Bifurcation in space: Emergence of functional modularity in the neocortex

Xiao-Jing Wang[1,†,*], Junjie Jiang[1,2,†,*], Roxana Zeraati[3], Aldo Battista[1], Julien Vezoli[4], Henry Kennedy[4], Ulises Pereira-Obilinovic[1,5]

[1] Center for Neural Science, New York University, 4 Washington Place, New York 10003, USA

[2] Present address: The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Institute of Health and Rehabilitation Science, School of Life Science and Technology, Research Center for Brain-inspired Intelligence, Xi'an Jiaotong University, No.28, West Xianning Road, Xi'an, 710049, Shaanxi, P. R. China.

[3] Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany

[4] INSERM, Stem Cell and Brain Research Institute U1208, 69500 Bron, France

[5] The Allen Institute for Neural Dynamics, Seattle, WA 98109, USA

† These authors contributed equally to this work.

*To whom correspondence should be addressed; E-mail:xjwang@nyu.edu,jiangjj@xjtu.edu.cn.

Abstract

Recent studies have shown that neural representation and processing are widely distributed in the brains of behaving animals [1, 2, 3, 4]. These observations challenge functional specialization as a central tenet of Neuroscience, which refers to the notion that distinct brain regions are dedicated to specific aspects of cognition such as working memory or subjective decision-making. Here we develop the concept of *bifurcation in space* to mechanistically account for the emergence of functional specialization that is compatible with distributed neural coding in a large-scale neocortex. Our theory starts with a departure from the canonical local circuit principle [5] by highlighting differences between cortical areas in the form of experimentally quantified heterogeneities of synaptic excitation and inhibition. We investigated connectome-based modelling of a multiregional cortex for both macaque monkeys and mice, as well as a generative model of a spatially embedded neocortex. During working memory in a simulated delayed response task, surprisingly, we found an inverted-V-shaped pattern of neuronal timescales across the cortical hierarchy as a signature of functional modularity, in sharp contrast to an increasing pattern of timescales during the resting state, as reported previously [6]. Furthermore, our model cortex simultaneously and robustly displays a plethora of bifurcations in space and their associated rich repertoire of timescale profiles across a large-scale cortex; the corresponding functionally defined modules (spatial attractors) could potentially subserve various internal mental processes. This work yields several specific experimentally testable predictions, including an inverted-V pattern of timescales, a measure of comparison between functional modules and structural modules defined by the graph theory, and a new plot for revealing bifurcation in space in neural activity recorded from animals performing different tasks that engage various functional modules. We propose that bifurcation in space, resulting from the connectome and macroscopic gradients of neurobiological properties across the cortex, represents a fundamental principle for understanding the brain's functional specialization and modular organization.

Introduction

Technical advances are enabling neuroscientists to image calcium signals or record the spiking activity of large populations of single cells in behaving animals, opening a new era for the investigation of distributed neural computation across the multi-regional brain [7]. Experimentally observed widespread neural signals of task-relevant information raised a major challenge to the concept of *functional modularity*, namely specialization of brain areas dedicated to distinct processes such as face recognition in vision, subjective decision-making, or complex cognitive processes like music perception [8, 9]. In the neocortex, connectomic analysis showed that 67% and 97% of all possible area-to-area projections are present in the macaque [10] and mouse [11, 12], respectively; therefore, neural signals could in principle be shared across the cortical mantle within one or two synapses. A mechanistic understanding is needed to explain how functional specialization/modularity may emerge in such a large-scale cortical circuit endowed with an abundance of long-range inter-areal connections.

In this work, we tackle the challenge using connectome-based computational modeling of the multi-regional neocortex and mathematical analysis. For concreteness, consider working memory, our brain's ability to maintain and manipulate information internally without external stimulation [13]. Working memory represents an excellent case study because it is a core cognitive function and has been extensively investigated in tasks where information must be held in the brain across a short time period in order to guide a later response. The underlying neural mechanism involves persistent neural firing that is self-sustained [14, 15, 16] and distributed over some cortical areas but not others [17, 18]. In particular, for working memory of visual motion information, there is evidence that the middle temporal (MT) area does not show persistent activity during a mnemonic

delay, while its monosynaptic projection target, the medial superior temporal (MST) area does [19], suggesting a sharp onset of working memory representation along the cortical hierarchy. What is the mechanism for the emergence of such functional modularity?

We have previously developed large-scale cortex models for distributed working memory in both monkey [20, 21] and mouse [22] on the basis of recently available area-to-area connectivity data. Critically, unlike the conventional view of a cortical network as a graph where parcellated areas (nodes) are identical and differ from each other only by their connections, our model incorporated heterogeneities in different parts of the cortex [23, 24], which are not random but display macroscopic gradients along low-dimensional axes such as the anatomically defined cortical hierarchy [25]. Model simulations revealed an abrupt transition at some stage of the cortical hierarchy that separates cortical areas exhibiting information-coding self-sustained persistent activity from those that do not. These results suggested that the sharp transition is "akin to a bifurcation in cortical space" [20, 7].

The mathematical term "bifurcation" denotes the sudden onset of a qualitatively novel behavior under a graded change of the properties of a dynamical system [26]. For instance, as the amplitude of an input current is gradually increased to above a threshold level, a single neuron displays a sudden transition from a steady state to repetitive firing of action potentials. The idea of bifurcation in space implies a bifurcation localized at a position in the cortical space, which would be highly non-trivial. First, our model is set up so that no isolated area can maintain persistent activity when disconnected. Consequently, depending on long-distance connection loops, the observed working memory representation must be a collective phenomenon. Second, while the connectivity of the cortical graph is dense [10], bifurcation occurs locally in space. Third, a bifurcation in space does not require fine-tuning of a parameter. However, this idea has remained purely speculative up to now.

The novelty of the present work, absent in the previous ones, is sixfold. First, we incorporate hitherto unpublished connectomic data which enables us to reproduce and explain mechanistically the key experiment on sharp emergence of working memory representation between MT and MST [19]. Second, we show functional modularity both for working memory and subjective decision-making in the macaque cortex, thereby demonstrating bifurcation in space beyond working memory for the first time. Third, as an observable signature of this novel phenomenon, we report that the time constants of neural firing fluctuations are maximal at a spatial location of sharp transition separating areas devoid of persistent activity and those that exhibit robust persistent activity. Thus, cortex displays an inverted-V shaped pattern of time constants of neural fluctuations during a mnemonic state, in striking contrast to our previous finding of a roughly monotonic increase of time constants along the cortical hierarchy during a resting state [6, 27, 28]. Fourth, using a generative model of spatially embedded cortex, we carry out analytical calculations of a normal form (according to the mathematical theory of bifurcations) that establishes rigorously the novel concept of bifurcation in space. Fifth, we explicitly address the question of how functional modules defined here by spatial attractors may or may not be related to structural modules defined by a community detection algorithm in graph theory. The proposed comparison measure is broadly useful in neuroscience. Sixth, we report a diversity of timescale patterns across the cortical hierarchy, which nevertheless show a rather common profile when plotted against a differential neural firing rate with respect to a baseline. This new metric is applicable to the analysis of neurophysiological data as a probe for bifurcation in space.

Remarkably, a single large-scale neocortex model (with fixed parameter values) displays simultaneously many bifurcations in space, characterizing specialized modules that could potentially subserve distinct internally driven brain functions. On the basis of these

findings, we propose that modularity can be defined as a selective subset of cortical areas, which are not necessarily spatially congruent, engaged in a distinct brain function. This theory is compatible with distributed neural representation and processing across multiple brain regions, but against neo-equipotentiality à la Lashley [29] with merely graded variations of a behavioral attribute's neural coding across the entire cortical mantle.

Results

A connectome-based monkey cortex model displays bifurcation in space during decision-making and working memory

We start with an extension of a macaque cortex model [20, 21] to 41 parcellated areas (Fig. 1a, Fig. S1). The model incorporates hitherto unpublished connectivity data on the area MST, specifically motivated by the physiological observations that neurons in MST, but not their main monosynaptic afferents from the neighboring MT area, display delay period activity in a working memory task [19]. In a simple version of our model, suitable for stimulus-selective decision-making and working memory, each area has two excitatory populations encoding different options A and B, and one inhibitory population [20]. Both long-range and local recurrent excitation strength follows a macroscopic gradient proportional to the hierarchical position of each area [6, 25].

We simulated the model for a task that depends on perceptual decision-making and working memory. Concretely, we focused on a classic perceptual decision task used in monkey experiments that requires a subjective judgment of a random-dot stimulus's net motion direction [30, 31]. From trial to trial, the task difficulty varies by changing the motion coherence, the fraction of dots moving coherently from 100% to 0% (when there is no correct answer), the subject has to decide whether the net motion direction is A

(e.g., left) or B (right) (two alternative-forced choice task). In the model, the motion coherence is implemented by the relative strength of external inputs into the two (A and B) selective excitatory neural populations in V1 [32, 33]. Therefore, neural signals in V1 encode the physical stimulation veridically, whereas a subset of other areas in the multi-regional model are expected to determine a categorical choice by winner-take-all between the two selective neural populations. The latter is subjective (e.g., when the motion coherence is zero) and opposite to the objective evidence in error trials. How does the subjective decision emerge in the large-scale cortical circuit model?

As in the monkey experiments [30], we simulated two versions of the direction discrimination task. In the reaction time (RT) version, a decision is read out when the accumulated information reaches a threshold (for instance, when the neural activity in the lateral intraparietal area (LIP) ramps up to a preset level [30]. A different decision readout from a multi-regional system is conceivable). In the fixed duration (FD) version, a stimulus is presented for two seconds, followed by a delay period when the choice must be held in working memory to guide a later behavioral response (Fig. 1b). This task thus engages decision-making and working memory. The psychometric functions of the model for the RT and FD tasks (Fig. 1c) are comparable to the observed monkey performance.

