 5 time bins); during the delay period (blue line, 3 time bins) and in the time interval going from the start to the end of the movement (red line, 2 time bins). **(f)** Pairwise correlation of trial averaged activity during movement preparation (blue) and execution (red) of pairs of neurons as a function of the difference ΔPD in their preferred directions during preparation (top) and execution (bottom).

the whole instruction and movement periods. In line with previous studies (e.g., [21]), the preparatory and movement-related epochs were defined as two 300ms time intervals beginning, respectively, 100ms after target onset and 50ms before the start of the movement.

Correlations and tuning properties of neural activity

As previously reported in [21], the correlation coefficient between the activity of pairs of neurons during movement preparation weakly correlates with the correlation coefficient of the same two neurons during movement execution (Fig. 1.b); moreover, subspaces of neural activity dedicated to movement-preparation and execution are close to being orthogonal (Fig. 1.c). We aim to understand how computations during movement preparation and execution are linked, while correlation patterns are fundamentally reorganized between epochs. To this end, we first analyzed how neurons tuning properties relate in the two epochs of motion. Studies of motor cortex have shown that tuning to movement direction is not a time-invariant property of motor cortical neurons, but rather varies in time throughout the course of the motor action; single-neuron encoding of entire movement trajectories has been also reported (e.g., [36, 37]). We measured temporal variations in neurons preferred

direction by binning time into $290ms$ time bins and fitting the binned trial-averaged spike counts as a function of movement direction with a cosine function. The cumulative density function of circular variances of preferred directions (Fig. 2.e) shows that the variability in preferred direction during the delay period alone and during movement execution alone is smaller than the variability during the entire duration of the task. This justifies our choice to characterize neurons tuning properties only in terms of their preferred direction during movement preparation and their preferred direction during movement execution. In each epoch of motion, pairs of neurons have positively correlated activity if their preferred direction is similar, and negatively correlated activity if their preferred direction is opposite (Fig. 2.f). Conversely, correlations of preparatory activities depend less strongly on the preferred direction during movement execution, and vice versa (Fig. 2.f). We next built a network model whose architecture is shaped by neurons tuning properties.

Two-cylinder model

We considered a recurrent network model where two distinct but correlated maps, denoted by A and B, encode the direction of movement during movement preparation and during movement execution, respectively. We will refer to the units in the network as neurons, even though each unit in the model rather represents a group of M1 neurons with similar functional properties, and the connection between two units in our model represents the effective connection between the two functionally similar groups of neurons in M1. The model is an extension of the ring model: while previously studied ring models are defined on one [41] or multiple [42] one-dimensional circular spaces, we added another dimension to the feature space, representing the degree of participation of each neuron to encoding the circular variable. Neurons are identified by four coordinates: $\theta_A \in [0, 2\pi)$ and $\theta_B \in [0, 2\pi)$, representing their preferred direction during movement preparation and during movement execution, and $\eta_A \in [0, 1]$ and $\eta_B \in [0, 1]$, describing how strongly the neuron participates into encoding the direction of motion in the two epochs of movement. Maps A and B are thus two cylindrical maps defined by one angular coordinate θ and one linear coordinate η. Neurons receive input currents I^{ext} both from outside of the network - either homogeneous or anisotropic in maps A and B – and recurrent inputs I^{rec}, mediated by direction-specific cortical interactions. The strength of synaptic connections from a pre-synaptic neuron with preferred directions (θ_A, θ_B) and participation strengths (η_A, η_B) to a post-synaptic neuron with preferred directions (θ'_A, θ'_B) and participation strengths (η'_A, η'_B) is denoted by $J(\theta_A, \theta'_A, \theta_B, \theta'_B, \eta_A, \eta'_A, \eta_B, \eta'_B)$. Since the two maps encode two non-concurrent features of the motor action, we assume that the couplings encode the maps additively [43]. In analogy with previously studied ring models [41, 44, 42], the interactions are given by:

$$J(\theta_A, \theta'_A, \theta_B, \theta'_B, \eta_A, \eta'_A, \eta_B, \eta'_B) = j_0 + \sum_{\nu=A,B} j_s^\nu \eta_\nu \eta'_\nu \cos(\theta_\nu - \theta'_\nu) + j_a \eta_B \eta'_A \cos(\theta_B - \theta'_A). \quad (1)$$

j_0 represents a uniform inhibitory term; j_s^A and j_s^B measure the amplitude of the symmetric connections storing map A and map B, respectively; j_a measures the amplitude of asymmetric connections from map A to map B. Later in this section (also, Fig. 3), we will describe how the network can encode the direction of movement through a localized pattern of activity. The last term of (1) contributes to align location of the bump in map B with the location of the bump in map A. In the mean-field limit of a very large network, the dynamics of the activity rate is defined by:

$$\tau \frac{d}{dt} r(x;t) = -r(x;t) + \left[I^{\text{tot}}(x;t) \right]_+ , \quad (2)$$

where $[\ ]_+$ is the threshold-linear (a.k.a. relu) function. We set the time constant to $\tau =$ 25ms, which is of the same order of magnitude of the membrane time constant, and we

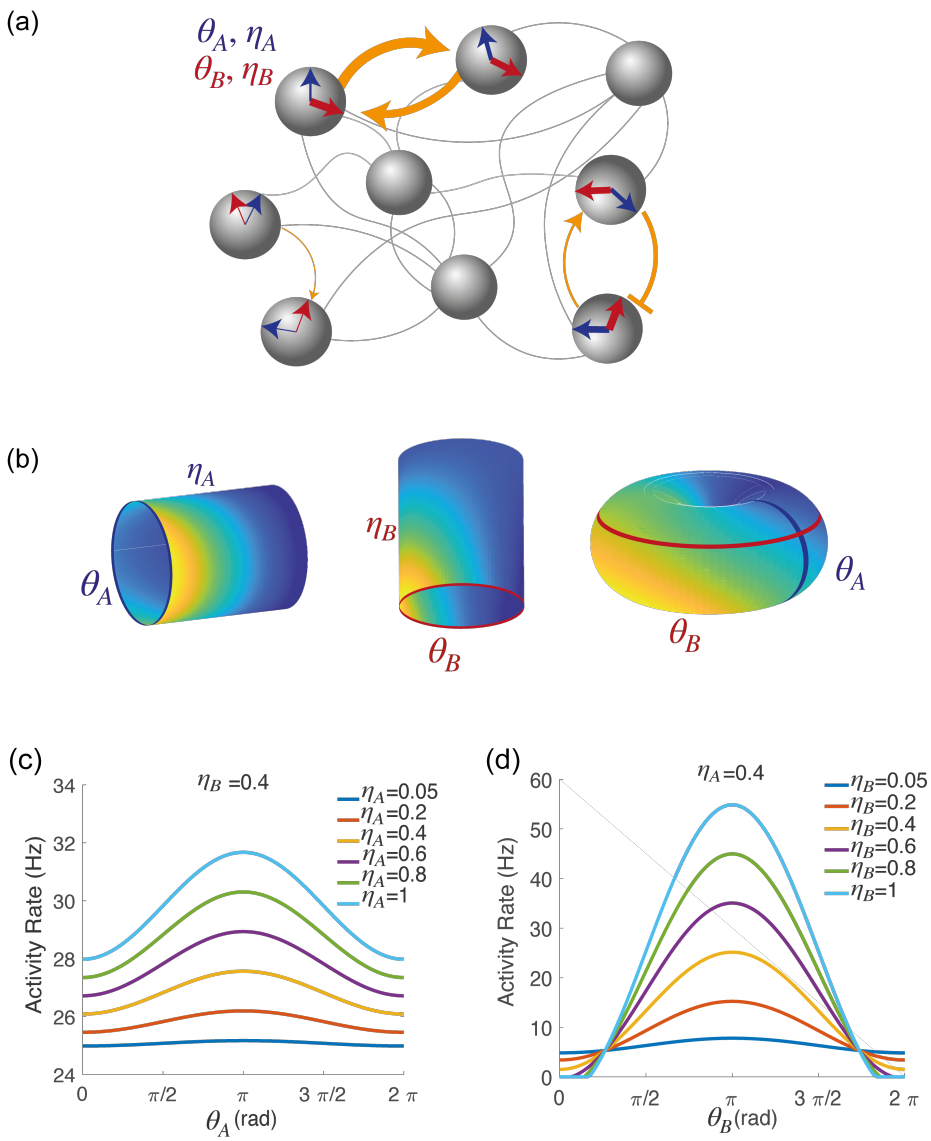


Figure 3: **(a)** Schematic of the recurrent neural network model. The model is defined on a 4−dimensional space, with coordinates: θ_A, θ_B, representing neurons preferred direction during movement preparation and execution, respectively; and η_A, η_B, representing neurons degree of participation to encoding the direction of motion during movement preparation and execution, respectively. In the schematic, the direction of the arrows represents the neurons preferred direction, while its thickness represents the degree of participation. Synaptic connectivity is defined by equation 1: it contains both a symmetric component, that depends on the distance between preferred directions of pre and post-synaptic neurons separately for the two epochs, and an asymmetric component that depends on the distance between the preferred preparatory direction of the pre-synaptic neuron and the preferred execution direction of the post-synaptic one. **(b)** Localized 4−dimensional stationary network activity shown in three different cross-sections. Left: activity plotted on map A (θ_A, η_A) at fixed θ_B, η_B. Center: activity plotted on map B (θ_B, η_B) at fixed θ_B, η_B. Right: activity plotted as a function of θ_A, θ_B, at fixed η_A, η_B;. **(c)** Localized 4−dimensional stationary network activity plotted as a function of θ_A, at fixed θ_B, η_B and for different values of η_A. The activity profiles represent tuning curves of different neurons during movement preparation. **(d)** Localized 4−dimensional stationary network activity plotted as a function of θ_B, at fixed θ_A, η_A and for different values of η_B. The activity profiles represent tuning curves of different neurons during movement execution. The couplings parameters used to generate the activity shown in **(b-c-d)** correspond to the solution **(d)** in Fig. 6

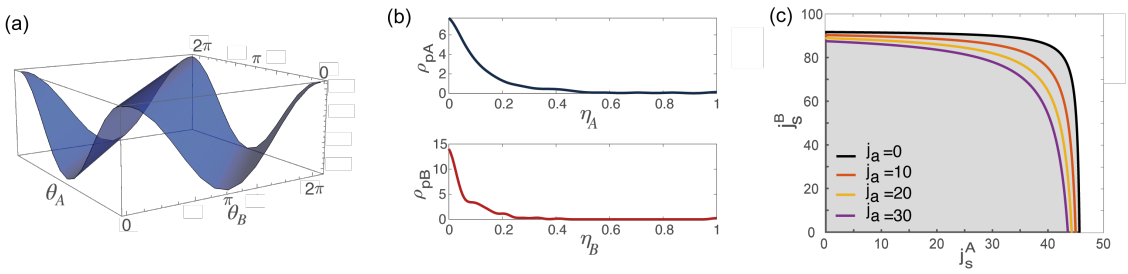


Figure 4: **(a)** Joint distribution of coordinates θ_A and θ_B inferred from the data shown in Fig. 2.b. **(b)** Distribution of coordinates η_A and η_B inferred from the data shown in Fig. 2.d. η_A and η_B are assumed to be independent. **(c)** Phase diagram of the model with homogeneous and constant external inputs, shown as a function of the parameters j_s^A, j_s^B, modulating the two symmetric terms in the couplings (see equation 1). Different curves correspond to bifurcation lines for different values of the parameter j_a, modulating the asymmetric term in the couplings. The area underneath each line (gray) corresponds to the homogeneous phase; the area beyond each line (white) corresponds to the marginal phase, where the network exhibit a narrowly localized activity pattern even in absence of external tuned inputs.

checked that for values of τ in the range $10ms - 100ms$ our results did not quantitatively change. The total input to a neuron is

$$I^{\text{tot}}(x;t) = \int dx' \rho(x') J(x,x') r(x';t) + I^{\text{ext}}(x;t), \tag{3}$$

where x represents the coordinates vector $x = \{\theta_A, \theta_B, \eta_A, \eta_B\}$ and $\rho(x)$ is the joint probability density of the preferred directions and participation strengths, that we will specify later. The mean-field formalism allowed us to reduce the dimensionality of the system to only five order parameters: the total average activity, $r_0(t)$; the spatial modulation of the activity profile in either maps, $r_A(t)$ and $r_B(t)$; the location of the peak of the activity profile in either map: ψ_A and ψ_B. The equations governing the temporal evolution of the order parameters are derived in the Methods.

We first characterize the model by studying the fixed-points of the network dynamics when subject to a constant and homogeneous external input. If the external field is independent on θ_A, θ_B, the fixed points can be of two kinds, depending on the value of the couplings parameters. One solution is homogeneous, meaning that the activity rate is independent on θ_A, θ_B; in this case, the order parameters measuring the spatial modulation of the activity are zero: $r_A = 0$ and $r_B = 0$ (see Methods for details). If the coupling parameters j_s^A, j_s^B, j_a modulating the cosine terms in the couplings (1) exceed a certain threshold, the activity rate is zero in a subset of the $(\theta_A, \theta_B, \eta_A, \eta_B)-$space. The only nonzero stationary states are localized patterns of activity (bump), which can be localized either more strongly in map A ($r_A > r_B > 0$), in map B ($r_B > r_A > 0$) or at the same level in both maps ($r_A = r_B > 0$). Since maps A and B are correlated, the location of the stationary bump of activity is constrained to be the same in the two maps: $\psi_A = \psi_B \equiv \psi$, with arbitrary ψ. The system can relax to a continuous manifold of fixed points parameterized by ψ that are marginally stable. In this continuous attractor regime, the system can store in memory any direction of motion ψ, as the activity is localized in absence of tuned inputs. In the space of parameters j_s^A, j_s^B, j_a, we computed the bifurcation surface separating the homogeneous from the bump phase. The result is shown in Fig. 4.c where for different values of j_a we plotted the bifurcation line in the space j_s^A, j_s^B, that is implicitly defined by the following

set of inequalities:

$$\begin{cases} j_a \frac{x}{4}\langle\eta_A\rangle\langle\eta_B\rangle + j_s^A \frac{1}{2}\langle\eta_A^2\rangle + j_s^B \frac{1}{2}\langle\eta_B^2\rangle < 2 \\ \left(1 - j_s^A \frac{1}{2}\langle\eta_A^2\rangle\right)\left(1 - j_s^B \frac{1}{2}\langle\eta_B^2\rangle\right) - \left(j_a + j_s^A j_s^B \frac{x}{4}\langle\eta_A\rangle\langle\eta_B\rangle\right)\frac{x}{4}\langle\eta_A\rangle\langle\eta_B\rangle > 0, \end{cases} \tag{4}$$

$\langle\dots\rangle$ denoting average over the distribution of $\theta_A, \theta_B, \eta_A, \eta_B$. Moreover, j_0 has to be smaller than 1 to avoid rate instability. Examples of the cross-section of the four-dimensional stationary activity profile $r(\theta_A, \theta_B, \eta_A, \eta_B)$ in the marginal phase are shown in Fig. 3. The activity as a function of θ_A, θ_B at fixed values of η_A, η_B can be strictly positive, with a full cosine modulation profile (e.g., Fig. 3.c-d, for $\eta_A = 0.4, \eta_B = 0.4$); this is a major difference with respect to the double ring model [42], where the activity has a cosine threshold profile at least in one of the two rings.