To quantify the complex neural dynamics during decision-making in space and time, we computed the normalized difference between the activity of two excitatory populations as $\Delta_i(t) = (r^A_{E,i}(t) - r^B_{E,i}(t))/(r^A_{E,i}(t) + r^B_{E,i}(t))$, in each parcellated area i of our model. In model simulations of the RT task (with evidence in favor of A), the RT is read out when $\Delta_i(t)$ in LIP reaches a (positive or negative) threshold (see Methods). The average RT as a function of the coherence is shown in Fig. 1d, with longer RTs in error (open circle) than correct (filled circle) trials, in accordance with experimentally observed behavior.

Fig. 1f and g (see also Movies 1, 2) show the spatiotemporal dynamics of $\Delta_i(t)$ in a

correct trial and an error trial, respectively, in model simulations with evidence in favor of A. At the stimulus onset (left panel), the input to neural population A is larger than to B, and $\Delta(t)$ is positive in V1 (red); the signal propagates to higher areas where strong recurrent dynamics leads to a categorical winner (A or B) (middle and right panels). This was quantified by the time-to-threshold that varies from area to area (Fig. 1e); areas TEpd, TEO, and LIP are the first to reach the threshold in the correct (Fig. 1f, Movie 1) and error (Fig. 1g, Movie 2) trials. Compared to the correct trial, in the error trial, there is a period (Fig. 1g, middle panel) when some areas, such as V1 and MT, represent the veridical evidence A (with positive $\Delta_i(t)$, red). In contrast, others signal the subjective choice B (with negative $\Delta_i(t)$, blue). This suggests a bifurcation in space that unfolds in time. Late in time, MT also reflects the subjective choice (Fig. 1g, right panel), reproducing the observed choice correlate in MT neurons of behaving monkeys [34]. Choice signals in MT have been suggested to result from top-down inputs [35], which naturally occur in our connectome-based multiregional cortex model.

The inverted-V shaped profile of time constants in macaque monkey and mouse

The FD task allowed us to examine working memory of a choice across a delay period in the absence of external stimulation. The model's performance in the FD task is similar to that of the RT task (Fig. 1c). The connectome-based model of macaque cortex (Fig. 2a) exhibits the coexistence of a resting state (a firing rate around 1.5 Hz) and stimulus-selective persistent activity states encoding working memory of A or B (Fig. 2b, Fig. 6a, Fig. S2a-e). The multiregional persistent activity is dynamically stable as a spatial attractor, which defines a working memory module with only some areas displaying persistent firing (Fig. 6a). Capturing macaque monkey physiological experimental observations [17],

association cortical areas but not early sensory areas are engaged in stimulus-selective persistent firing during working memory states (Fig. 2b). The area MST exhibits sustained activity throughout the delay period, whereas MT does not manifest such persistent activity during this period, as evidenced in Fig. 2b. Thus, our model successfully reproduces the sharp transition in the cortical hierarchy observed by Mendoza-Halliday et al. (2014) [19], and offers a general account for the emergence of functional specialization for working memory [17].

We systematically analyzed how model behavior depends on two parameters: J_s^{max} that denotes the largest strength of local excitatory-to-excitatory connections among cortical areas, and z, which controls the counterstream inhibitory bias [6, 7] (the smaller the z, the more top-down projections target inhibitory versus excitatory neurons in a recipient area). We found that, in the (J_s^{max}, z) plane (Fig. S3a), the parameter set that satisfies the requirements for decision-making tasks constitutes a subset of the parameter space that meets the criteria for the working memory tasks (Fig. S3b,c). Consequently, we selected the parameter set that is tailored to the decision-making task, which inherently satisfies the requirements for the working memory task as well. The "reference parameter" values of J_s^{max} and z were chosen so that the model's performance fits well with the monkey's psychometric function (Fig. 1c).

We examined the timescale pattern along the cortical hierarchy in both the resting state and the working memory state of our model. We calculated the autocorrelation function of stochastic neural activity, from which a timescale is extracted (see Methods). In the resting state prior to a brief stimulus, the time constants dominating neural fluctuations were roughly a monotonic function of the hierarchy (Fig. 2d, upper panel, see also Fig. S2e, left), in agreement with previous modeling [6] and empirical [27, 28, 36] findings. When the analysis was repeated for the mnemonic delay period, we found that

the timescale of stochastic neuronal activity showed a peak, with short time constants for areas both at high and low hierarchical positions. The peak is located around the sharp transition in the hierarchical space where the working memory module emerges, around Brodmann area F4 (a premotor cortex) for the reference parameter set (Fig. 2d, lower panel; see also Fig. S2e, right). These findings show that the timescale pattern varies depending on the internal state of a strongly recurrent neural system. For a working memory state, the time constant of neural fluctuations can become very long near the transition (see also the next Section), a phenomenon called critical slowing down [37, 38]. Note that the particular area that displays the maximal time constant of mnemonic firing fluctuations depends on the model's parameter values, but the phenomenon is robust against parameter changes.

We then asked if critical slowing down also occurs in the connectome-based large-scale model of the mouse cortex [22]. The model contains 43 cortical areas in the common coordinate framework v3 atlas [39] (Fig. 2e) with a quantified hierarchy. Synaptic inhibition mediated by parvalbumin-expressing interneurons [40] displays a decreasing macroscopic gradient along the hierarchy. The mouse model also exhibits the coexistence of resting and persistent activity states appropriate for working memory function (Fig. 2f, Fig. S2f). Once again, we performed an autocorrelation analysis of the resting state and mnemonic persistent activity during working memory delay. We found an overall hierarchy of time constants in the resting state (Fig. 2h, upper panel) and a qualitatively similar inverted-V shaped profile of time constants during working memory (Fig. 2h, lower panel, see also Fig. S2g, right). The inverted-V pattern of time constants in the mouse and macaque monkey cortical models offers a strong model prediction that is testable experimentally.

A generative model for the mammalian neocortex

To investigate the mathematics of bifurcation in space, we need to "zoom in" close to an abrupt transition point, similar to a phase transition like the melting of ice when the temperature is adjusted to precisely zero degrees Celsius, but here at a particular spatial location along the cortical hierarchy that separates areas engaged in a working memory representation from those that are not. However, a connectome-based model of the mouse or macaque monkey cortex has a relatively small number of areas; thus, the distance between any pair of areas along the hierarchy cannot be reduced as much as desired. To overcome this limitation, we used a generative model of a spatially embedded mammalian cortex [41], unlike a purely topographic graph.

The model (Fig. 3a) is both generative and random, thus can be used to produce multiple *realizations* of a mammalian neocortical network model with sample connectivity matrices (see Fig. 3b for one network realization) that share the same statistical distributions empirically observed in the inter-areal connectivity [42, 10, 41]. This network model has three advantages. First, conclusions drawn from this model can be applied to different mammalian species [10, 11, 12, 43]. Second, in this model, the number of brain areas can be arbitrarily large, enabling us to examine the bifurcation phenomenon close to a transition point. Third, our results are robust by studying the network dynamics over multiple network realizations, i.e., our results depend only on the connection statistics but not qualitatively on the specific network realizations.

We define the hierarchical distance using the diffusion map embedding method [44, 45] applied to our model. In the embedding space obtained through this nonlinear dimensionality reduction method, closer areas share more paths connecting them with stronger connections, while areas further apart share fewer paths and weaker connections (see Methods). Interestingly, the embedding of the generated connectivity conforms to a

low-dimensional hyperbolic shape (see Fig. 3c). To define a hierarchical distance, we arbitrarily select the cortical area at one of the tips of the hyperbolic shape as the start of a hierarchy. We found that for the hyperbolic distance (the distance defined along the hyperbolic shape), cortical areas display a smooth progression evenly distributed across hierarchical positions (see Fig. 3d blue trace), in contrast to the Euclidean distance where a significant fraction of cortical areas are concentrated around the hierarchy value 0.8 (see Fig. 3d brown trace). Therefore, we use the hyperbolic distance to define the hierarchical position. Strikingly, after remapping each brain area's hyperbolic hierarchical position into the ellipsoid's position (see Fig. 3a), we found that the hierarchical position increases along the major axis of the ellipsoid, similar to the hierarchy of the mammalian cortex along the rostro-caudal axis (Fig. S5a). The alignment of hierarchical positions with the ellipsoid's major axis likely reflects the underlying network connectivity and embedding geometry. This network structure drives the primary gradient of the diffusion map to align with the longest axis, corresponding to the rostro-caudal axis of the mammalian cortex.

Using the above defined hierarchy, we built a simplified yet biologically realistic model of the neocortex incorporating the macroscopic properties of the canonical circuit corresponding to the different hierarchy levels. Each brain area is modeled as a local canonical circuit of recurrently connected excitatory and inhibitory populations (Fig. 3e and Methods). Consistent with the macroscopic gradient of excitation observed in the cortex [25], the local and long-range excitatory weights are scaled by the hierarchical location (i.e., $J \propto h$, inset in Fig. 3f). When decoupled, each cortical area exhibits a low firing rate resting state when synaptic excitation level is below a threshold corresponding to $J_{\text{threshold}} \approx 1.32$ ($J < J_{\text{threshold}}$) and exhibits bistability between a resting state and an elevated persistent activity state when $J > J_{\text{threshold}}$. In the case when the maximal

value of J at the top of the hierarchy is smaller than $J_{\text{threshold}}$ (Fig. 3f), isolated areas are unable to generate elevated persistent activity, but working memory representation may emerge from long-distance area-to-area connection loops, thereby extending the concept of synaptic reverberation [15, 16] to the large-scale multi-regional brain.

The connected network of interacting areas exhibits a coexistence of a resting state (where all areas of the network exhibit a low firing rate state $\sim$1 Hz) and active states (where some cortical areas display persistent high firing rates while others display low firing rates). An example is shown in Fig. 3g, where the resting state (brown) and a persistent firing state suitable to underlie working memory representation (blue) are shown as a function of the cortical hierarchy. For the active state, those areas engaged in persistent activity are located higher in the hierarchy and separated by a firing rate gap, indicating a transition zone. The size of this transition zone systematically shrinks with the increasing size of the network (Fig. S6b). We expect that in the limit of infinitely large networks this transition zone shrinks to a point, which we denote as the bifurcation location in hierarchical space (Fig. 3g, Fig. S6b). Interestingly, at the bifurcation point, there is a firing rate gap reminiscent of a first-order phase transition in statistical physics [46].

Mapped into our generative model's ellipsoid where connectivity was originally embedded, the firing rate of both active and resting states increases along its major axis (Fig. 3h). In contrast to the resting state, where the firing rate increases smoothly along the hierarchy (Fig. 3h, lower panel), in the persistent activity state, there is a firing rate gap between the rostral and the caudal areas (Fig. 3h, upper panel). Furthermore, the persistent firing rate increases along the minor axis z (Fig. 3h).

We quantified the timescale of persistent activity fluctuations in the generative network model, using the same autocorrelation analysis as for the connectome-based cortex models (Fig. 2). We found that the time scale of neural fluctuations increases from milliseconds

to tens of seconds for areas near the transition point (Fig. 3i), displaying the critical slowing down phenomenon. The timescale profile along the hierarchy is of an inverted V-shape with fast fluctuations for areas low and high in the hierarchy and very slow fluctuations for areas in the bifurcation region. This inverted V-shape divergence in time scales is characteristic of critical slowing down and is only observed for the active state. In contrast, in the resting state, the timescales increase monotonically with the hierarchy (left panel of Fig. S6h), reproducing the result in the resting state [6, 47] .

Why do areas high in the hierarchy, with long time constants in the resting state, now exhibit short time constants? Imagine that each area can be described by an effective energy potential with the persistent activity corresponding to its minimum (schematically illustrated in Fig. 3i). The time constant quantifies the speed at which the system returns to the bottom of the potential (corresponding to an attractor) after a brief perturbation. The potential would be steep for areas near the top of the hierarchy that display robust mnemonic activity thus with short time constants. By contrast, it would be very shallow for areas near the transition that recover slowly from perturbations [38]. This explanation is not rigorous because areas are not isolated from each other, but offers an intuitive way to understand the inverted-V shaped pattern of timescales along the cortical hierarchy during working memory.

Normal form analysis of bifurcation in space

We observed (Fig. 4a) that the firing rate gap shown in Fig. 3g depends on the gain parameter d if the input-output transfer function of our model as Eq. (8) (see Methods): it is nearly zero when $d \simeq 0.15$, increases progressively with d and saturates at large d where the transfer function becomes threshold-linear (see Methods, Eq. (9)) [48]. With $d = 0.15$, in the spatial profile of persistent activity firing rate is constant and low in areas

below a hierarchical (bifurcation) location, then starts to increase (without a discrete jump) in areas higher in the hierarchy. Nevertheless, the working memory state continues to be characterized by an inverted-V shaped time scale profile (Fig. 4b), suggesting that critical slowing down does not require the presence of a firing gap in the persistent activity state.

The model behavior with large d is approximately reproduced when the transfer function is threshold-linear (Fig. 4c). In this case, the model is much simplified and we undertook analytical calculation of a "normal form" according to mathematical bifurcation theory [49], in order to elucidate rigorously the transition observed in Figs. 2-3 as a new form of bifurcation. In our model, cortical areas indexed by $i = 1, 2, ..., N$ receive long-range excitatory input currents from the network's recurrent interactions, $L_E^i = \sum_j FLN_{ij} S_E^j$ where FLN_{ij} is the connection weight from area j to i and S_E^j is the output synaptic variable of area j (see illustration in Fig. 4d). The quantity L_E^i is a weighted average long-range input; if its value is known, the equation for area i can be solved. However, the input current for each area i depends on its connections (the FLN_{ij} values) and the collective activity state (all S_E^j values). Indeed, the firing rates are driven by L_E^i, which in turn depend on the firing rates themselves; the two must be solved in a self-consistent manner for the entire system (not separately for each area). At the limit of infinite gain, equivalent to a threshold-linear transfer function (see Methods, Eq. (9)), the firing rate as a function of h and L_E can be readily solved, yielding a bistability surface as a function of h and L_E (see Fig. 4e). The normal form of bifurcation in hierarchical space is obtained (see Methods, in particular, Eq. (31)). Using the normal form, we solved self-consistent equations for the N firing rates and N L_E variables. The analytical results match perfectly our numerical simulations (Fig. 4e).

Having rigorously established the notion of bifurcation in space by the normal form

analytical calculations, we further explored the model generally with moderate gain parameter d values. Similar to connectome-based macaque cortex model [20], numerous spatial attractors coexist, each corresponding to persistent activity in a distinct subset of areas. Interestingly, some of these self-sustained states do not show monotonic increase of firing rate along the hierarchy as in Fig. 2a (see also Fig. S9a). Instead, they display a localized *bump* of activity in hierarchy space (Fig. 4f, and Fig. S9c, left panel). Strikingly, unlike the monotonic active states where persistent firing roughly follows the rostro-caudal axis of the model ellipsoid, bump active states generally display more scattered spatial patterns of persistent activity (Fig. 4f and Fig. S9c, right panel). The time constant of each area's neuronal fluctuations in those bump states is maximal and exceptionally long at both edges of the bump (Fig. 4f and Fig. S9c). This double-peaked pattern of timescales further demonstrates critical slowing down as a signature bifurcation in space, which in this case of a bump attractor takes place at two locations.

For each spatial attractor, the firing rate of all cortical areas must lie on a surface parameterized by the hierarchy h and the L_E (Fig. 4g, Fig. S10, Fig. S11, and Fig. S12). The $r(h, L_E)$ surface does not depend on the network size N. In the resting state, both firing rates and weighted long-range excitatory inputs L_E are low (Fig. 4g). In a persistent activity state (Fig. 3g with $d = 0.17$), the more an area is active, the higher its output synaptic gating variable. Therefore, those areas below the transition receive weak input currents from nearby in the hierarchy [41, 42] (Fig. 4g), whereas L_E^i are large for areas above the transition since they receive strong inputs from nearby areas with elevated persistent activity.

We find that the $r(h, L_E)$ surface folds at a *cusp*, corresponding to a point in the two-dimensional space of h and L_E (Fig. S12). This geometry is reminiscent of the cusp described in classical bifurcation theory for non-linear dynamical systems [50, 49]. For

a system that undergoes a cusp bifurcation, the $r(h, L_E)$ surface representing the steady state solution in a two-dimensional parameter space folds at a cusp point [49]. From this point to the folded region in parameter space, the system transitions from having a single to three (two stable and one unstable) steady states. Indeed, in our network, areas with hierarchy values h and L_E within the folded region exhibit bistability (see Methods and Fig. 4g). However, it is essential to highlight that the bifurcation in space in our model is conceptually different from the conventional cusp because L_E is not a control parameter and depends on the firing rates themselves, which are shown in the cloud of dots in Fig. 3g. Recall that when decoupled from each other, none of the isolated areas show elevated persistent activity. Therefore, bifurcation in space is a collective behavior emerging from the complex area-to-area interactions, yet, importantly, the transition occurs at a particular location of the hierarchy.

To further test our model, we considered two "null" models. First, we randomly shuffled area-to-area interactions. In that case, modularity is absent and all areas participate in the persistent activity state (Fig. S8a, blue), even though the firing rate increases along the hierarchy as a result of the macroscopic gradient of excitation. There is no clear pattern for the timescale of neural fluctuations in the persistent state (Fig. S8b) or the resting state (Fig. S8c). Second, the macroscopic gradient of excitation is abolished by randomly shuffling the parameter J between areas. In this case, all the salient behavioral characteristics of our model disappeared (Fig. S8d-f). These results demonstrate that the desired functional modularity depends critically on both the connectomic properties and the macroscopic gradient.

Coexistence of multiple functional modules (spatial attractors)

Fig. 5a-c shows a collection of coexisting spatial attractors, with a given set of fixed parameter values for the generative model, macaque model and mouse model, respectively. Note that spatial attractors are only defined by participating areas, independent of the number of selective neural populations within each area. We shall refer to these spatial attractors as functional modules, as they are well-defined activity patterns, each is self-sustained among a selective subset of cortical areas.

In network analysis, graph theoretical measures were applied to connectomic data [51, 52], *Structural modules* are defined by communities corresponding to mutually exclusive subsets of nodes (e.g., cortical areas) with many more connections among themselves than with the rest of the graph. To assess the relationship between functional modules defined by spatial attractors and structural modules, we quantified the latter by applying the Louvain community detection algorithm [53] to the connectivity matrix, revealing 11, 5 and 4 structural modules for the generative model (Fig. 5d), macaque model (Fig. 5e) and mouse model (Fig. 5f), respectively. Note that for the macaque model, the 5 communities are identified only for the subgraph of 41 areas (out of 91 parcellated cortical areas [10]) for which the FLN data are currently available. Fig. 5g-i illustrate how relations between a functional module and 5 structural modules are quantified. For each of two sample spatial attractors (upper and lower plots of Fig. 5g), a functional module consists of areas with persistent firing rates above a threshold of 10 Hz. In Fig. 5h, Areas (x-axis) are divided according to their memberships to each of five communities (structural modules) (y-axis), shown in green or purple depending on whether they belong to a given spatial attractor (functional module). The Hamming distance between the two (Fig. 5i) defines their similarity. For instance, all members of community 3 are part of the attractor 1, therefore, the Hamming distance is 0; for the attractor 2, no member of

community 1 shows persistent activity, and the Hamming distance is 1. As is clear in these two examples, a spatial attractor typically includes areas in multiple communities.