Time-dependent activity profiles

The activity of motor cortical neurons shows no signature of being in a stationary state but instead displays complex transients. To model the data, we assumed that the neurons are responding both to recurrent inputs and to fluctuating external inputs that can be either homogeneous or tuned to θ_A, θ_B, with peak at constant location $\Phi_A = \Phi_B \equiv \Phi$. In the limit of an infinitely large network, we derived a simplified description of the dynamics, where recurrent inputs are replaced by *effective local inputs*, that take into account the average effect of the many recurrent inputs the neuron is receiving. The total input in (3) is rewritten as the sum three local inputs: the first one is homogeneous, the second is tuned to map A and the third is tuned to map B:

$$I^{\text{tot}}(\theta_A, \theta_B, \eta_A, \eta_B; t) = I_0(t) + I_A(t)\,\eta_A \cos(\theta_A - \Phi) + I_B(t)\,\eta_B \cos(\theta_B - \Phi), \tag{5}$$

where

$$\begin{aligned} I_0(\eta_A, \eta_B; t) =& C_0(t) + \eta_A C_A(t) + \eta_B C_B(t) + j_0 r_0(t) \\ I_A(t) =& j_s^A r_A(t) + \epsilon_A(t) \\ I_B(t) =& j_s^B r_B(t) + j_a r_A(t) + \epsilon_B(t), \end{aligned} \tag{6}$$

and where we have introduced the external fields parameters $C_0, C_A, C_B, \epsilon_A, \epsilon_B$ representing, respectively, the magnitude of: an homogeneous input; an untuned preparatory input (proportional to η_A but homogeneous in θ_A); an untuned execution input (proportional to η_B but homogeneous in θ_B); a tuned preparatory input (proportional to η_A and tuned to θ_A); and a tuned execution input (proportional to η_B and tuned to θ_B). Equations (2) and (5) show that the pattern of activity is localized at a constant location Φ throughout the dynamics, while its shape changes in time. For example, the activity pattern can evolve from being strongly localized in map A to being strongly localized in map B; accordingly, the order parameters measuring the spatial modulation in either map, r_A, r_B, are also time dependent. In order to draw a correspondence between the model and the data, we note that the activity of each neuron in the data corresponds to the activity rate at a specific coordinate $\theta_A, \theta_B, \eta_A, \eta_B$ in the model. From equations (2) and (5) we see that, if we assume that changes in the external inputs happen on a time scale larger than τ, the modulation of the activity profile in map A at each time t is centered at $\theta_A - \Phi$ and has amplitude proportional to η_A, while the modulation of the activity profile in map B is centered at $\theta_B - \Phi$ and has amplitude proportional to η_B. Accordingly, for each recorded neuron θ_A and θ_B represent the location of the peak of the neuron's tuning curve computed during the preparatory and movement-related epoch, respectively, while η_A and η_B are proportional to the amplitude of the tuning curves. From the empirical distribution of the amplitude and location of the tuning curves that we measured from data, we inferred the probability

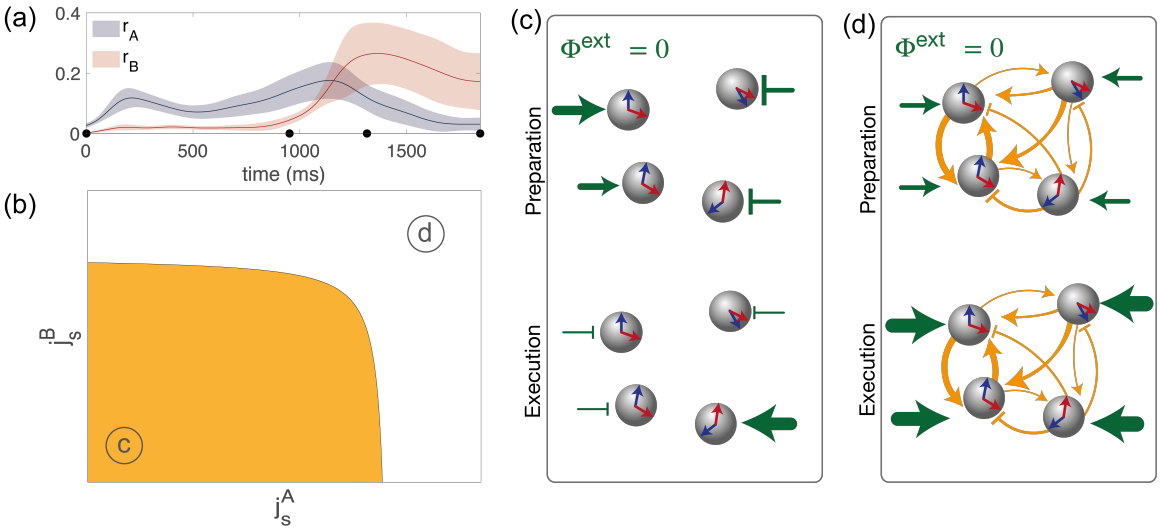


Figure 5: **(a)** Dynamics of the order parameters r_A (degree of spatial modulation of the activity in map A) and r_B (degree of spatial modulation in map B) computed from data (solid line; shaded area: $\pm$ SEM across the population). Black dots on the x-axes represent the trial-averaged time of: target onset, go cue, start of movement and end of movement. **(b)** Our model can qualitatively explain the switching trajectories of the order parameters for different values of the coupling parameters j_s^A, j_s^B, j_a. We illustrate here the schematic of two opposite scenarios: very small couplings (c), and very strong couplings, away from the bifurcation line (d). **(c)** Very small couplings scenario: the observed dynamics of the order parameters can result from external inputs that are tuned to θ_A during movement preparation and to θ_B during movement execution. **(d)** Very strong couplings scenario: the strong recurrent connections sustain the localized pattern of activity; a change in *untuned* external inputs makes the activity to be localized along the θ_A during preparation and along θ_B during preparation.

density of the four coordinates $\rho(\theta_A, \theta_B, \eta_A, \eta_B)$ that we used in our mean-field analysis (see Fig. 4 and Methods for details). It factorizes as follow:

$$\rho(\theta_A, \theta_B, \eta_A, \eta_B) = \rho_d(\theta_A, \theta_B)\ \rho_{pA}(\eta_A)\ \rho_{pB}(\eta_B); \tag{7}$$

where the distribution of preferred directions is well fitted by:

$$\rho_d(\theta_A, \theta_B) = \frac{1}{4\pi^2}\left[1 + \frac{2}{3}\cos(\theta_A - \theta_B)\right], \tag{8}$$

as shown in (Fig. S1), and where $\rho_{pA}(\eta_A), \rho_{pB}(\eta_B)$ are estimated non-parametrically using kernel density estimation. Next, we notice that either of the two fields $I_A(t)$ and $I_B(t)$ modulating the amplitude of the activity are a sum of two terms (6): the strength of tuned external inputs and the strength of recurrent inputs, which is proportional to the couplings parameters j_s^A, j_s^B, j_a. Hence, the localized pattern of activity can be sustained by either strongly tuned external inputs or by strong recurrent connections.

External inputs inferred from the dynamics of the order parameters

The dynamics of the order parameters r_A, r_B computed from data (Fig.5) shows that during the delay period neural activity is spatially modulated in map A but not in map B, while during movement execution the activity is strongly modulated in map B, with the degree of modulation in map A slowly decreasing to very small values. This behaviour can be reproduced by the model for different choices of recurrent connections and external inputs parameters, ranging in between the following two opposite scenarios (Fig.5):

- The couplings are not direction specific ($j_s^A = j_s^B = j_a = 0$) and the localized activity is sustained by an external input that is tuned to θ_A during movement preparation and to θ_B during movement execution.

- The bump is self-sustained via strong recurrent connections and fluctuating homogeneous external inputs allow the bump to be localized in map A during movement preparation and in map B during movement execution.

The value of the parameters that best fit the data are inferred through minimization of a cost function (E_{tot}) composed of two terms: one is the reconstruction error of the temporal evolution of the order parameters (E_{rec}) and the other represents an energetic cost penalizing large external inputs (E_{ext}):

$$E_{\text{tot}} = \alpha E_{\text{rec}} + E_{\text{ext}},$$

where α is an hyperparameter of the fitting algorithm. To reduce the degeneracy in the results of the fit, we impose not only that the model reconstruct the value of the order parameters r_0, r_A, r_B, but also the value of two additional parameters. We denote them by r_{0A} and r_{0B} and they represent the overlap between the variable η_A (or η_B) and the neural activity. The result of the fit depends on the hyperparameter α (Fig.5). If α is chosen so that the two terms αE_{rec} and E_{ext} have equal magnitude – after the cost function is minimized – then the inferred values of the couplings parameters are above, but very close to, the bifurcation surface. In absence of tuned external inputs, the activity sustained by recurrent connections is localized in map B. The inferred value of the external fields is shown in Fig. 6.e. During the delay period, an external input *tuned* to map A sustains the bump of activity localized in map A and sets the direction of motion; during movement execution, the activity is driven by homogeneous external inputs and the activity localized in map B is sustained by the strong recurrent connections. The correlation between the two maps allows the information about the direction of motion to be encoded stably throughout the whole duration of the task. On the other hand, if we chose a larger value of the hyperparameter α, so to impose a better reconstruction of the recorded dynamics but penalize less for large external inputs, the solutions that we obtain are still close to the bifurcation surface but below it, and a small input tuned to θ_B is present during movement execution. Importantly, the same analysis applied to a second dataset recorded from a different macaque monkey performing the same task yielded qualitatively similar results (see Fig. S3).

Simulations of the dynamics of the network

The fitting procedure and the results of Fig. 5 are based on the equations governing the dynamics of the order parameters that we derived in the mean-field limit of infinitely many neurons. We next checked that a network of finite size with the same parameters that we inferred in the last section reproduces the dynamics seen in the data. We built a network of 16000 neurons by assigning to each neuron i the coordinates $\theta_A^{(i)}, \theta_B^{(i)}, \eta_A^{(i)}, \eta_B^{(i)}$ so to match the empirical distribution of coordinates of Fig. 3. In particular, the values of $\theta_A^{(i)}$ and $\theta_B^{(i)}$ are chosen to be equally spaced along the lines $\theta_A^{(i)} - \theta_B^{(i)} = \text{cost}$ (see Fig. S5). Neuronal firing rates obeyed standard rate equations (29), were a noise term modelled as an Ornstein-Uhlenbeck process is added to the total input. Simulations of a network whose coupling parameters correspond to the solution of Fig. 6d reproduce well the dynamics of the order parameters (Fig. 7) and the location of the bump fluctuates only weakly around a stable location (Fig. S5). We analyzed the results of the simulations similarly as how we analyzed the data. In particular, we computed tuning curves during the preparatory and execution epochs and estimated the values $\theta_A^{(i)}, \theta_B^{(i)}, \eta_A^{(i)}, \eta_B^{(i)}$ from the location and the amplitude of the tuning functions. The reconstructed values of neurons tuning parameters are consistent with the values initially assigned to them when we built the network (Fig. S5); moreover, simulating the dynamics with additive noise produces tuning curves whose shape resembles

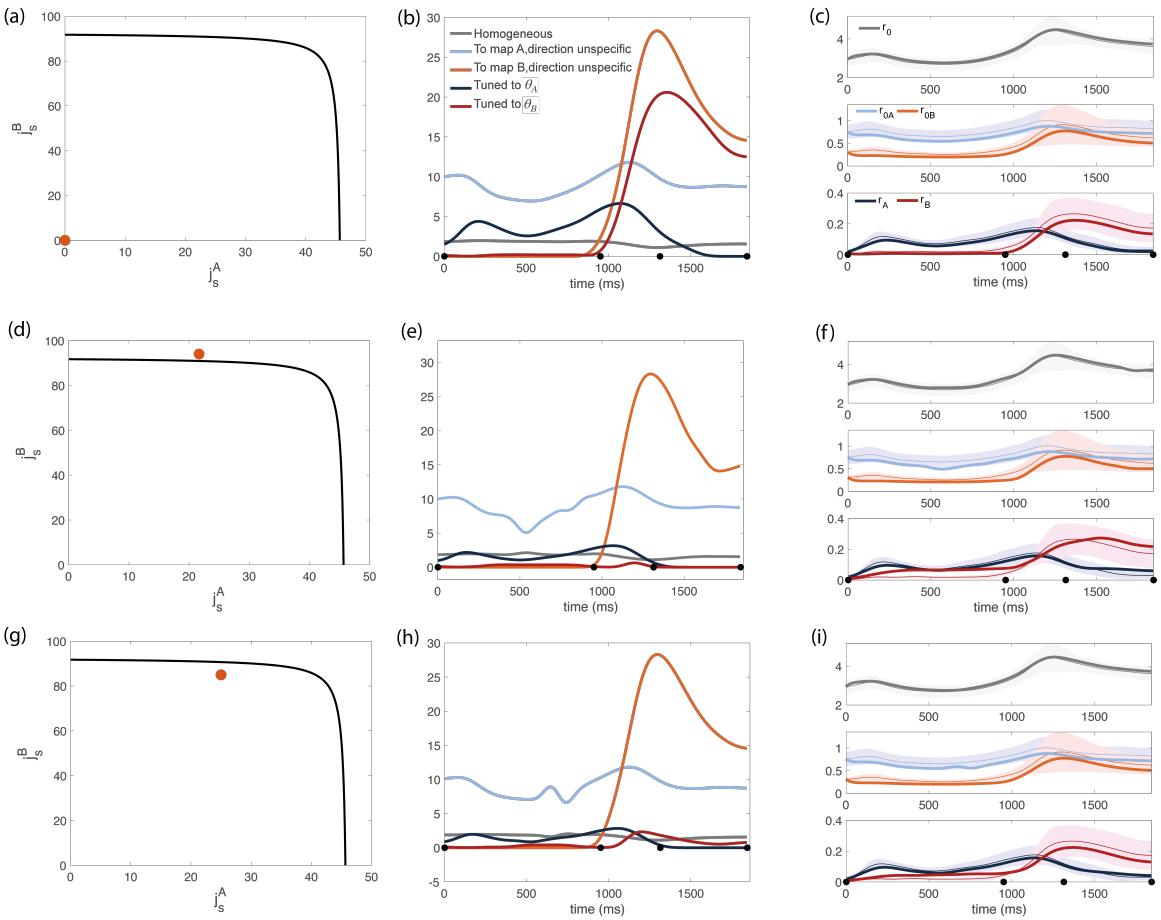


Figure 6: Inferred dynamics of the external fields required to sustain the observed dynamics of the order parameters. **(a-b-c)** Scenario where neurons are connected by uniform inhibitory connections, in absence of direction specific couplings. **(a)** Couplings parameters j_s^A, j_s^B, j_a are set to zero (orange dot on the phase diagram). The black line represents the bifurcation surface in the space j_s^A, j_s^B, at $j_a = 0$. **(b)** Dynamics of the external inputs inferred from the data. Gray line: homogeneous input; light blue (\ light red) line: input that is direction unspecific, but proportional to η_A (\$\eta_B$); dark blue (\ dark red) line: input that is tuned to the preferred direction θ_A (\$\theta_B$) and proportional to η_A (\$\eta_B$). Black dots on the x-axes represent the trial-averaged time of: target onset, go cue, start of movement and end of movement. **(c)** Dynamics of the order parameters $r_0, r_{0A}, r_{0B}, r_A, r_B$ computed from data (thin line; shaded area: $\pm$ standard error across the population) and analytical prediction from mean-field analysis (thick line). **(d-e-f-g-h-i)** Both couplings parameters and external inputs are inferred from data, through minimization of a cost function composed of two terms: $E_{\text{tot}} = \alpha E_{\text{rec}} + E_{\text{ext}}$. E_{rec} represents the reconstruction error of the order parameters, while E_{ext} is the magnitude of the external inputs required to sustain the activity. α is an hyperparameter of the fitting algorithm. **(d)** Value of the couplings parameters (orange dot) inferred by choosing the hyperparameter α in such a way that the two terms of the cost function have equal magnitude. The obtained couplings parameters are above, but close to, the bifurcation line. **(e)** Dynamics of the external inputs inferred from the data. **(f)** Dynamics of the order parameters r_0, r_A, r_B computed from data (thin line) analytical prediction from mean-field analysis (thick line). **(g)** Value of the couplings parameters (orange dot) inferred by choosing the hyperparameter α in such a way that the term E_{rec} has a slightly larger weight than the term E_{ext} in the cost function. The obtained couplings parameters are below, but close to, the bifurcation line. **(h)** Dynamics of the external inputs inferred from the data. **(i)** Dynamics of the order parameters r_0, r_A, r_B computed from data (thin line) analytical prediction from mean-field analysis (thick line).

the data (Fig. S6). Next, we added noise to the variables $\theta_A^{(i)}, \theta_B^{(i)}$, to break the rotational symmetry of the model. If the network has coupling parameters above the bifurcation surface and no tuned external input is present during movement execution, the location of the bump after movement initiation starts to drift towards a few discrete attractors. When the noise is weak, the drift happen on a time scale that is much larger that the time of movement execution; the larger the level of the noise, the faster the drift. If we instead consider a solution like the one of Fig. 6.g, where tuned inputs are weak but non-zero, the location of the bump remains stable throughout the dynamics. Hence, adding stability constraints to the cost function that we used for our minimization procedure will favor network parameters that are close to, but below the bifurcation surface.

PCA subspaces dedicated to movement preparation and execution

We performed principal component analysis (PCA) on the trial averaged activity to show that the two subspaces that capture most of the variance of neural activity during movement preparation and execution are close to being orthogonal. After identifying the preparatory and movement-related principal components (PCs) (see Methods for details), we quantified the level of orthogonality of the two supspaces by the aligment index as defined in [21], that measures the percentage of variance of each epoch's activity explained by both sets of PCs. Fig 7 shows that the alignment index is much smaller than the one computed between two subspaces drawn at random (random alignment index, explained in the Methods), both for the simulations and for the data. The level of orthogonality of the preparatory and movement-related subspaces can be understood as follows. The activity of the ring model encoding the value of a singular angular variable is two-dimensional in the Euclidean space. Similarly, the activity of the double-ring model encoding two distinct circular maps is four-dimensional. Our model is an extension of the double-ring model, where both the connectivity matrix and the external fields () are a sum of several terms, each one composed of an η-dependent term multiplying a θ-dependent term. The connectivity matrix is still rank-four, but its eigenvectors are modulated by the η variables. Although the dynamics is four-dimensional, we have shown that during movement preparation the activity is localized only in map A, while during movement execution it is predominantly localized in map B: in either epoch, we expect only two eigenvalues to explain most of the activity variance, as we indeed see from simulations in absence of additive noise (Fig. S7). The degree of orthogonality between the preparatory and movement-related subspaces is determined by the level of orthogonality between map A and map B, which can be quantified in terms of the following correlation:

$$C_{AB} = \frac{\langle \eta_A \cos(\theta_A) \eta_B \cos(\theta_B) \rangle}{\sqrt{\langle \eta_A^2 \cos^2(\theta_A) \eta_B^2 \cos^2(\theta_B) \rangle}} = \frac{\langle \eta_A \rangle \langle \eta_B \rangle}{3\sqrt{\langle \eta_A^2 \rangle \langle \eta_B^2 \rangle}} \sim 0.19, \tag{9}$$

where brackets $\langle \dots \rangle$ denote the average with respect to the distribution of (7-8), and where we used the distribution of η_A, η_B that we inferred from data (Fig. 4) to compute the last term on the right-hand side. The correlation between maps is smaller with respect to both the case where $\eta_A = 1, \eta_B = 1$ for all neurons ($C_{AB} = 0.33$) and the case where η_A, η_B are uniformly distributed ($C_{AB} = 0.25$). Finally, the noise term added to the dynamics introduces extra random dimensions; as the dynamics gets higher dimensional, both the alignment index and the random alignment index get smaller.

To conclude, we observed that from the PCA analysis alone it is not possible to discriminate between the scenario where neurons tuning is induced by tuned external inputs from the one where it emerges from recurrent connections. Simulations of the network dynamics in the two scenarios yielded very similar results in terms of the variance captured by the first principal components (Fig. S7).

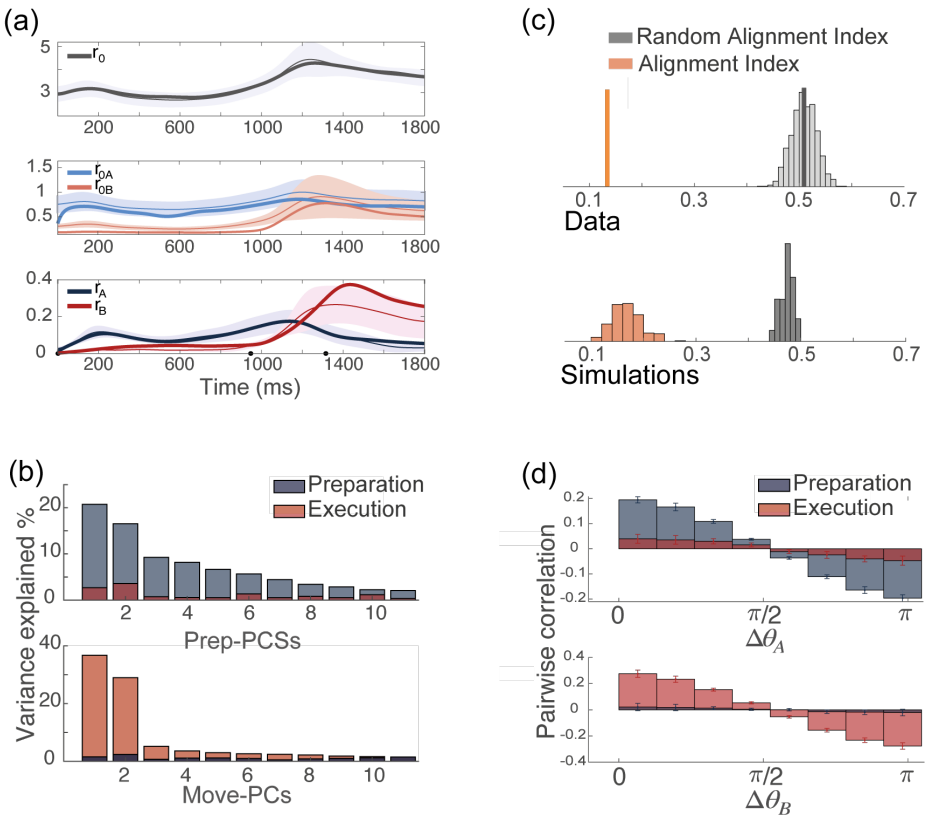


Figure 7: **(a)** Dynamics of the order parameters r_0, r_A, r_B computed from data (thin line; shaded area: $\pm$ SEM of the population) and numerical simulations (thick line) of the dynamics of a finite-size network model with additive noise. The couplings parameters of the network correspond to the scenario of Fig. 6, panel d: they are above, but close to, the bifurcation line. **(b)** Percentage of variance of the preparatory (blue) and movement-related (red) activity from simulations explained by the first 11 principal components calculated from preparatory (top) and movement-related (bottom) trial-averaged activity. **(c)** The alignment index quantifies the degree of orthogonality between two subspaces. Top: alignment index between the preparatory and movement-related activities computed from data, compared to the randomized test (random alignment index, distribution in light gray and average in dark grey). Bottom: alignment index computed from simulations, for 100 subsets of 140 neurons each, sampled uniformly at random from the larger network we simulated; the gray histogram shows the average random alignment index for all subsets. **(d)** Pairwise correlation of the trial-averaged activity from simulations during movement preparation (blue) and execution (red) of pairs of neurons as a function of the difference in their preferred directions $\Delta\theta_A$ (top) and $\Delta\theta_B$ (bottom).

Discussion

Studies on the dynamics of motor cortical activity during delayed reach tasks have shown that the primary motor cortex employs an 'orthogonal but linked strategy' [22, 21] to coordinate planning and execution of movements. In this work, we explored the hypothesis that this strategy emerges as result of a specific Hebbian-like recurrent functional architecture, in which synaptic connections store information about two distinct patterns of activity that underlie movement preparation and movement execution. In a simplifying modeling setting based on recordings in Macaque monkeys performing a straight reach task, we characterized response patterns in terms of their covariance with the direction of motion. Hence, the preparatory and movement related patterns stored in the couplings are formalized in terms of two *maps* (A and B) in the space defined by two features of neurons tuning properties: their preferred direction and their level of participation to the population encoding. We inferred the distribution of these tuning features from data and showed that the degree of correlation between the two maps is small enough to allow for almost orthogonal subspaces, which is thought to be important for the preparatory activity not to cause premature movement; at the same time, having a non-zero correlation between maps allows the activity to flow from the preparatory to the movement-related subspaces with minimal external inputs.

By using a mean field analysis, we derived a simple description of neural dynamics in terms of a few order parameters that can be easily computed from data. Different combinations of the strength of direction-specific recurrent connections and of tuned external inputs allow the model to accurately reproduce the dynamics of the order parameters, ranging from a scenario where neurons tuning properties emerge solely from recurrent inputs to one where the motor cortex is simply integrating tuned external inputs. The addition of an external input strength minimization constraint breaks the degeneracy of the space of solutions, leading to a solution where synaptic couplings depend on the tuning properties of the pre- and post- synaptic neurons, in such a way that in the absence of a tuned input, neural activity is localized in map B. During movement preparation, an external input tuned to map A sustains a localized activity in map A, and sets the direction of motion encoded at the population level. During movement execution, movement direction is stably encoded at the population level, thanks to recurrent inputs that sustain a localized activity in map B. The correlation between maps A and B allows the activity in map B during movement execution to be localized around the same location as it was in map A during movement preparation. These results, based on a mean-field analysis valid for infinitely large networks, were confirmed by simulating the dynamics of a finite-size network of 16000 neurons, with additive noise whose level was chosen so that the dimensionality of the activity in the neural state space matched the one from data. The presence of noise did not disrupt the temporal evolution of the order parameters; on single trials, the direction of motion encoded by the network slightly diffuses around the value predicted by the mean field analysis, but remains stable for trial-averaged activity. However, our solution requires an implausible fine tuning of the recurrent connections. Heterogeneity in the connectivity causes a systematic drift of the encoded direction of motion on a typical time scales of seconds - the larger the structural noise in the couplings, the faster the drift, as has been extensively studied in the literature of continuous attractor models [45, 46, 47, 48, 49]. It has been shown that homeostatic mechanisms could compensate for the heterogeneity in cellular excitability and synaptic inputs to reduce systematic drifts of the activity [47] and that short-term synaptic facilitation in recurrent connections could also significantly improve the robustness of the model [48] – even when combined with short-term depression [49]. While a full characterization of our model in the presence of structural heterogeneity is beyond the scope of this work, we considered a second version of the optimization procedure to infer the model parameters that takes into account the stability of the encoded direction of motion with respect to perturbations in the couplings. We showed that a solution that improves the stability is one where tuned inputs are also present during movement execution. Interestingly, the inferred tuned

inputs are still much weaker than the untuned ones, during movement execution. Also, the inferred direction-specific couplings are strong and amplify the weak external inputs tuned to map B, therefore still playing a major role into shaping the observed dynamics during movement execution.