Figure 5j-l display the histograms of Hamming distances between spatial attractors and communities for the generative model, macaque and mouse connectome-based models, respectively. A peak around 0 for the generative model is indicative of numerous spatial attractors that engage all areas of some communities; but this peak is absent for the macaque and mouse models, which means that when a structural module partakes in a functional module, only some but not all of its member areas are involved. The wide distributions demonstrate that structural modules are insufficient to account for functional modules that are generated by complex large-scale cortical dynamics. As a matter of fact, a persistent activity pattern may depend on reverberatory interactions through short as well as long loops among several areas that are not captured by communities defined by monosynaptic interareal connections.

A diversity of bifurcations in space

Fig. 6 shows 4 sample spatial attractors of the macaque cortex model and 2 spatial attractors of the mouse cortex model, each engaging a subset of areas (right panels: activity map). The timescale τ of neural fluctuations is plotted as a function of the difference in the firing rate $\delta r = r_{\text{active}} - r_{\text{rest}}$ between the persistent activity r_{active} and baseline resting state r_{rest}, revealing a similar pattern (left panels), while its dependence on the hierarchy varies considerably from one to another (inset). Importantly, long timescales correspond to relatively small firing rate differences which correspond to a spatial location separating those areas not involved in the functional module (with $\delta r \simeq 0$) from those that are, demonstrating bifurcation in space in each of the 6 examples. This plot can be used to uncover a bifurcation in space even when the estimated timescale from neural

activity displays complex variations along the cortical hierarchy.

In summary, bifurcation in space is defined separately for each of numerous spatial attractors of distributed persistent activity. In other words, a single network has many bifurcations in space, each for the emergence of a subset of areas (a functional module) engaged in the corresponding persistent activity state and manifested by critical slowing down at its transition spatial location. A *diversity* of timescale patterns along the cortical hierarchy highlights a rich repertoire of observable manifestations of bifurcations in space.

Discussion

Motivated by the experimental advances, theoretical modeling of a multi-regional cortex has come to the fore [54, 7]. Connectome-based models were developed for distributed working memory [21, 20, 22], and more recently are beginning to be applied to perceptual decision-making [55, 56], in line with the idea that working memory and decision-making share a common cortical substrate [57]. The present work aimed at uncovering a specific mechanism for the emergence of functional specialization, under the biological assumptions that in the neocortex, parcellated areas share a common architecture but differ considerably from each other in synaptic excitation and inhibition and that the cortical connectome is dense with an abundance of long-range and weak inter-areal projections. The main findings are fivefold. First, our macaque cortex model incorporates new connectivity data of the area MST and reproduces the experimental observation of a sharp emergence of mnemonic persistent activity between the monosynaptically connected MT and MST areas [19], validating the model. It offers a large-scale circuit account of choice correlates in sensory areas like MT that result from top-down signaling [35]. Second, we show that a connectome-based model of macaque monkey cortex displays functional mod-

ularity for working memory and subjective choice in decision-making. In both macaque and mouse models, as well as a generative cortex model, a transient input to the primary visual cortex triggers self-sustained persistent activity distributed across a subset of areas; timescales of mnemonic neural fluctuations display an inverted-V shaped pattern along the cortical hierarchy. Third, a mathematical analysis using the normal form theory of the generative model rigorously establishes the novel concept of bifurcation in space. The transition occurs locally in spite of the fact that the spatial attractor is a collective phenomenon arising from widespread long-range area-to-area interactions. Fourth, numerous spatial attractors coexist in a single cortical system, each defining a functional module which cannot be inferred from structural modules alone. Fifth, a plethora of bifurcations in space, each corresponding to a spatial attractor, gives rise to a variety of timescale patterns which may be observed in different experiments depending on the internal state and specific task performed by animals.

What are the implications of this theoretical work for experimental studies of distributed representations and functional modularity? (1) The surprising and specific prediction of an inverted-V pattern of timescales (Figs. 2-3) is directly testable with recordings from many cortical areas in a working memory task. (2) Certain cortical areas including the prefrontal cortex exhibit short time constants as an indication of robust mnemonic activity against perturbations [58, 59, 21]. This raises the possibility that while playing a major role in sustaining working memory representations, paradoxically, they may not be sensitive to optogenetic inactivation in behaving animals. (3) In our model, during a mnemonic delay, areas engaged in stimulus-selective persistent activity send sustained top-down synaptic inputs to both excitatory and inhibitory neurons in a sensory area that does not show noticeably elevated spiking activity. Therefore, a specialized module is compatible with the observation [60, 61] (which has been interpreted as evidence

against functional specialization) that the information content of working memory can be decoded in primary sensory areas like V1 by Blood Oxygen Level Dependent (BOLD) fMRI signals, which reflect better synaptic currents than neural spiking activity [62]. (4) The inverted-V shaped timescale profile is likely to hold during decision processes, which would explain why areas like the posterior parietal cortex at intermediate hierarchical locations are particularly suitable for slow time integration of evidence in favor of choice options, whereas areas near the top of the hierarchy operate as strong attractors leading to a categorical decision [31, 63, 59]. (5) When structural and functional data are both available, for instance with structural connectivity measured by diffusion tensor imaging and resting state functional connectivity from fMRI, the functional and structural modules can be compared using the measure of similarity introduced in Fig. 5. (6) In a specific experiment and during a particular behavioral epoch, the timescale may display a complex pattern along the hierarchy (e.g., [2]). In such a case, a plot as a function of the differential firing rate with respect to a baseline (Fig. 6) provides a test for revealing a bifurcation in space associated with the emergence of functional specialization required for ongoing behavior.

To investigate how functional modularity emerges in a multiregional cortex we chose to address working memory because of its central importance in cognition and behavioral flexibility. The question of modularity for working memory encoding is still a matter of debate. At the single-neuron level, there is ample evidence that a subset of cortical areas are involved in working memory maintenance [17, 18]; to what extent primary sensory areas participate in working memory storage as required by behavioral demands remains unclear [64, 65, 61]. The present work addresses how functional modularity, the existence of which is not a given (e.g., absent in our model when the experimentally measured connection weights are randomly shuffled (Fig. S8)), may emerge under the assumption

that the cortex is composed of repeats of a canonical local circuit. We found that the mechanism is mathematically described as a novel form of bifurcation occurring at a transition location in the spatially embedded cortex. The idea of computation near criticality has been proposed [66] for an entire neural system that requires fine-tuning parameters. By contrast, a bifurcation in space is characterized by criticality locally for areas close to the transition, and is robust against parameter changes, which would merely move the spatial location of bifurcation. Moreover, bifurcation is defined for each of numerous spatially extended persistent activity states that can potentially serve various internally driven cognitive processes. For example, one spatially distributed attractor could store sensory information, while another could maintain a behavioral rule that guides sensorimotor mapping, etc. Each of these activity states would be modularly organized in the sense that they selectively engage subsets of areas, mathematically determined by the location of the bifurcation in space.

An observable manifestation of bifurcation in space is critical slowing down near the transition point. Consequently, along the cortical hierarchy, an inverted-V shaped pattern of time constants dominates neural fluctuations during working memory. This contrasts our previous report that during a resting state, the dynamical timescale roughly increases along the cortical hierarchy [6, 7]. But there is no contradiction, as time constants are uniquely defined mathematically only for a linear dynamical system; they critically depend on the internal state under consideration in a strongly recurrent neural circuit described by a highly nonlinear system. The present work thus extends our previous finding of a hierarchy of time constants. We propose that an inverted-V shaped pattern of time constants during working memory represents a sensitive test of the absence or presence of functional modularity. Testing this model prediction warrants the following considerations. First, as noted above, time constant estimates may vary in different behavioral

epochs and tasks. Second, there is heterogeneity of time constants across single cells within an area, therefore sufficient statistical power is needed for a cross-area comparison [67, 68, 36]. Third, a mnemonic delay period must be much longer than to-be-assessed autocorrelation times. Fourth, critical slowing down is manifested near a critical point, but the number of recorded cortical areas is limited; it remains to be seen in experiments how close one can get to a bifurcation locus in the cortical system.

This work focuses on spatial patterns of modular neural representations that are mathematically described as attractor states. The concept is not limited to steady state attractors, however. Generally, attractors can display complex temporal dynamics such as chaos rather than steady states [26], and the attractor paradigm can be consistent with considerable temporal variations of neuronal delay period activity and cell-to-cell heterogeneities [69, 70, 71, 72, 73, 16, 74]. Future research is needed to extend the concept of bifurcation in space beyond steady states.

A limitation of this study is the consideration of a cortex by itself. Previous work [22] demonstrated that incorporating thalamic inputs into a thalamocortical mouse model resulted in similar dynamical properties, suggesting that the thalamocortical connection loop effectively enhances corticocortical connections. Given the thalamus's significant role in modulating cortical excitability in working memory [75], a detailed investigation of its influence, and that of other subcortical structures including the basal ganglia, on large-scale dynamics could be pursued in future studies.