Our prediction that external inputs are direction-specific during movement preparation but non-specific during movement execution agrees with several other studies on the activity of the primary motor cortex during limb movement. In particular, [28] showed that the changes in neural activity that characterize the transition from movement preparation to execution reflect when movement is made but are invariant to movement direction and type; [24] argued that external input transients signalling the motor cortical network a new movement direction are more strongly time-locked to target appearance as compared to the start of movement; finally, [39] measured the correlations in firing between motor thalamic (m-Th) and motor cortical cells, in monkeys performing the same task considered in the present work. In order to investigate if the observed cofiring patterns resulted from neurons tuning properties, the authors looked at the effect of the difference in preferred directions on cofiring patterns between couples of M1–mTh cells. Interestingly, the study reported significant correlation between cofiring patterns and difference in preferred direction solely before movement, but not during movement execution (see Appendix, Fig. S8 of ref. [39]).

Comparison with the ring model

The idea that the tuning properties of motor cortical neurons could emerge from direction-specific synaptic connections goes back to the work of Lukashin and Georgopulos [50]. However, it was with the theoretical analysis of the so called ring model [51, 41, 52, 44] that localized patterns of activity were formalized as attractor states of the dynamics in networks with strongly specific recurrent connections. Related models were later used to describe maintenance of internal representations of continuous variables in various brain regions [53, 54, 55, 56, 57, 58, 59, 43, 60, 61] and were extended to allow for storage of multiple continuous manifolds [62, 42, 63] to model the firing patterns of place cells the hippocampus of rodents exploring multiple environments. While our formalism is built on the same theoretical framework of these works, we would like to stress two main differences between our model and the ones previously considered in the literature. First, we studied the dynamic interplay between fluctuating external inputs and recurrent currents, that causes the activity to flow from the preparatory map to the movement-related one and, consequently, neurons tuning curves and order parameters to change over time, while maintaining stable encoding of the direction of motion at the population level. Moreover, we introduced an extra dimension representing the degree of participation of single neurons to the population encoding of movement direction. We have discussed how the presence of this dimension is key to having tuning curves whose shape resembles the one computed from data, and decreases the level of orthogonality between the subspaces dedicated to the preparatory and movement-related activity.

Extension to modeling richer movements

By analysing neural patterns of covariation with the direction of motion in the simple context of delayed straight reaches, we were able to model the temporal evolution of neural trajectories in low-dimensional latent spaces. However, neurons directional tuning properties have been shown to be influenced by many contextual factors that we neglected in our analysis [64, 65], to depend on the acquisition of new motor skills [66, 67, 68] and other features of movement such as the shoulder abduction/adduction angle even for similar hand kinematic profiles [34, 35]. Moreover,a large body of work [69, 35, 70, 64, 71, 72, 73] has shown that the activity in the primary motor cortex covary with many parameters of movement other than the hand kinematics – for a review, see [74]. More recent studies have also provided evidence that encoding of movement-related variables is insufficient to explain the

rich dynamics of neurons in M1 [75, 76, 73, 77]. These and other works [74] argue that the largest signals in motor cortex might not 'represent' task-relevant variables at all, but if they do, surely not in a simple way that allows for different features to be decoded by linear decoders [77]. Based on these observations, we will now speculate on two possible ways of extending our model to more realistic scenarios.

In the current model, maps A and B each live in the low-dimensional space defined by neurons preferred directions: θ_A for map A and θ_B for map B. In a more general model, maps A and B could be parameterized by additional latent variables other than movement direction. For different tasks, the external input would have a different structure and select the dimensions that are relevant for the required pattern of kinematics and muscle activation. Neurons tuning to one single feature of the motor action would then vary across tasks in a highly non-linear manner, depending on the other latent variables relevant to that type of movement.

Another simplifying assumption allowed by the study of a delayed straight limb task is that the motor action can be clearly separated into a preparatory and movement related phase, where the direction of motion encoded at the population level is constant. We hypothesize that our simple model could also explain encoding of the direction of motion during more complex trajectories, that can be seen as a concatenation of short segments [20], where the network keeps an internal representation of the ongoing movement direction in map B while the next movement in a different direction is being prepared in map A. As suggested in [20], this kind of analysis can also provide a framework to interpret the population encoding of dynamic hand trajectories [78, 79, 36], that could emerge from the superposition of preparatory and movement-related activities relative to consecutive segments.

Acknowledgements

We thank Jason MacLean, Alex P. Vaz, Subhadra Mokashe, Alessandro Sanzeni for helpful discussions, and Stephen H. Scott for pointing the work of [39] to our attention. This work has been supported by NIH R01NS104898.

Author contributions

LBR and NB conceptualized the project. LBR performed the analytical calculations, implemented the computer codes and supporting algorithms. NB supervised the project. NGH provided the data. LBR and NB wrote the paper, and all authors edited the manuscript.

References

[1] Edward V Evarts. Relation of pyramidal tract activity to force exerted during voluntary movement. *Journal of neurophysiology*, 31(1):14–27, 1968.

[2] Ian Q Whishaw, Sergio M Pellis, Boguslaw Gorny, Bryan Kolb, and Wolfram Tetzlaff. Proximal and distal impairments in rat forelimb use in reaching follow unilateral pyramidal tract lesions. *Behavioural brain research*, 56(1):59–76, 1993.

[3] Ian Q Whishaw. Loss of the innate cortical engram for action patterns used in skilled reaching and the development of behavioral compensation following motor cortex lesions in the rat. *Neuropharmacology*, 39(5):788–805, 2000.

[4] Michael SA Graziano, Charlotte SR Taylor, and Tirin Moore. Complex movements evoked by microstimulation of precentral cortex. *Neuron*, 34(5):841–851, 2002.

[5] Thomas C Harrison, Oliver GS Ayling, and Timothy H Murphy. Distinct cortical circuit mechanisms for complex forelimb movement and motor map topography. *Neuron*, 74(2):397–409, 2012.

[6] Stephen H Scott. The computational and neural basis of voluntary motor control and planning. *Trends in cognitive sciences*, 16(11):541–549, 2012.

[7] Andrew R Brown and G Campbell Teskey. Motor cortex is functionally organized as a set of spatially distinct representations for complex movements. *Journal of Neuroscience*, 34(41):13574–13585, 2014.

[8] Doug P Hanes and Jeffrey D Schall. Neural control of voluntary movement initiation. *Science*, 274(5286):427–430, 1996.

[9] JUN Tanji and Edward V Evarts. Anticipatory activity of motor cortex neurons in relation to direction of an intended movement. *Journal of neurophysiology*, 39(5):1062–1068, 1976.

[10] Mark M Churchland, Afsheen Afshar, and Krishna V Shenoy. A central source of movement variability. *Neuron*, 52(6):1085–1096, 2006.

[11] Mark M Churchland, Gopal Santhanam, and Krishna V Shenoy. Preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach. *Journal of neurophysiology*, 96(6):3130–3146, 2006.

[12] Julie Messier and John F Kalaska. Covariation of primate dorsal premotor cell activity with direction and amplitude during a memorized-delay reaching task. *Journal of neurophysiology*, 84(1):152–165, 2000.

[13] Michael C Dorris, Martin Pare, and Douglas P Munoz. Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. *Journal of Neuroscience*, 17(21):8566–8579, 1997.

[14] Paul W Glimcher and David L Sparks. Movement selection in advance of action in the superior colliculus. *Nature*, 355(6360):542–545, 1992.

[15] ROBERT H Wurtz and MICHAEL E Goldberg. Activity of superior colliculus in behaving monkey. 3. cells discharging before eye movements. *Journal of Neurophysiology*, 35(4):575–586, 1972.

[16] Timothy R Darlington, Jeffrey M Beck, and Stephen G Lisberger. Neural implementation of bayesian inference in a sensorimotor behavior. *Nature neuroscience*, 21(10):1442–1451, 2018.

[17] Timothy R Darlington and Stephen G Lisberger. Mechanisms that allow cortical preparatory activity without inappropriate movement. *Elife*, 9:e50962, 2020.

[18] Antonio H Lara, Gamaleldin F Elsayed, Andrew J Zimnik, John P Cunningham, and Mark M Churchland. Conservation of preparatory neural events in monkey motor cortex regardless of how movement is initiated. *elife*, 7:e31826, 2018.

[19] K Cora Ames, Stephen I Ryu, and Krishna V Shenoy. Simultaneous motor preparation and execution in a last-moment reach correction task. *Nature communications*, 10(1):1–13, 2019.

[20] Andrew J Zimnik and Mark M Churchland. Independent generation of sequence elements by motor cortex. *Nature neuroscience*, 24(3):412–424, 2021.

[21] Gamaleldin F Elsayed, Antonio H Lara, Matthew T Kaufman, Mark M Churchland, and John P Cunningham. Reorganization between preparatory and movement population responses in motor cortex. *Nature communications*, 7(1):1–15, 2016.

[22] Matthew T Kaufman, Mark M Churchland, Stephen I Ryu, and Krishna V Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature neuroscience*, 17(3):440–448, 2014.

[23] Ta-Chu Kao, Mahdieh S Sadabadi, and Guillaume Hennequin. Optimal anticipatory control as a theory of motor preparation: a thalamo-cortical circuit model. *Neuron*, 109(9):1567–1581, 2021.

[24] Peter J Malonis, Nicholas G Hatsopoulos, Jason N MacLean, and Matthew T Kaufman. M1 dynamics share similar inputs for initiating and correcting movement. *bioRxiv*, 2021.

[25] Britton A Sauerbrei, Jian-Zhong Guo, Jeremy D Cohen, Matteo Mischiati, Wendy Guo, Mayank Kabra, Nakul Verma, Brett Mensh, Kristin Branson, and Adam W Hantman. Cortical pattern generation during dexterous movement is input-driven. *Nature*, 577(7790):386–391, 2020.

[26] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[27] Krishna V Shenoy, Maneesh Sahani, and Mark M Churchland. Cortical control of arm movements: a dynamical systems perspective. *Annual review of neuroscience*, 36:337–359, 2013.

[28] Matthew T Kaufman, Jeffrey S Seely, David Sussillo, Stephen I Ryu, Krishna V Shenoy, and Mark M Churchland. The largest response component in the motor cortex reflects movement timing but not movement type. *Eneuro*, 3(4), 2016.

[29] Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

[30] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[31] Andrew B Schwartz, Ronald E Kettner, and Apostolos P Georgopoulos. Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement. *Journal of Neuroscience*, 8(8):2913–2927, 1988.

[32] Apostolos P Georgopoulos, Joseph T Lurito, Michael Petrides, Andrew B Schwartz, and Joe T Massey. Mental rotation of the neuronal population vector. *Science*, 243(4888):234–236, 1989.

[33] Apostolos P Georgopoulos, Masato Taira, and Alexander Lukashin. Cognitive neurophysiology of the motor cortex. *Science*, 260(5104):47–52, 1993.

[34] Stephen H Scott and John F Kalaska. Reaching movements with similar hand paths but different arm orientations. i. activity of individual cells in motor cortex. *Journal of neurophysiology*, 77(2):826–852, 1997.

[35] Stephen H Scott, Paul L Gribble, Kirsten M Graham, and D William Cabel. Dissociation between hand motion and population vectors from neural activity in motor cortex. *Nature*, 413(6852):161–165, 2001.

[36] Nicholas G Hatsopoulos, Qingqing Xu, and Yali Amit. Encoding of movement fragments in the motor cortex. *Journal of Neuroscience*, 27(19):5105–5114, 2007.

[37] Mark M Churchland and Krishna V Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of neurophysiology*, 97(6):4235–4257, 2007.

[38] Jörn Rickert, Alexa Riehle, Ad Aertsen, Stefan Rotter, and Martin P Nawrot. Dynamic encoding of movement direction in motor cortical neurons. *Journal of Neuroscience*, 29(44):13870–13882, 2009.

[39] Abdulraheem Nashef, Rea Mitelman, Ran Harel, Mati Joshua, and Yifat Prut. Area-specific thalamocortical synchronization underlies the transition from motor planning to execution. *Proceedings of the National Academy of Sciences*, 118(6), 2021.

[40] Doug Rubino, Kay A Robbins, and Nicholas G Hatsopoulos. Propagating waves mediate information transfer in the motor cortex. *Nature neuroscience*, 9(12):1549–1557, 2006.

[41] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[42] Sandro Romani and Misha Tsodyks. Continuous attractors with morphed/correlated maps. *PLoS Comput Biol*, 6(8):e1000869, 2010.

[43] Alexei Samsonovich and Bruce L McNaughton. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, 17(15):5900–5920, 1997.

[44] David Hansel and Haim Sompolinsky. 13 modeling feature selectivity in local cortical circuits, 1998.

[45] Misha Tsodyks and Terrence Sejnowski. Associative memory and hippocampal place cells. *International journal of neural systems*, 6:81–86, 1995.

[46] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[47] Alfonso Renart, Pengcheng Song, and Xiao-Jing Wang. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485, 2003.

[48] Vladimir Itskov, David Hansel, and Misha Tsodyks. Short-term facilitation may stabilize parametric working memory trace. *Frontiers in computational neuroscience*, 5:40, 2011.

[49] Alexander Seeholzer, Moritz Deger, and Wulfram Gerstner. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS computational biology*, 15(4):e1006928, 2019.

[50] Alexander V Lukashin and Apostolos P Georgopoulos. A dynamical neural network model for motor cortical activity during movement: population coding of movement trajectories. *Biological cybernetics*, 69(5):517–524, 1993.

[51] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87, 1977.

[52] David C Somers, Sacha B Nelson, and Mriganka Sur. An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience*, 15(8):5448–5465, 1995.

[53] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[54] A David Redish, Adam N Elga, and David S Touretzky. A coupled attractor model of the rodent head direction system. *Network: computation in neural systems*, 7(4):671, 1996.

[55] Bruce L McNaughton, Carol A Barnes, Jason L Gerrard, Katalin Gothard, Min W Jung, James J Knierim, H Kudrimoti, Y Qin, WE Skaggs, M Suster, et al. Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *The Journal of experimental biology*, 199(1):173–185, 1996.

[56] H Sebastian Seung, Daniel D Lee, Ben Y Reis, and David W Tank. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26(1):259–271, 2000.

[57] Misha Tsodyks. Attractor neural network models of spatial maps in hippocampus. *Hippocampus*, 9(4):481–489, 1999.