The perspective of dynamical systems has been increasingly appreciated in neuroscience. The theory of dynamical systems emphasizes trajectories over time on a manifold of the state space of neural population activity, now commonly considered in both data analysis and modeling [76, 77, 78]. On the other hand, much less known by neuroscientists is the theory of bifurcation as a mathematical "engine" to explain how quantitative

changes lead to qualitatively distinct behavior via a sharp transition. Here, we rigorously demonstrate a novel type of bifurcation locally in space which, to our knowledge, is the first across disciplines.

In principle, bifurcation in space could take place in a spatially extended physical, chemical, or biological system endowed with a systematic gradient of property variations (e.g., climate catastrophe somewhere but not globally on earth). In contrast to local interactions through diffusion or chemical reactions, interareal cortical interactions involve long-range connections, making it all the more remarkable that criticality can occur in a spatially restricted fashion in a multi-regional cortex. A *recurrent* deep neural network (a hierarchical cortex with many feedback loops) and macroscopic gradients are sufficient to give rise to various spatially distributed persistent activity states, corresponding to several functionally modular networks in our model. Indeed, we found that bifurcation unfolds in space and time, underlying a modular subjective judgment in perceptual decision-making. A separate study on detection as a simple decision lends additional support to this idea [56]. Research along these lines should broadly help us understand the emergence of novel brain capabilities that are instantiated in certain parts of the brain as a result of quantitative changes of properties, providing a mechanistic foundation for functional modularity in the whole brain.

Acknowledgments

We thank Jorge Mejias and Xingyu Ding for help with the cortical model codes of the macaque and mouse, respectively. We thank Loïc Magrou for help with macaque brain maps. **Funding:** This work was supported by James Simons Foundation Grant 543057SPI and the NSF Neuronex grant 2015276 (to X.-J.W.); Swartz Foundation post-

doctoral fellowship (to U.P.-O. and A.B.); Marine Biology Laboratory award funded by William Morton Wheeler Family Founder's scholarship and Lola Ellis Robertson endowed scholarship, and Joachim Herz add-on fellowship (to R.Z.); Agence Nationale pour la Recherche ANR-24-CE37-5022-CONNECTOME (to H.K.); Funded by the European Union (ERC PREDICTION 101142153, to H.K.). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The bulk of the work was done when J.J. was a postdoctoral fellow at NYU.

Author contributions: X.-J.W. was responsible for the concept of bifurcation in space, designed the project, was actively involved in all details throughout the work, and wrote the bulk of the main text. X.-J.W., J.J., and U.P.-O. designed modeling; J.J., R.Z., U.P.-O., and A.B. performed the research with supervision and inputs from X.-J.W.; J.V. and H.K. performed the histological and anatomical analysis of the MST data; all the authors contributed to writing the paper. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All data and code will be available upon publication.

References

[1] Chen, S. *et al.* Brain-wide neural activity underlying memory-guided movement. *Cell* **187**, 676–691 (2024).

[2] Khilkevich, A. *et al.* Brain-wide dynamics linking sensation to action during decision-making. *Nature* **634**, 890–900 (2024).

[3] International Brain Laboratory *et al.* A brain-wide map of neural activity during complex behaviour. *Nature* **645**, 177–191 (2025).

[4] Findling, C. *et al.* Brain-wide representations of prior information in mouse decision-making. *Nature* **645**, 192–200 (2025).

[5] Douglas, R. J. & Martin, K. A. C. Neuronal circuits of the neocortex. *Annu Rev Neurosci* **27**, 419–451 (2004).

[6] Chaudhuri, R., Knoblauch, K., Gariel, M. A., Kennedy, H. & Wang, X.-J. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron* **88**, 419–431 (2015).

[7] Wang, X.-J. Theory of the multiregional neocortex: large-scale neural dynamics and distributed cognition. *Ann. Rev. Neurosci.* **45**, 533–560 (2022).

[8] Fodor, J. A. *The Modularity of Mind: An Essay on Faculty Psychology* (MIT Press: Cambridge, MA, 1983).

[9] Kanwisher, N. Animal models of the human brain: Successes, limitations, and alternatives. *Current Opinion in Neurobiology* **90**, 102969 (2025).

[10] Markov, N. T. *et al.* A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb. Cortex* **24**, 17–36 (2014).

[11] Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).

[12] Gămănuţ, R. *et al.* The mouse cortical connectome, characterized by an ultra-dense cortical graph, maintains specificity by distinct connectivity profiles. *Neuron* **97**, 698–715 (2018).

[13] Baddeley, A. *Working Memory* (Oxford, Britain: Oxford University Press, 1987).

[14] Fuster, J. M. & Alexander, G. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).

[15] Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).

[16] Wang, X.-J. 50 years of mnemonic persistent activity: Quo vadis? *Trends in Neurosci.* **44**, 888–902 (2021).

[17] Leavitt, M. L., Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Sustained activity encoding working memories: not fully distributed. *Trends in Neurosci.* **40**, 328–346 (2017).

[18] Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J. D. The distributed nature of working memory. *Trends Cogn. Sci.* **21**, 111–124 (2017).

[19] Mendoza-Halliday, D., Torres, S. & Martinez-Trujillo, J. C. Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat. Neurosci.* **17**, 1255–1262 (2014).

[20] Mejias, J. F. & Wang, X.-J. Mechanisms of distributed working memory in a large-scale model of the macaque neocortex. *eLife* **11**, e72136 (2022).

[21] Froudist-Walsh, S. *et al.* A dopamine gradient controls access to distributed working memory in monkey cortex. *Neuron* **109**, 3500–3520 (2021).

[22] Ding, X., Froudist-Walsh, S., Jaramillo, J., Jiang, J. & Wang, X.-J. Cell type-specific connectome predicts distributed working memory activity in the mouse brain. *elife* **13**, e85442 (2024).

[23] Amunts, K. & Zilles, K. Architectonic mapping of the human brain beyond Brodmann. *Neuron* **88**, 1086–1107 (2015).

[24] Barbas, H. General cortical and special prefrontal connections: principles from structure to function. *Annu. Rev. Neurosci.* **38**, 269–289 (2015).

[25] Wang, X.-J. Macroscopic gradients of synaptic excitation and inhibition in the neocortex. *Nature Reviews Neuroscience* **21**, 169–178 (2020).

[26] Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering* (Oxford, Britain: Taylor & Francis Group, 2016), second edition edn.

[27] Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).

[28] Song, M. *et al.* Hierarchical gradients of multiple timescales in the mammalian forebrain. *Proceedings of the National Academy of Sciences* **121**, e2415695121 (2024).

[29] Lashley, K. S. *Brain Mechanisms and Intelligence* (Chicago: Chicago University Press, 1929).

[30] Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).

[31] Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).

[32] Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).

[33] Wong, K. F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).

[34] Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Visual neuroscience* **13**, 87–100 (1996).

[35] Wimmer, K. *et al.* Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat Commun* **6**, 6177 (2015).

[36] Shi, Y.-L., Zeraati, R., International Brain Laboratory, Levina, A. & Engel, T. Brain-wide organization of intrinsic timescales at single-neuron resolution. *bioRxiv* 2025–08 (2025).

[37] Scheffer, M. *Critical Transitions in Nature and Society* (Princeton University Press, 2009).

[38] Tredicce, J. R. *et al.* Critical slowing down at a bifurcation. *American Journal of Physics* **72**, 799–809 (2004).

[39] Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).

[40] Kim, Y. *et al.* Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* **171**, 456–469 (2017).

[41] Song, H. F., Kennedy, H. & Wang, X.-J. Spatial embedding of similarity structure in the cerebral cortex. *Proc. Natl. Acad. Sci. (USA).* **111**, 16580–16585 (2014).

[42] Ercsey-Ravasz, M. *et al.* A predictive network model of cerebral cortical connectivity based on a distance rule. *Neuron* **80**, 184–197 (2013).

[43] Theodoni, P. *et al.* Structural attributes and principles of the neocortical connectome in the marmoset monkey. *Cerebral Cortex* **32**, 15–28 (2022).

[44] Coifman, R. R. & Lafon, S. Diffusion maps. *Applied and computational harmonic analysis* **21**, 5–30 (2006).

[45] Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12574–12579 (2016).

[46] Landau, L. D. & Lifshitz, E. M. *Statistical Physics* (Elsevier, 2013).

[47] Li, S. & Wang, X.-J. Hierarchical timescales in the neocortex: mathematical mechanism and biological insights. *Pro. Natl. Acad. Sci. (USA)* **119**, e2110274119 (2022).

[48] Abbott, L. F. & Chance, F. S. Drivers and modulators from push-pull and balanced synaptic input. *Progress in Brain Research* **149**, 147–155 (2005).

[49] Kuznetsov, Y. A., Kuznetsov, I. A. & Kuznetsov, Y. *Elements of Applied Bifurcation Theory*, vol. 112 (Springer, 1998).

[50] Thom, R. *Stabilité Structurelle et Morphogenèse* (New York: W. A. Benjamin Co, 1972).

[51] Sporns, O. Contributions and challenges for network models in cognitive neuroscience. *Nat. Neurosci.* **17**, 652–660 (2014).

[52] Fornito, A., Zalesky, A. & Bullmore, E. *Fundamentals of Brain Network Analysis* (Academic Press: London, 2016).

[53] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

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

[55] Zou, L., Palomero-Gallagher, N., Zhou, D., Li, S. & Mejias, J. F. Distributed evidence accumulation across macaque large-scale neocortical networks during perceptual decision making. *bioRxiv* 2023–12 (2023).

[56] Klatzmann, U. *et al.* A dynamic bifurcation mechanism explains cortex-wide neural correlates of conscious access. *Cell Reports* **44**, 115372 (2025).

[57] Wang, X.-J. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215–234 (2008).

[58] Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat. Neurosci.* **16**, 98–104 (2013).

[59] Murray, J. D., Jaramillo, J. & Wang, X. J. Working memory and decision-making in a frontoparietal circuit model. *J. Neurosci.* **37**, 12167–12186 (2017).

[60] Harrison, S. A. & Tong, F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–635 (2009).

[61] Sreenivasan, K. K. & D'Esposito, M. The what, where and how of delay activity. *Nat. Rev. Neurosci.* **20**, 466–481 (2019).

[62] Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–157 (2001).

[63] Brody, C. D. & Hanks, T. D. Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology* **37**, 149–157 (2016).

[64] Xu, Y. Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences* **21**, 794–815 (2017).

[65] Bouchacourt, F. & Buschman, T. J. A flexible model of working memory. *Neuron* **103**, 147–160 (2019).

[66] Shew, W. L. & Plenz, D. The functional benefits of criticality in the cortex. *The Neuroscientist* **19**, 88–100 (2013).

[67] Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat Commun* **9**, 3498 (2018).

[68] Soltani, A., Murray, J. D., Seo, H. & Lee, D. Timescales of cognition in the brain. *Current opinion in behavioral sciences* **41**, 30–37 (2021).

[69] Baeg, E. H. *et al.* Dynamics of population code for working memory in the prefrontal cortex. *Neuron* **40**, 177–188 (2003).

[70] Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).

[71] Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 394–399 (2017).

[72] Miller, E. K., Lundqvist, M. & Bastos, A. M. Working memory 2.0. *Neuron* **100**, 463–475 (2018).

[73] Barbosa, J. *et al.* Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience* **23**, 1016–1024 (2020).

[74] Panichello, M. F. *et al.* Intermittent rate coding and cue-specific ensembles support working memory. *Nature* **636**, 422–429 (2024).

[75] Guo, Z. V. *et al.* Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545**, 181–186 (2017).

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

[77] Chung, S. & Abbott, L. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology* **70**, 137–144 (2021).

[78] Perich, M. G., Narain, D. & Gallego, J. A. A neural manifold view of the brain. *Nature Neuroscience* https://doi.org/10.1038/s41593–025–02031–z (2025).

[79] Mazurek, M. E., Roitman, J. D., Ditterich, J. & Shadlen, M. N. A role for neural integrators in perceptual decision making. *Cereb Cortex.* **13**, 1257–1269 (2003).

[80] Markov, N. T. *et al.* Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* **522**, 225–259 (”2014”).

[81] Vezoli, J. *et al.* Cortical hierarchy, dual counterstream architecture and the importance of top-down generative networks. *Neuroimage* **225**, 117479 (2021).

Main Figure Captions

Figure 1. Connectome-based model of macaque monkey cortex for distributed perceptual decision-making and working memory. (a) Model scheme. Right: experimentally measured connections between areas as a graph (weight is indicated by line thickness); Left: local and long-range connections between two areas (illustrated by MT and MST) via directed projections to excitatory and inhibitory neural pools. (b) Timeline of a model simulated task that requires both perceptual decision-making and working memory. The model is initially in a resting state. A stimulus is presented for a fixed duration, with the task difficulty parametrically varied from trial to trial by the relative difference in the inputs onto the two stimulus-selective neural pools (blue: A; red: B) in V1, and a judgment is produced by winner-take-all attractor dynamics in a subset of cortical areas (functional module for subjective choice). The choice must be internally stored in working memory over a delay period. The memory-guided response could be an eye saccade or moving a hand towards one of two choice targets, not explicitly modeled here. (c) Psychometric function (accuracy) for the RT and FD tasks and (d) chronometric function (average reaction times in correct and error trials) of the model, compatible with monkey behavior in random dots motion direction discrimination experiments [79]. (e) Time-to-threshold for all brain areas. (f, g) Activity maps across different brain areas exhibit distinct structures in correct (f) and error (g) trials depending on the time from the stimulus onset (indicated in panel titles). Red indicates areas (indexed by i) signaling A ($r^A_{E,i} > r^B_{E,i}$) when input to excitatory population A is larger), and blue indicates areas signaling B ($r^A_{E,i} < r^B_{E,i}$). The location of areas on graph maps is computed by projecting the 3D location of each area in the inflated surface map to 2D, based on [10]. For the temporal dynamics of the activity maps, see Movie 1 (correct trials) and Movie 2 (error trials)

Figure 2. Bifurcation in space of connectome-based cortical models of the macaque monkey (a-d) and mouse (e-h). (a) Lateral view of macaque neocortex surface with areas in color included in the model. (b) Firing rate of 41 brain areas, ranked by the hierarchical position, when the neocortex model is in the delay period working memory state. (c) Firing rate time series of 9 chosen brain areas. (d) Time constants of neural fluctuations in 9 selected brain areas for resting (upper panel) and delay period working memory (lower panel) states. (e) Superior and lateral view of mouse cortex surface with model areas in color. (f) Firing rate as a function of the hierarchical position of 43 brain areas in the delay period working memory state. (g) Firing rate time series of 8 sample areas. (h) Time constants of neural fluctuations in 8 selected areas in resting (upper panel) and delay period working memory (bottom panel) states. During working memory, the timescales of neural activity fluctuations exhibit an inverted-V shaped pattern along the hierarchy, with a peak indicating critical slowing down near the sharp transition locally in space.

Figure 3. Spatially embedded generative model of a cortex. (a) Network connectivity of 1000 parcellated cortical areas produced by a generative model of the mammalian cortical connectivity [41] which is spatially dependent, directed, and weighted. (b) Connectivity matrix of the model. (c) Three-dimensional diffusion map embedding of the generative cortical network. Red dots: $1,000$ cortical areas, black lines: the nearest neighbor links in the embedding space. The axes correspond to the first three principal gradients from the diffusion map embedding (see Methods). Using the diffusion map, the hierarchical position of any area is defined based on Euclidean (brown) or hyperbolic (blue) distance to a starting area at the bottom of the hierarchy (see Methods). (d)

Hyperbolic metric yields a more linear increase than Euclidean metric along the hierarchy. (e) Local circuit model scheme for each cortical region, with recurrently connected excitatory (E) and inhibitory (I) populations. The strength of local and long-range connections' strength is indicated by line thickness. (f) Bifurcation diagram of an isolated cortical circuit (see Eq. (7) in SI). In the large-scale system, the local E-to-E and E-to-I weights are scaled by a factor J, which displays a macroscopic gradient as a function of the hierarchical position h (Inset), with the maximal value J below the threshold. (g) Resting (brown) and persistent activity states (blue) are shown with firing rates plotted against areas ranked by hierarchical position. In the persistent activity state, a subset of areas represents working memory and are separated from the rest of the network by a firing rate gap. (h) Front view (left) and side view (right) of the spatial distribution of the persistent activity state (top) and resting state (bottom) in the generative model's ellipsoid. (i) The time constant of all brain areas for the persistent activity state of panel (b). Potential plots illustrate schematically the idea that neural activity state is more stable for areas at both low and high hierarchical positions (left and right) and marginally stable near the bifurcation (middle), giving rise to a inverted-V-shaped pattern of timescales. (j) Left, time constant of 10 selected brain areas; right, firing rate time series of 8 selected cortical areas for the cortical network in the persistent activity state.

Figure 4. Normal form analysis that establishes mathematically the concept of bifurcation in space. (a) As the gain parameter d increases from 0.157 to ∞, the firing rate gap in the spatial profile of persistent activity progressively widens from zero. (b) Model behavior with the gain parameter $d = 0.157$. When there is no rate gap, firing rates of persistent activity (right) of 8 selected brain areas shows a smooth variation, yet the time constant still exhibits an inverted-V pattern (left). (c) With a sufficiently large gain

parameter d of the input-output transfer function, persistent activity as a function of the hierarchy (orange) is similar to that of the model with a threshold-linear transfer function (purple) in the generative cortex model. (d-e) Normal form analysis with threshold-linear transfer function. (d) Illustration of the mean long-range input current to an area i as a sum of synaptic inputs weighted by FLN_j from all source areas j. The gradient of red and blue colors corresponds to hierarchical positions of areas and the weight of long-distance connections FLN_{ij} from area j to area i, respectively. (e) The bifurcation in hierarchical space occurs at the critical line in the plane of hierarchy h and weighted long-range excitatory input L_E. Inset: The firing rate from network simulations versus the predicted firing rate from the normal form analysis show perfect agreement. (f) A bump spatial attractor in the cortical hierarchy of a generative model. Right: activity pattern in the 3D cortical space with firing rate indicated by color. Left: persistent firing rate (blue) and timescale of neural fluctuations (orange) along the hierarchy h. (g) Firing rates of all the areas in the resting state, a monotonic state, and a bump state lie on the $r(h, L_E)$ surface.

Figure 5. A diversity of spatially distributed attractor states. (a) Activity patterns of 4333 spatial attractors (x-axis) shown along the cortical hierarchy (y-axis). Distributed working memory states display either a monotonic pattern or a bump pattern with elevated activity only in the middle region of the hierarchy. The x-axis corresponds to the rank of all persistent activity states according to the number of areas with firing rates larger than 10 Hz (see Methods). All spatially distributed persistent states coexist in a single realization of the generative network model. Similarly, the macaque cortex model (b) and mouse cortex model exhibit many spatial attractors. (d-f) Structural modules are

defined by the Louvain community detection algorithm applied to the connectivity matrix for the generative model (d), macaque model (e) and mouse model (f). (g-i) Hamming distance between a functional module and each of five structural modules (communities) in the macaque cortex model. (g) Mnemonic persistent activity in the hierarchical space for two sample spatial attractors. A functional module consists of areas with firing rate above a threshold of 10 Hz (dashed horizontal line). (h) Membership to structural modules (y-axis) of individual cortical areas (x-axis). Green indicates participation in the functional module. (i) Hamming distance between a functional module and each of the five structural modules. (j-l) Histogram of Hamming distances of all functional modules (spatial attractors in the top row) with structural modules for the generative model (j), macaque model (k) and mouse model (l). There is no simple relationship between communities and spatial attractors, as functional modules are determined by complex recurrent neural circuit dynamics in the multiregional cortex, which cannot be inferred from structural modules alone.