[58] Marcelo Camperi and Xiao-Jing Wang. A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. *Journal of computational neuroscience*, 5(4):383–405, 1998.

[59] Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923, 2000.

[60] SM Stringer, TP Trappenberg, ET Rolls, and IETd Araujo. Self-organizing continuous attractor networks and path integration: one-dimensional models of head direction cells. *Network: Computation in Neural Systems*, 13(2):217–242, 2002.

[61] Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[62] Francesco P Battaglia and Alessandro Treves. Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Physical Review E*, 58(6):7738, 1998.

[63] Rémi Monasson and Sophie Rosay. Transitions between spatial attractors in place-cell models. *Physical review letters*, 115(9):098101, 2015.

[64] M-C Hepp-Reymond, M Kirkpatrick-Tanner, L Gabernet, H-X Qi, and B Weber. Context-dependent force coding in motor and premotor cortical areas. *Experimental brain research*, 128(1-2):123–133, 1999.

[65] RB Muir and RN Lemon. Corticospinal neurons with a special role in precision grip. *Brain research*, 261(2):312–316, 1983.

[66] Rony Paz, Thomas Boraud, Chen Natan, Hagai Bergman, and Eilon Vaadia. Preparatory activity in motor cortex reflects learning of local visuomotor skills. *Nature neuroscience*, 6(8):882–890, 2003.

[67] Chiang-Shan Ray Li, Camillo Padoa-Schioppa, and Emilio Bizzi. Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron*, 30(2):593–607, 2001.

[68] SP Wise, SL Moody, KJ Blomstrom, and AR Mitz. Changes in motor cortical activity during visuomotor adaptation. *Experimental Brain Research*, 121(3):285–299, 1998.

[69] Robert Ajemian, Daniel Bullock, and Stephen Grossberg. Kinematic coordinates in which motor cortical cells encode movement direction. *Journal of Neurophysiology*, 84(5):2191–2203, 2000.

[70] Paul L Gribble and Stephen H Scott. Overlap of internal models in motor cortex for mechanical loads during reaching. *Nature*, 417(6892):938–941, 2002.

[71] Emanuel Todorov. Direct cortical control of muscle activation in voluntary arm movements: a model. *Nature neuroscience*, 3(4):391–398, 2000.

[72] RN Holdefer and LE Miller. Primary motor cortical neurons encode functional muscle synergies. *Experimental Brain Research*, 146(2):233–243, 2002.

[73] Lauren E Sergio, Catherine Hamel-Pâquet, and John F Kalaska. Motor cortex neural correlates of output kinematics and kinetics during isometric-force and arm-reaching tasks. *Journal of neurophysiology*, 94(4):2353–2378, 2005.

[74] Stephen H Scott. The role of primary motor cortex in goal-directed movements: insights from neurophysiological studies on non-human primates. *Current opinion in neurobiology*, 13(6):671–677, 2003.

[75] Jonathan A Michaels, Benjamin Dann, and Hansjörg Scherberger. Neural population dynamics during reaching are better explained by a dynamical system than representational tuning. *PLoS computational biology*, 12(11):e1005175, 2016.

[76] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966, 2018.

[77] Karen E Schroeder, Sean M Perkins, Qi Wang, and Mark M Churchland. Cortical control of virtual self-motion using task-specific subspaces. *bioRxiv*, pages 2019–12, 2021.

[78] Yoram Ben-Shaul, Rotem Drori, Itay Asher, Eran Stark, Zoltan Nadasdy, and Moshe Abeles. Neuronal activity in motor cortical areas reflects the sequential context of movement. *Journal of Neurophysiology*, 91(4):1748–1762, 2004.

[79] Xiaofeng Lu and James Ashe. Anticipatory activity in primary motor cortex codes memorized movement sequences. *Neuron*, 45(6):967–973, 2005.

Methods

Analysis of the model

We studied the rate model with couplings defined by (1) with two complementary approaches. First, an analytic approximation based on mean-field arguments and valid in the limit of large network size allowed us to derive a low-dimensional description of the network dynamics in terms of a few latent variables, and to fit the model parameters to the data. Next, we checked that simulations of the dynamics of a network of $N = 10^4$ neurons reproduce the results that we derived in the limit $N \to \infty$.

The mean-field equations are derived following the methods introduced in [41, 42]. In the limit where the number of neurons is large, the average activity of a neuron with coordinates

$\theta_A, \theta_B, \eta_A, \eta_B$ is described by equations (2) and (3). In the following, we will denote the integration measure by the shorthand

$$d\mu = d\theta_A\, d\theta_B\, d\eta_A\, d\eta_B\; \rho_d(\theta_A, \theta_B)\; \rho_{pA}(\eta_A)\; \rho_{pB}(\eta_B).$$

In order to derive a lower dimensional description of the dynamics, we rewrite (2) in terms of the average activity rate

$$r_0 = \int d\mu\; r(\theta_A, \theta_B, \eta_A, \eta_B), \tag{10}$$

and of the second Fourier components of the activity rate modulated by $\eta_{A/B}$:

$$\begin{aligned} Z_A &\equiv \int d\mu\; \eta_A\, r(\theta_A, \theta_B, \eta_A, \eta_B)\, e^{i\theta_A} \equiv r_A e^{i\psi_A} \\ Z_B &\equiv \int d\mu\; \eta_B\, r(\theta_A, \theta_B, \eta_A, \eta_B)\, e^{i\theta_B} \equiv r_B e^{i\psi_B}. \end{aligned} \tag{11}$$

The phase $\psi_{A(/B)}$ is defined so that the parameter $r_{A(/B)}$ is a real nonnegative number. Together with (10), the order parameters of the dynamics are thus defined by:

$$\begin{aligned} r_A &= \int d\mu\; \eta_A\, \cos(\theta_A - \psi_A)\, r(\theta_A, \theta_B, \eta_A, \eta_B) \\ r_B &= \int d\mu\; \eta_B\, \cos(\theta_B - \psi_B)\, r(\theta_A, \theta_B, \eta_A, \eta_B) \\ 0 &= \int d\mu\; \eta_A\, \sin(\theta_A - \psi_A)\, r(\theta_A, \theta_B, \eta_A, \eta_B) \\ 0 &= \int d\mu\; \eta_B\, \sin(\theta_B - \psi_B)\, r(\theta_A, \theta_B, \eta_A, \eta_B); \end{aligned} \tag{12}$$

$r_{A(/B)}$ is interpreted as a measure of the spatial modulation of the activity profile, while $\psi_{A(/B)}$ represents the position of the peak of the activity profile in map $A(/B)$. From (2), we see that the order parameters evolve in time according to the following set of equations,

$$\begin{aligned} \tau \frac{d}{dt} r_0(t) &= -r_0(t) + \int d\mu\; [I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t)]_+ \\ \tau \frac{d}{dt} r_A(t) &= -r_A(t) + \int d\mu\; \eta_A\, \cos(\theta_A - \psi_A(t))\, [I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t)]_+ \\ \tau \frac{d}{dt} r_B(t) &= -r_B(t) + \int d\mu\; \eta_B\, \cos(\theta_B - \psi_B(t))\, [I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t)]_+ \\ \tau r_A(t) \frac{d}{dt} \psi_A(t) &= \int d\mu\; \eta_A\, \sin(\theta_A - \psi_A(t))\, [I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t)]_+ \\ \tau r_B(t) \frac{d}{dt} \psi_B(t) &= \int d\mu\; \eta_B\, \sin(\theta_B - \psi_B(t))\, [I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t)]_+, \end{aligned} \tag{13}$$

where the total input (3) is rewritten as:

$$\begin{aligned} I^{\rm tot}(\theta_A, \theta_B, \eta_A, \eta_B; t) =& I^{\rm ext}(\theta_A, \theta_B, \eta_A, \eta_B; t) + j_0 r_0(t) + \\ & \sum_{\nu=A,B} j_s^\nu \eta_\nu \cos\left(\theta_\nu - \psi_\nu(t)\right) r_\nu + j_a \eta_B \cos\left(\theta_B - \psi_A(t)\right) r_A(t). \end{aligned} \tag{14}$$

.

Stationary states for homogeneous external inputs

We first study the properties of the fixed point solution focusing on the scenario where the joint distribution (7) of the θ_A, θ_B is of the form

$$\rho_d(\theta_A, \theta_B; x) = \frac{1}{4\pi^2} \left[1 + x \cos(\theta_A - \theta_B)\right], \tag{15}$$

which fits the empirical distribution of the data well for $x = 2/3$ (Fig. S1). For now, we leave the distributions $\rho_{pA}(\eta_A)$ and $\rho_{pB}(\eta_B)$ unspecified. We first consider the case where the external input to the network is a constant that is independent of θ_A, θ_B:

$$I^{\text{ext}}(\eta_A, \eta_B) = C_0 + \eta_A C_A + \eta_B C_B. \tag{16}$$

The stationary solutions of (2,14) are of the form:

$$r(\theta_A, \theta_B, \eta_A, \eta_B) = \left[I^{\text{tot}}\right]_+ = \left[I_0 + I_A \cos(\theta_A - \psi_A) + I_B \cos(\theta_B - \psi_B)\right]_+, \tag{17}$$

where we have defined the fields

$$\begin{aligned} I_0 =& C_0 + \eta_A C_A + \eta_B C_B + j_0 r_0 \\ I_A =& j_s^A \eta_A r_A \\ I_B =& j_s^B \eta_B r_B + j_a \eta_B r_A. \end{aligned} \tag{18}$$

Here, $\{r_0, r_A, r_B\}$ are solutions of the system (13) with the left hand side set to zero. The second term on the r.h.s in the last equation of (18) is obtained from (14) by observing that in the stationary state $\psi_A = \psi_B$ if either $j_a > 0$ or if θ_A and θ_B are correlated (as we will assume in the following). As in [41, 44], we can distinguish broad from narrow activity profiles. The term *broad activity profile* refers to the scenario where the activity of all of the neurons is above threshold, the dynamics is linear and the stationary state reduces to:

$$r(\theta_A, \theta_B, \eta_A, \eta_B) = I_0 + I_A \cos(\theta_A - \psi_A) + I_B \cos(\theta_B - \psi_B). \tag{19}$$

By inserting the above equation in (12), we find that the only solution is homogeneous over the maps θ_A and θ_B:

$$\begin{aligned} r_0 &= \frac{C_0 + \langle \eta_A \rangle C_A + \langle \eta_B \rangle C_B}{1 - j_0}, \\ r_A &= 0, \\ r_B &= 0, \\ r(\eta_A, \eta_B) &= C_0 + \eta_A C_A + \eta_B C_B + j_0 r_0, \end{aligned} \tag{20}$$

where the notation $\langle . \rangle$ represents an average over the measure $d\mu$, i.e.

$$\langle \eta_A \rangle = \int_0^1 d\eta\, \rho_{pA}(\eta)\, \eta, \quad \langle \eta_B \rangle = \int_0^1 d\eta\, \rho_{pB}(\eta)\, \eta.$$

First, we notice that a nonzero homogeneous state is present if

$$\frac{C_0 + \langle \eta_A \rangle C_A + \langle \eta_B \rangle C_B}{1 - j_0} > 0.$$

Then, the stability of this state with respect to a small perturbation $\{\delta r_0, \delta r_A, \delta r_B\}$ can be studied by linearizing (13) around the stationary solution, at fixed $\psi_A = \psi_B = 0$. The resulting Jacobian matrix is

$$\begin{pmatrix} -1 + j_0 & 0 & 0 \\ 0 & -1 + j_s^A F_{AA} + j_a F_{AB} & j_s^B F_{AB} \\ 0 & j_s^A F_{AB} + j_a F_{BB} & -1 + j_s^B F_{BB} \end{pmatrix}$$

where

$$\begin{aligned} F_{AA} &= \int d\mu\, \eta_A^2 \cos(\theta_A)^2 = \frac{1}{2}\langle \eta_A^2 \rangle \\ F_{BB} &= \int d\mu\, \eta_B^2 \cos(\theta_B)^2 = \frac{1}{2}\langle \eta_B^2 \rangle \\ F_{AB} &= \int d\mu\, \eta_A \eta_B \cos(\theta_A)\cos(\theta_B) = \frac{x}{4}\langle \eta_A \rangle \langle \eta_B \rangle. \end{aligned}$$

The homogeneous solution (20) is stable if $j_0 < 1$ and if the couplings parameters $\{j_s^A, j_s^B, j_a\}$ satisfy the system of inequalities (4). If $j_0 \geq 1$, the system undergoes an amplitude instability. For values of $\{j_s^A, j_s^B, j_a\}$ that exceed the threshold implicitly defined by (4), the dynamics is no longer linear and the activity profile at the fixed point is narrowly localized and characterized by positive stationary values of the order parameters r_A, r_B.

Tuned and time-dependent external inputs

We next considered the case where the system is subject to a tuned time-dependent external input. While the external inputs have time-varying magnitude, their location Φ is constant and is the same in the two maps: $\Phi_A = \Phi_B \equiv \Phi$. Analogously to the ring model [44], the external input pins the location of the bump of activity. We assume that at the initial time the location of the bump equals the location of the external input, so that $\psi_A(t) = \psi_B(t) = \Phi$ at all times t. Although the position of the bump is stable, the direction modulation of the activity rates change in time in response to time-varying external inputs. We parameterized the external input in analogy with the recurrent input:

$$\begin{aligned} I^{\text{ext}}(\theta_A, \theta_B, \eta_A, \eta_B; t) =& C_0(t) + \eta_A C_A(t) + \eta_B C_B(t) + \\ & \eta_A \cos(\theta_A - \Phi)\epsilon_A(t) + \eta_B \cos(\theta_B - \Phi)\epsilon_B(t), \end{aligned} \tag{21}$$

where $C_0, C_A, C_B, \epsilon_A, \epsilon_B$ represent, respectively, the magnitude of: an homogeneous input; an input that is proportional to η_A ($\backslash\eta_B$) but homogeneous in θ_A ($\backslash\theta_B$); an input that is proportional to η_A ($\backslash\eta_B$) and tuned to θ_A ($\backslash\theta_B$). The total inputs can be rewritten as in (5).

Fitting the model to the data

Neurons activity rates were computed by smoothing the spike trains with a Gaussian kernel with s.d. of 25ms and averaging them across all trials with the same condition; $r_k^{(i)}(t)$ denotes the rate of neuron i at time t for condition k, each condition corresponding to one of the 8 angular locations of the target on the screen. Since trials had highly variable length, we normalized the responses along the temporal dimension before averaging them over trials, as follows. We divided the activity into three temporal intervals: from the target onset to the go cue; from the go cue to the start of the movement; from the start of the movement to the end of the movement. For each interval, we normalized the response times to the average length of the interval across trials. We then aggregated the three intervals together. We defined the preparatory and execution epochs – denoted by A and B – as two 300ms time intervals beginning, respectively, 100ms after target onset and 50ms before the start of the movement, in line with [21]. Fig. S2 shows that our results do not change qualitatively when the lengths of the preparatory and execution intervals are increased. For each neuron i, we fitted the activity rate averaged across time within each epoch as a function of the angular position Φ of the target with a cosine function:

$$a_\nu^{(i)} + b_\nu^{(i)} \cos(\theta_\nu^{(i)} - \Phi), \quad \nu = A, B,$$

where the parameters $\theta_A^{(i)}$ and $\theta_B^{(i)}$ represent the neuron's preferred direction during the preparatory and execution epochs. In our the model, the direction modulation of the rates, see (6), is proportional to $\eta_{A\backslash B}$, that measures how strongly the neuron participates in the two epochs of movement; hence, we defined $\eta_{A\backslash B}^{(i)}$ to be proportional to the amplitude of the tuning curve:

$$\eta_\nu^{(i)} = b_\nu^{(i)} / \max_i(b_\nu^{(i)}), \quad \nu = A, B. \tag{22}$$

The scatter plot of η_A, η_B (Fig. 2.d) shows an outlier with $\eta_B \sim 1$. We checked that our results hold true if we discard that point. Fig. S1 shows that η_A and η_B are not significantly

correlated with either θ_A, θ_B, nor $\theta_A - \theta_B$, while Fig. 2 shows that η_A and η_B have significant but weak mutual correlation. Based on these observations, we assumed for simplicity that η_A and η_B are independent variables. The order parameters (10,12) are computed at each time t by approximating the integrals with the sums:

$$\begin{aligned} r_0^{\text{data}}(t) &= \frac{1}{N}\frac{1}{n_c}\sum_i \sum_k r_k^{(i)}(t), \\ r_{A/B}^{\text{data}}(t) &= \frac{1}{N}\frac{1}{n_c}\sum_i \sum_k \eta_{A/B}^{(i)} \cos\left(\theta_{A/B}^{(i)} - \psi_{A/B}^{(k)}(t)\right) r_k^{(i)}(t) \quad \text{for} \quad k = 1, \ldots, n_c, \end{aligned} \tag{23}$$

where N is the number of neurons and $n_c = 8$ is the number of conditions. The angular location of the localized activity $\psi_{A/B}^{(k)}(t)$ can be computed from (12) as

$$0 = \frac{1}{N}\sum_i \eta_{A/B}^{(i)} \sin\left(\theta_{A/B}^{(i)} - \psi_{A/B}^{(k)}(t)\right) r_k^{(i)}(t). \tag{24}$$

However, this estimate is strongly affected by the heterogeneity in the distribution of θ_A, θ_B - deviating from the rotational symmetry of the model. That is why in Fig. 3 we approximated $\psi_{A/B}^{(k)}(t)$ by the $k-th$ angular location of the target on the screen. Fig. S2 shows that computing $\psi_{A/B}^{(k)}(t)$ by using either method does not affect the dynamics of the order parameters $r_{A/B}^{\text{data}}(t)$.

We assumed that the observed dynamics of the order parameters $r_0(r), r_A(t), r_B(t)$ obeys equations (13,14) with time-dependent external inputs of the form (21). We inferred the value of the parameters $\{C_0(t), C_A(t), C_B(t), \epsilon_A(t), \epsilon_B(t)\}_t$ of the external fields and j_0, j_s^A, j_s^B, j_a of the coupling matrix that allow us to reconstruct the dynamics of the order parameters r_0, r_A, r_B computed from data; since this inference problem is undetermined, we required as further constraint that the model reconstruct the dynamics of the following two additional order parameters:

$$\begin{aligned} \tau \frac{d}{dt} r_{0A}(t) &= -r_A(t) + \int d\mu\, \eta_A \left[I^{\text{tot}}(\theta_A, \theta_B, \eta_A, \eta_B; t)\right]_+ \\ \tau \frac{d}{dt} r_{0B}(t) &= -r_B(t) + \int d\mu\, \eta_B \left[I^{\text{tot}}(\theta_A, \theta_B, \eta_A, \eta_B; t)\right]_+. \end{aligned} \tag{25}$$

In this way, for given coupling parameters j_0, j_s^A, j_s^B, j_a, we can uniquely identify the external fields parameters that produced the observed dynamics. Still, an equally good reconstruction of $r_0, r_A, r_B, r_{0A}, r_{0B}$ can be obtained for different choices of coupling parameters (3). Hence, we inferred the model parameters by minimizing a cost function composed of two terms: one that is proportional to the reconstruction error of the temporal evolution of the order parameters and the other that represents an energetic cost penalizing large external inputs.

Fitting the model to the data: details.

The fitting procedure was divided in the following steps:

1. The time interval T going from the target onset till the end of the movement was binned into $\Delta t = 5ms$ time bins: $T = \{\Delta t_1, \Delta t_2, \ldots \Delta t_T\}$.

2. The couplings parameters were initialized to zero: $j_0 = 0, j_s^A = 0, j_s^B = 0, j_a = 0$. At the first time bin, the external fields parameters were initialized to zero:

$$C_0(\Delta t_1) = C_A(\Delta t_1) = C_B(\Delta t_1) = \epsilon_A(\Delta t_1) = \epsilon_B(\Delta t_1) = 0$$

and the reconstructed order parameters $(r_0, \dots)$ were initialized to the order parameters estimated from the data $(r_0^{\text{data}}, \dots)$:

$$r_0(\Delta t_1) = r_0^{\text{data}}(\Delta t_1), \quad r_A(\Delta t_1) = r_A^{\text{data}}(\Delta t_1), \quad r_B(\Delta t_1) = r_B^{\text{data}}(\Delta t_1),$$
$$r_{0A}(\Delta t_1) = r_{0A}^{\text{data}}(\Delta t_1), \quad r_{0B}(\Delta t_1) = r_{0B}^{\text{data}}(\Delta t_1).$$

3. For each time step $\Delta t_i,\ i = 2, \dots, T$:

   - We started from the reconstructed order parameters at the previous time step:

     $$r_0(\Delta t_{i-1}), r_A(\Delta t_{i-1}), r_B(\Delta t_{i-1}), r_{0A}(\Delta t_{i-1}), r_{0A}(\Delta t_{i-1}),$$

     and we let the dynamical system (13, 25) with external fields parameters $C_0, C_A, C_B, \epsilon_A, \epsilon_B$ evolve for $\Delta t = 5ms$ to estimate the order parameters at the current time step:

     $$r_0(\Delta t_i), r_A(\Delta t_i), r_B(\Delta t_i), r_{0A}(\Delta t_i), r_{0A}(\Delta t_i).$$

   - We inferred the value of the external fields parameters

     $$C_0(\Delta t_i), C_A(\Delta t_i), C_B(\Delta t_i), \epsilon_A(\Delta t_i), \epsilon_B(\Delta t_i)$$

     by minimizing the reconstruction error:

     $$E_{\Delta t}(C_0, C_A, C_B, \epsilon_A, \epsilon_B) = \frac{[r_0(\Delta t) - r_0^{\text{data}}(\Delta t)]^2}{\frac{1}{T}\sum_i r_0^{\text{data}}(\Delta t_i)} + \frac{[r_A(\Delta) - r_A^{\text{data}}(\Delta t)]^2}{\frac{1}{T}\sum_i r_A^{\text{data}}(\Delta t_i)}$$
     $$+ \frac{[r_B(\Delta t) - r_B^{\text{data}}(\Delta t)]^2}{\frac{1}{T}\sum_i r_B^{\text{data}}(\Delta t_i)} + \frac{[r_{0A}(\Delta t) - r_{0A}^{\text{data}}(\Delta t)]^2}{\frac{1}{T}\sum_i r_{0A}^{\text{data}}(\Delta t_i)} + \frac{[r_{0B}(\Delta t) - r_{0B}^{\text{data}}(\Delta t)]^2}{\frac{1}{T}\sum_i r_{0B}^{\text{data}}(\Delta t_i)}, \tag{26}$$

     that quantifies the difference between the order parameters estimated from the data and the reconstructed ones; note that the dependence of the cost function E on $C_0, C_A, C_B, \epsilon_A, \epsilon_B$ is implicitly contained in the reconstructed order parameters. We minimized the cost function (26) by using an interior point method algorithm [Byrd et al, 1999] starting from the initial condition

     $$C_0(\Delta t_{i-1}), C_A(\Delta t_{i-1}), C_B(\Delta t_{i-1}), \epsilon_A(\Delta t_{i-1}), \epsilon_B(\Delta t_{i-1});$$

     we imposed that $\epsilon_A > 0, \epsilon_B > 0$ and added a $L1$ regularization term to stabilize the solution.

   The external fields inferred with step 3 depend on our initial choice of of the couplings parameters j_0, j_s^A, j_s^B, j_a.

4. Using step 3, the value of the couplings parameters j_0, j_s^A, j_s^B, j_a is inferred by minimizing the cost function

   $$E_{\text{tot}} = \alpha E_{\text{ rec}} + E_{\text{ ext}}$$

   composed of two terms: the reconstruction error and a term that favors small external fields:

   $$\begin{aligned} E_{\text{rec}} &= \sum_{i=2}^{T} \left\{ E_{\Delta t_i} \left[C_0(\Delta t_i), C_A(\Delta t_i), C_B(\Delta t_i), \epsilon_A(\Delta t_i), \epsilon_B(\Delta t_i)) \right] \right\}, \\ E_{\text{ext}} &= a \sum_{i=2}^{T} \left[|C_0(\Delta t_i)| + |C_A(\Delta t_i)| + |C_B(\Delta t_i)| + \epsilon_A(\Delta t_i) + \epsilon_B(\Delta t_i) \right]. \end{aligned} \tag{27}$$

   The hyperparameter α was chosen so that the two terms have equal magnitude after minimization (see Fig. S4 for details) and $a = 0.001$ is a parameter rescaling the external fields. The minimization is done using a surrogate optimization algorithm [Gutmann et al., 2001; Wang et al., 2014].

The result does not depend on the choice of the time bin Δt. Also, the result weakly depends on the time constant τ in the mean field equations (e.g., 13), if τ varies on in the range: $10 - 100$ms. We set $\tau = 25$ms.

Simulations of the model

We simulated the dynamics of a finite network of N neurons. To each neuron i, we assigned the variables $\theta_A^{(i)}, \theta_B^{(i)}, \eta_A^{(i)}, \eta_B^{(i)}$ as follows.

- In the $\theta_{A\backslash B}$-space, we sampled N_θ points $\{\theta_A^{(i)}, \theta_B^{(i)}\}_{i=1}^{N_\theta}$ equally spaced along the lines

$$\theta_A - \theta_B = \text{cost}$$

  in such a way that their joint distribution matches the distribution of the data (8), as shown in Fig.S5.

- In the $\eta_{A\backslash B}$-space, we drew N_η points $\{\eta_A^{(i)}, \eta_B^{(i)}\}_{i=1}^{N_\eta}$ at random from the empirical distribution $\rho_{pA}(\eta_A)\rho_{pB}(\eta_B)$.

- For $i = 1, 2, \ldots N_\theta$, we assigned to a block of N_η neurons the same coordinates $\theta_A^{(i)}, \theta_B^{(i)}$ in the $\theta_{A\backslash B}$-space, and all possible coordinates $\{\eta_A^{(j)}, \eta_B^{(j)}\}_{j=1}^{N_\eta}$ in the $\eta_{A\backslash B}$-space, so that the overall number of neurons is $N = N_\theta N_\eta$.

The network dynamics we simulated is defined by the following stochastic differential equation:

$$\tau \frac{dr_i(t)}{dt} = -r_i(t) + \left[\frac{1}{N} \sum_{j=1}^{N} J_{ij} r_j(t) + I_i^{\text{ext}}(t) + \xi_i(t) \right]_+ , \tag{28}$$

$$d\xi_i(t) = -\gamma \xi_i(t) dt + \sigma_n dW, \tag{29}$$

where

$$J_{ij} = j_0 + \sum_{\nu=A,B} j_s^\nu \eta_\nu^{(i)} \eta_\nu^{(j)} \cos(\theta_\nu^{(i)} - \theta_\nu^{(j)}) + j_a \eta_B^{(i)} \eta_A^{(j)} \cos(\theta_B^{(i)} - \theta_A^{(j)}), \tag{30}$$

$$I_i^{\text{ext}} = C_0(t) + \eta_A^{(i)} C_A(t) + \eta_A^{(i)} \cos(\theta_A^{(i)} - \Phi)\epsilon_A(t) \tag{31}$$

$$+ \eta_B^{(i)} C_B(t) + \eta_B^{(i)} \cos(\theta_B^{(i)} - \Phi)\epsilon_B(t); \tag{32}$$

. W in (29) is a Wiener process, and the parameters $\gamma = 75/s, \sigma_n = 0.35Hz/\sqrt{s}$ set the magnitude of the noise fluctuations. The level of noise is chosen so that the results of the PCA analysis (see next section) match the data. Note that by setting the noise to zero and taking the limit $N \to \infty$ we recover the mean-field equations (2). The results of Fig. 4 are obtained from a network of $N = 16000$ neurons; we simulate the network dynamics for 8 location of the external input Φ and 20 trials for each of the 8 conditions, i.e. 20 instances of the noisy dynamics. The order parameters shown in Fig. 4 are computed from single trial activity, and then averaged over trials. The correlation-based analysis, instead, is obtained from trial-averaged activity.

Correlations and PCA analysis.