Figure 6. A plethora of bifurcations in space coexisting in a single cortical system. (a-d) Four sample spatial attractors of a macaque cortex model; (e-f) two sample spatial attractors of a mouse cortex model. Right: activity map showing a subset of areas that defines a functional module. Left: timescale of neural fluctuations as a function of the differential firing rate with respect to the resting state, with a peak at relatively small firing rate differences indicative of critical slowing down as a manifestation of bifurcation in space associated with an individual spatial attractor. Inset: timescale as a function of hierarchy.

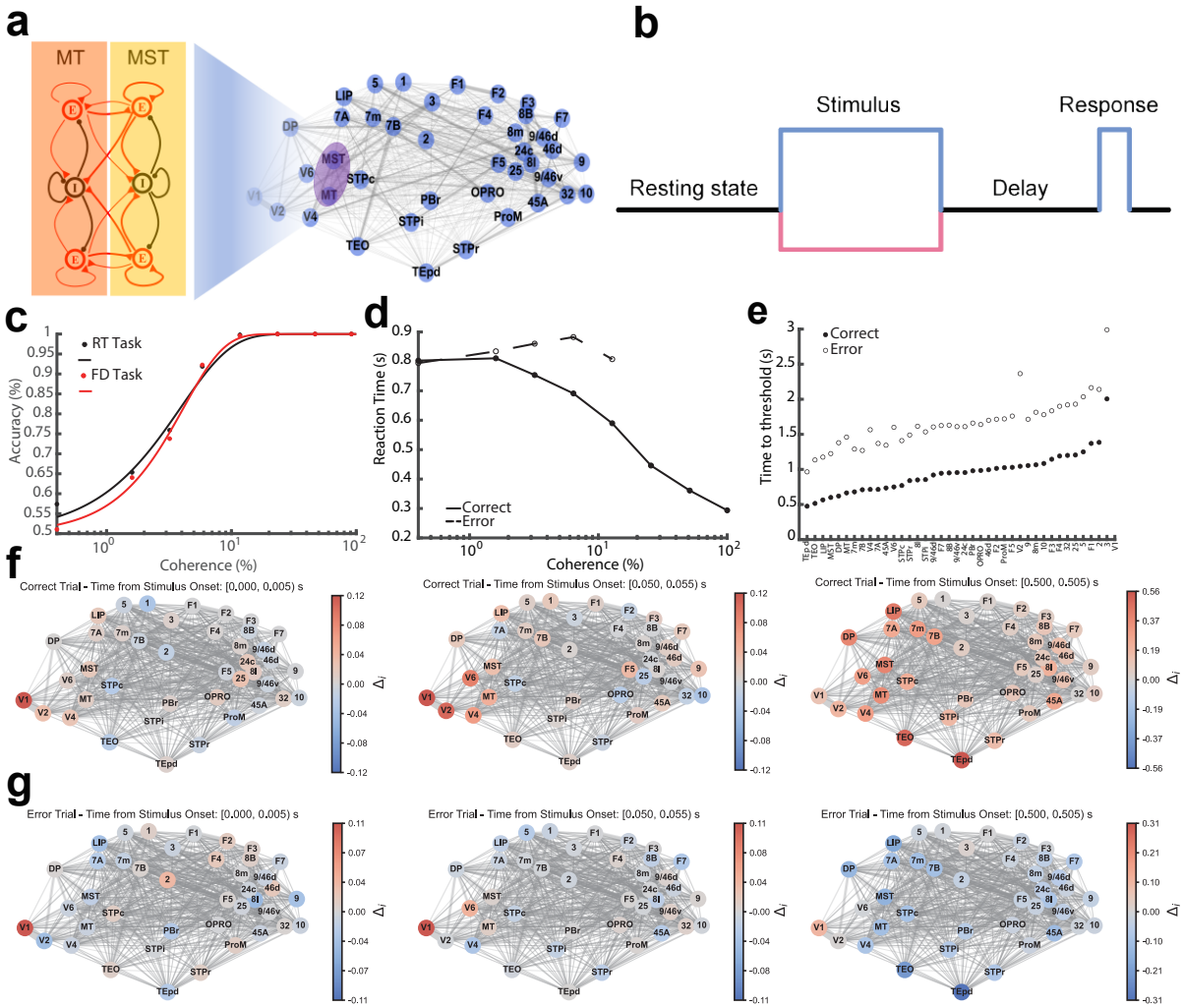
a
MT
MST
b
Stimulus
Response
Resting state
Delay
c
RT Task
FD Task
Accuracy (%)
Coherence (%)
d
Reaction Time (s)
Correct
Error
Coherence (%)
e
Correct
Error
Time to threshold (s)
f
Correct Trial - Time from Stimulus Onset: [0.000, 0.005) s
Correct Trial - Time from Stimulus Onset: [0.050, 0.055) s
Correct Trial - Time from Stimulus Onset: [0.500, 0.505) s
g
Error Trial - Time from Stimulus Onset: [0.000, 0.005) s
Error Trial - Time from Stimulus Onset: [0.050, 0.055) s
Error Trial - Time from Stimulus Onset: [0.500, 0.505) s

Figure 1

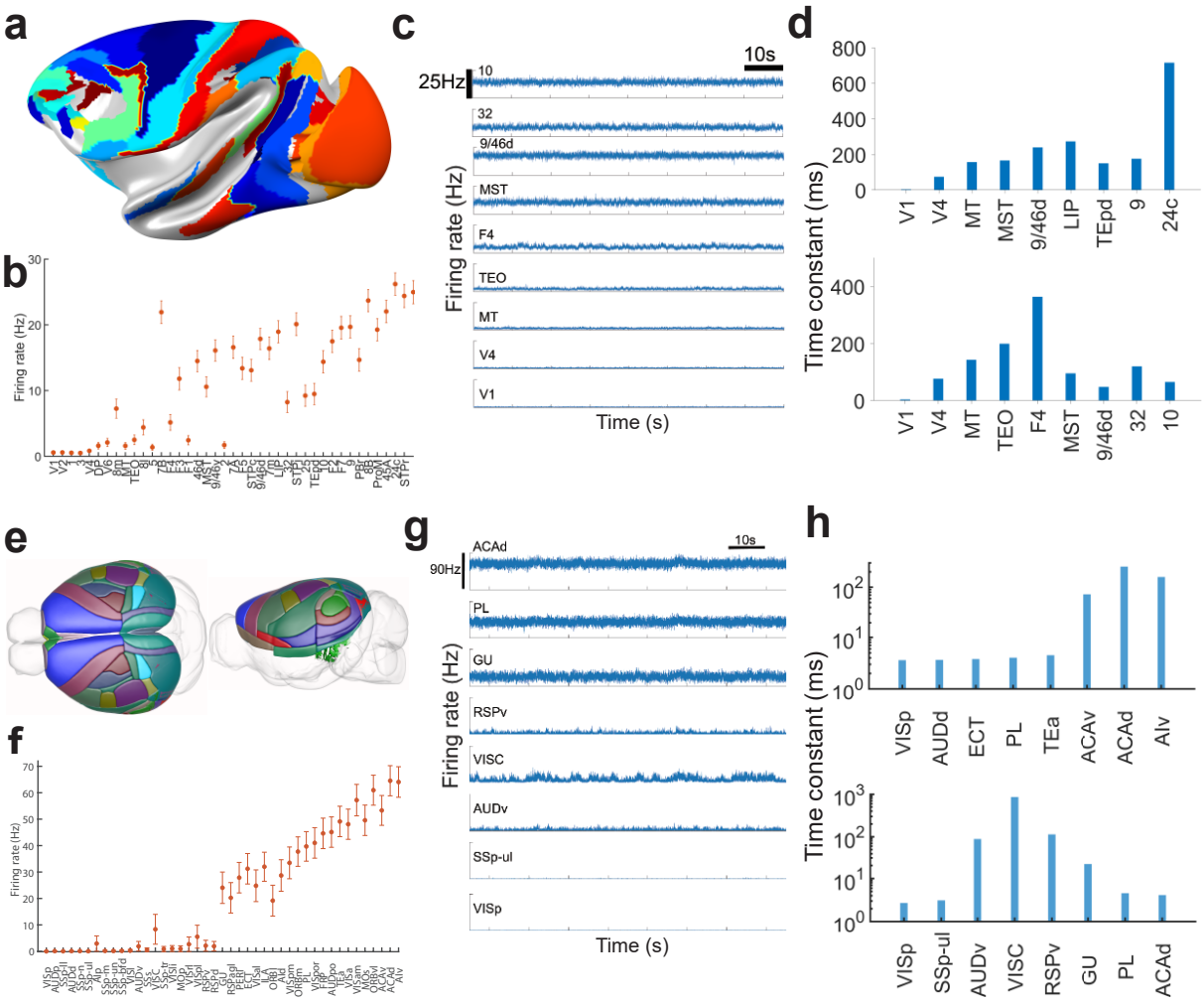
a
b
Firing rate (Hz)
c
25Hz
10s
10
32
9/46d
MST
F4
TEO
MT
V4
V1
Firing rate (Hz)
Time (s)
d
Time constant (ms)
V1
V4
MT
MST
9/46d
LIP
TEpd
9
24c
V1
V4
MT
TEO
F4
MST
9/46d
32
10
e
f
Firing rate (Hz)
g
ACAd
90Hz
10s
PL
GU
RSPv
VISC
AUDv
SSp-ul
VISp
Firing rate (Hz)
Time (s)
h
Time constant (ms)
VISp
AUDd
ECT
PL
TEa
ACAv
ACAd
AIv
VISp
SSp-ul
AUDv
VISC
RSPv
GU
PL
ACAd