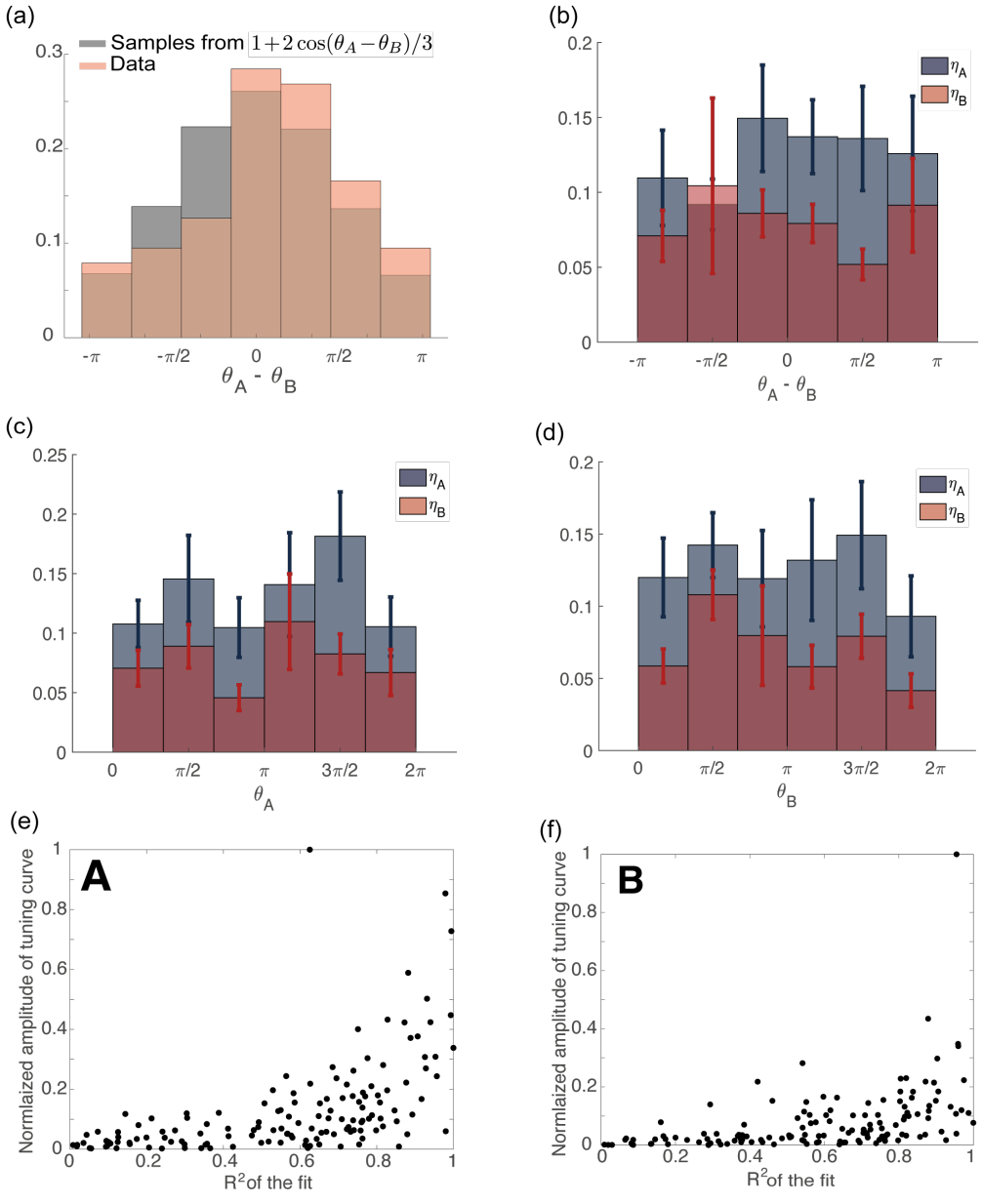
The correlation-based analysis explained in this section was performed both on the smoothed and trial-averaged spike trains from recordings, and on the trial-averaged activity rates from simulations. To compute signal correlations, we preprocessed the data as follows - the same procedure holds both for the recordings and for the simulations. For each neuron, we normalized the activity by its standard deviation (computed across all times and all

conditions); then, we mean-centered the activity across conditions. The $T = 300$ms long preparatory activity for all $C = 8$ conditions was concatenated into a $N \times TC$ matrix denoted by P, and the movement-related activity was grouped into an analogous matrix M. We obtained correlation matrices relative to preparatory and movement related activity by computing the correlations between the rows of the respective matrices. We then identified the prep-PCs and move-PCs by performing PCA separately on the matrices P and M. The degree of orthogonality between the prep- and move- subspaces was quantified by the Alignment Index A [21], measuring the amount of variance of the preparatory activity explained by the first K move-PCs:

$$A = \frac{\mathrm{Tr}(E_{\mathrm{mov}}^T C_{\mathrm{prep}} E_{\mathrm{mov}})}{\sum_{i=1}^{K} \sigma_{\mathrm{prep}}(i)},$$

where E_{mov} is the matrix defined by the top K move-PCs, C_{prep} is the covariance matrix of the preparatory activity and $\sigma_{\mathrm{prep}}(i)$ is the i-th eigenvalue of C_{prep}. K was set to the number of principal components needed to explain 88% of the execution activity variance. Hence, the Alignment Index ranges from 0 (orthogonal subspaces) to 1 (aligned subspaces). As random test, we computed the Random Alignment Index between two sets of K dimensions drawn at random within the space occupied by neural activity, using the Monte Carlo procedure described in [21]. We performed the same analysis on both the data ($K = 12$) and the model ($K = 9$) trial averaged activity. Since the number of recorded neurons was of the order of $N_{\mathrm{data}} \sim 10^2$, while the simulated network was composed of $N_{\mathrm{sim}} \sim 10^4$ neurons, we computed signal correlations and performed the PCA analysis on a subset of N_{data} neurons randomly sampled from the larger simulated network. We repeated this procedure 100 times and averaged the Alignment Index and the Random Alignment Index across the 100 random samples.

We also quantified the roational structure present in the data by applying the jPCA [26] dimensionality reduction technique to both the simulated activity and the recordings, and showed (Fig. S8) that the trajectories from simulations qualitatively resembled the ones from data when projected onto the dimensions that capture rotational dynamics.



h!

Figure S1: **(a)** Distribution of $\theta_A - \theta_B$ computed from recorded data (orange) and data generated from the distribution: $1+2\cos(\theta_A-\theta_B)/3$ (gray). **(b)** Variables η_A, η_B plotted as a function of $\theta_A-\theta_B$. The circular correlation is $r = 0.13$ ($P = 0.3$), for both η_A and η_B. **(c)** Variables η_A, η_B plotted as a function of θ_A. The circular correlation is $r = 0.08$ ($P = 0.6$) for η_A and $r = 0.06$ ($P = 0.7$) for η_B. **(d)** Variables η_A, η_B plotted as a function of θ_B. The circular correlation is $r = 0.01$ ($P = 0.9$) for η_A and $r = 0.1$ ($P = 0.2$) for η_B. **(e)** Scatter plot of the amplitude of a cosine function fitted to the tuning curve of the preparatory activity vs the R^2 coefficient of the fit. The neurons with larger tuning amplitude are the ones with higher R^2. **(f)** Same as in **(c)** but for movement-related activity.

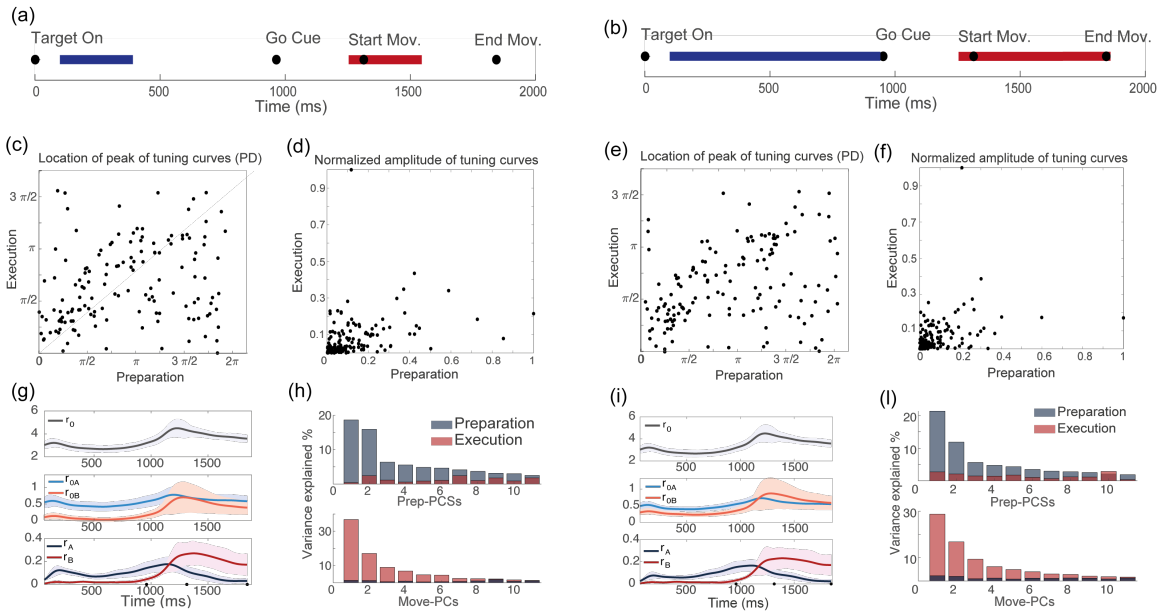


Figure S2: For two different definitions of the preparatory (blue segment) and movement-related (red segment) epochs: **(a)** vs **(b)**, we show the corresponding distributions of θ_A, θ_B: **(c)** vs **(e)**; the distributions of η_A, η_B: **(d)** vs **(f)**; the dynamics of the order parameters: **(g)** vs **(i)**; and the results of the PCA analysis: **(h)** vs **(l)**.

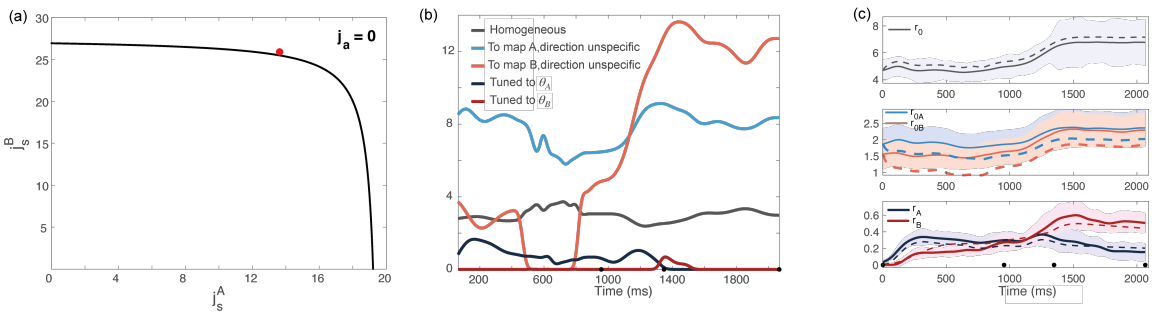


Figure S3: Analysis based on data recorded from monkey Rj. **(a)** Couplings parameters and **(b)** dynamics of the external inputs inferred from data. **(c)** Dynamics of the order parameters (solid line: data, shaded area: $\pm$ standard error across the population; dotted line: analytical prediction).

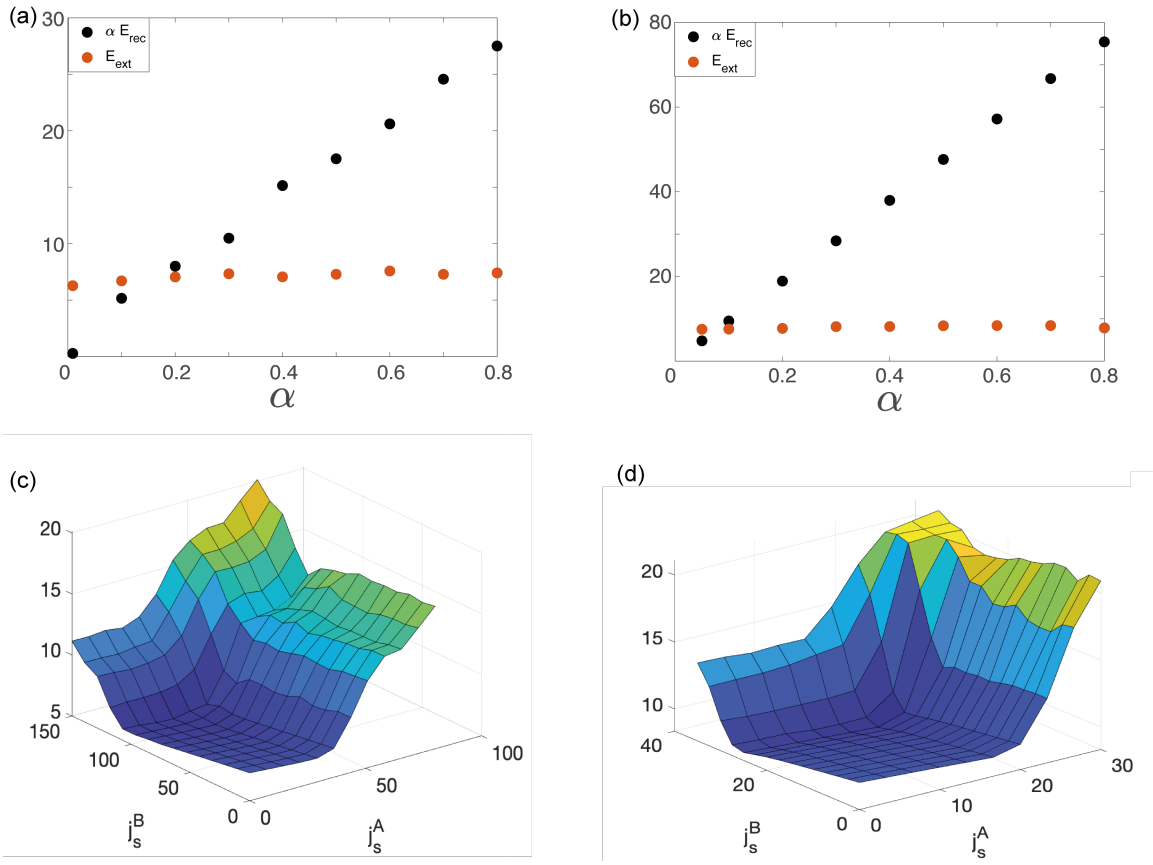


Figure S4: The two terms composing the cost function $E_{\text{tot}} = \alpha E_{\text{rec}} + E_{\text{ext}}$ (see) as a function of α, obtained from fitting the model to data recorded from monkey Rk **(a)** and Rj **(b)**. In both cases, the hyperparameter α is chosen so that the two terms composing the cost function have equal magnitude. **(c)** Cost function E_{tot} computed at different values of j_s^A and j_s^B. **(d)** Same as in (c), but for monkey Rj.

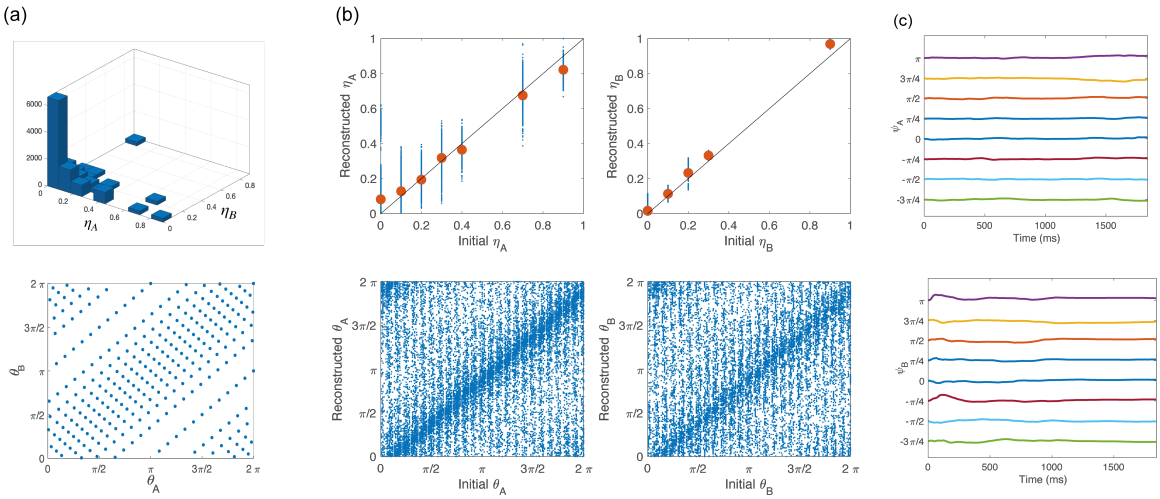


Figure S5: **(a)** To build the network architecture, we assigned to 16000 neurons coordinates η_A, η_B (top) and θ_A, θ_B (bottom) drawn from the their respective empirical distribution. **(b)** Scatter plot of the coordinates initially assigned to the neurons (x-axes) vs the corresponding variables that we computed from simulations of the network activity during the preparatory and movement-related epochs (y-axes). Points clustered on the diagonal show self-consistency of the model. **(c)** Dynamics of the order parameters ψ_A (top) and ψ_B (bottom) for 8 different locations Φ of the external input; ψ_A and ψ_B represent the location of the bump of activity in map A and B, respectively.

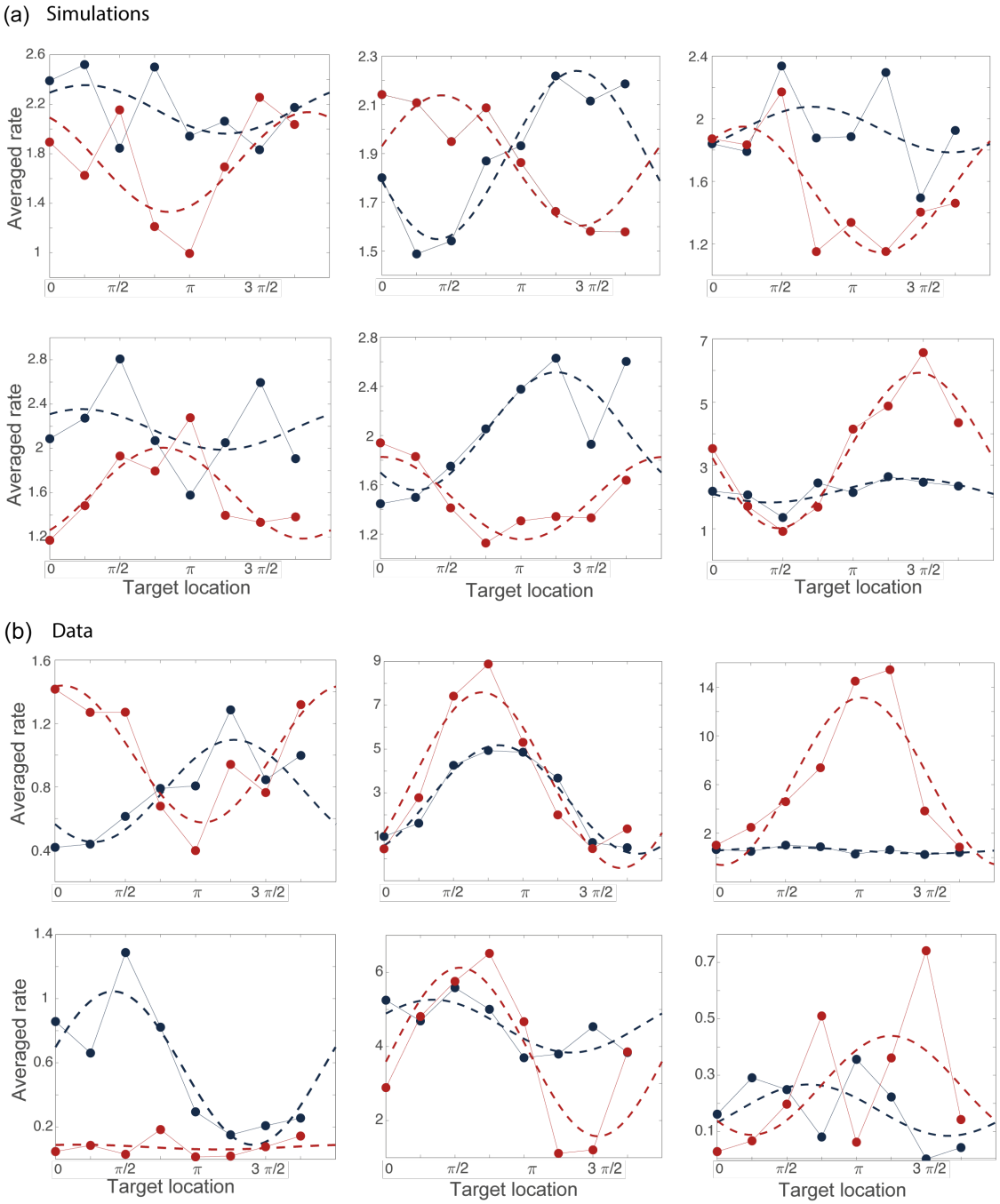


Figure S6: Examples of tuning curves during the preparatory (blue) and movement-related (red) epochs computed from simulations **(a)** and from data **(b)**.

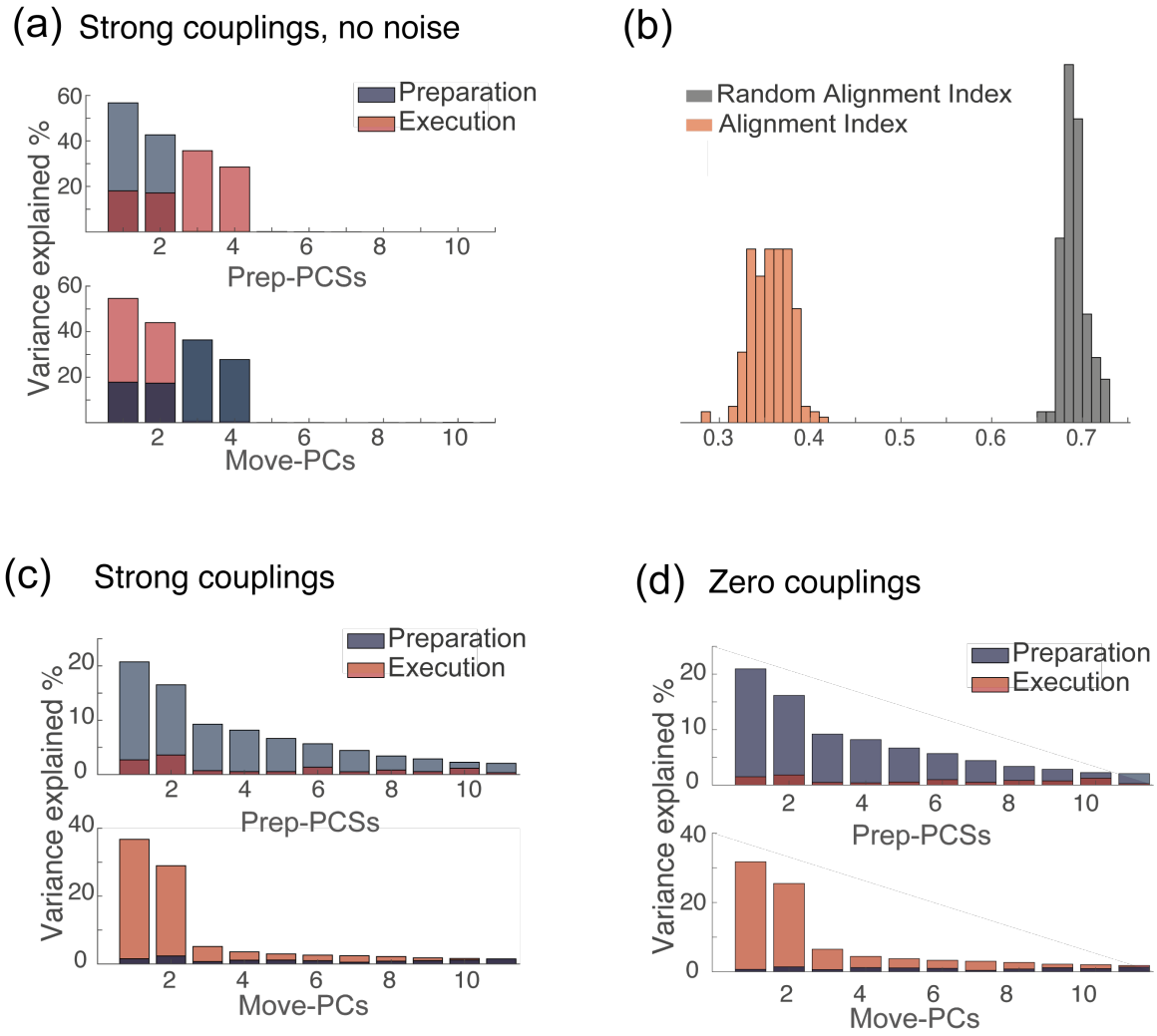


Figure S7: **(a)** PCA analysis on the results of simulations for the solution of Fig.7.g (strong couplings) when no noise is added to the dynamics. **(b)** Alignment index quantifying the degree of orthogonality between the preparatory and movement-related subspaces from simulations, for the same solution as in (a): strong couplings, no noise added to the dynamics. Orange: alignment index for 100 subsets of 140 neurons each, sampled uniformly at random from the larger network we simulated; gray: the average random alignment index for each subset. In absence of noise, the dynamics is confined onto a four-dimensional space; the random alignment index quantifies the alignment between two two-dimensional subspaces drawn at random within the four-dimensional space occupied by neural activity. Note that the alignment index is significantly smaller than the random test. **(c)** PCA analysis on the results of simulations for the solution of Fig.7.g (strong couplings) and **(d)** of Fig.7.a (zero couplings).

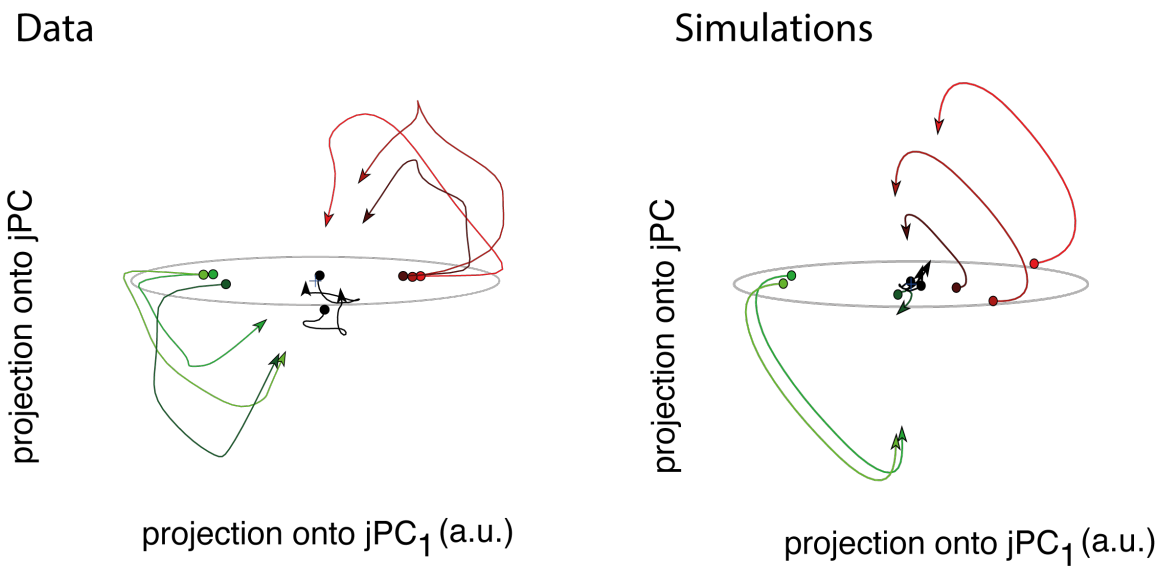


Figure S8: **(a)** Trial-averaged activity rates during movement execution (from $50ms$ before the start of the movement, to $250ms$ after the start of the movement) projected onto the first two jPCA dimensions, defined as in [26]. In order to compare the data to the theory, we averaged the activity of all trials corresponding to the same condition – defining 8 groups (each denoted in a different color) corresponding to the 8 different conditions. However, hand kinematics corresponding to the activity within the same group can differ quite substantially, even if they end up at the same target location. This is a major difference with respect to the analysis of [26], where many more groups were defined, each group containing the activity corresponding to very similar hand trajectories. **(b)** Same as in (a), but for activity from simulations.