Figure 2

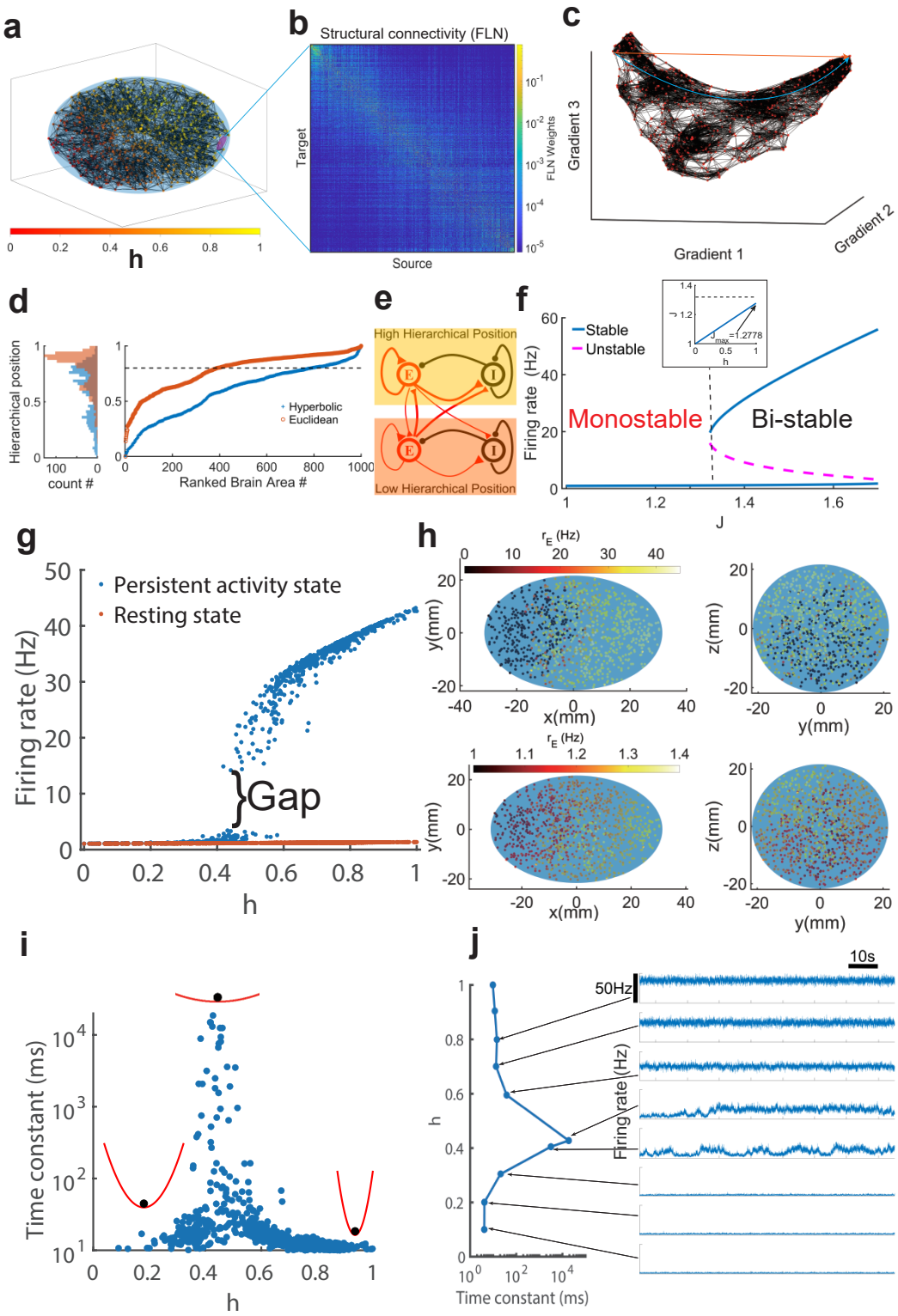
a
b
Structural connectivity (FLN)
Target
Source
FLN Weights
c
Gradient 3
Gradient 1
Gradient 2
h
d
Hierarchical position
count #
Ranked Brain Area #
Hyperbolic
Euclidean
e
High Hierarchical Position
Low Hierarchical Position
f
Stable
Unstable
Monostable
Bi-stable
Firing rate (Hz)
J
g
Persistent activity state
Resting state
Firing rate (Hz)
Gap
h
x(mm)
y(mm)
z(mm)
i
Time constant (ms)
j
10s
50Hz
Firing rate (Hz)
Time constant (ms)

Figure 3

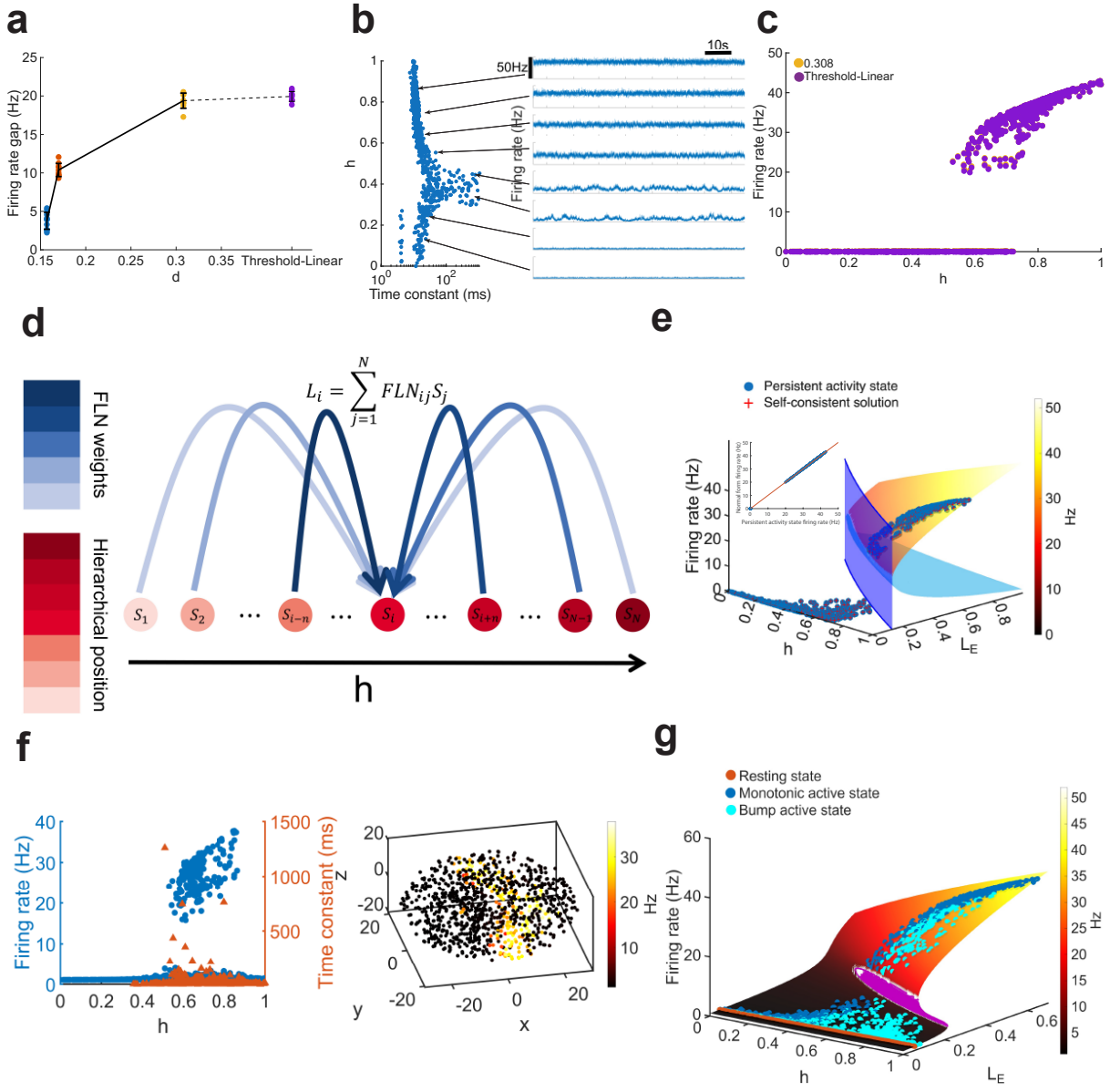
a
Firing rate gap (Hz)
d
Threshold-Linear
b
50Hz
10s
h
Firing rate (Hz)
Time constant (ms)
c
0.308
Threshold-Linear
Firing rate (Hz)
h
d
FLN weights
Hierarchical position
$L_i = \sum_{j=1}^{N} FLN_{ij} S_j$
h
e
Persistent activity state
Self-consistent solution
Firing rate (Hz)
h
L_E
Hz
f
Firing rate (Hz)
Time constant (ms)
h
z
y
x
Hz
g
Resting state
Monotonic active state
Bump active state
Firing rate (Hz)
h
L_E
Hz

Figure 4

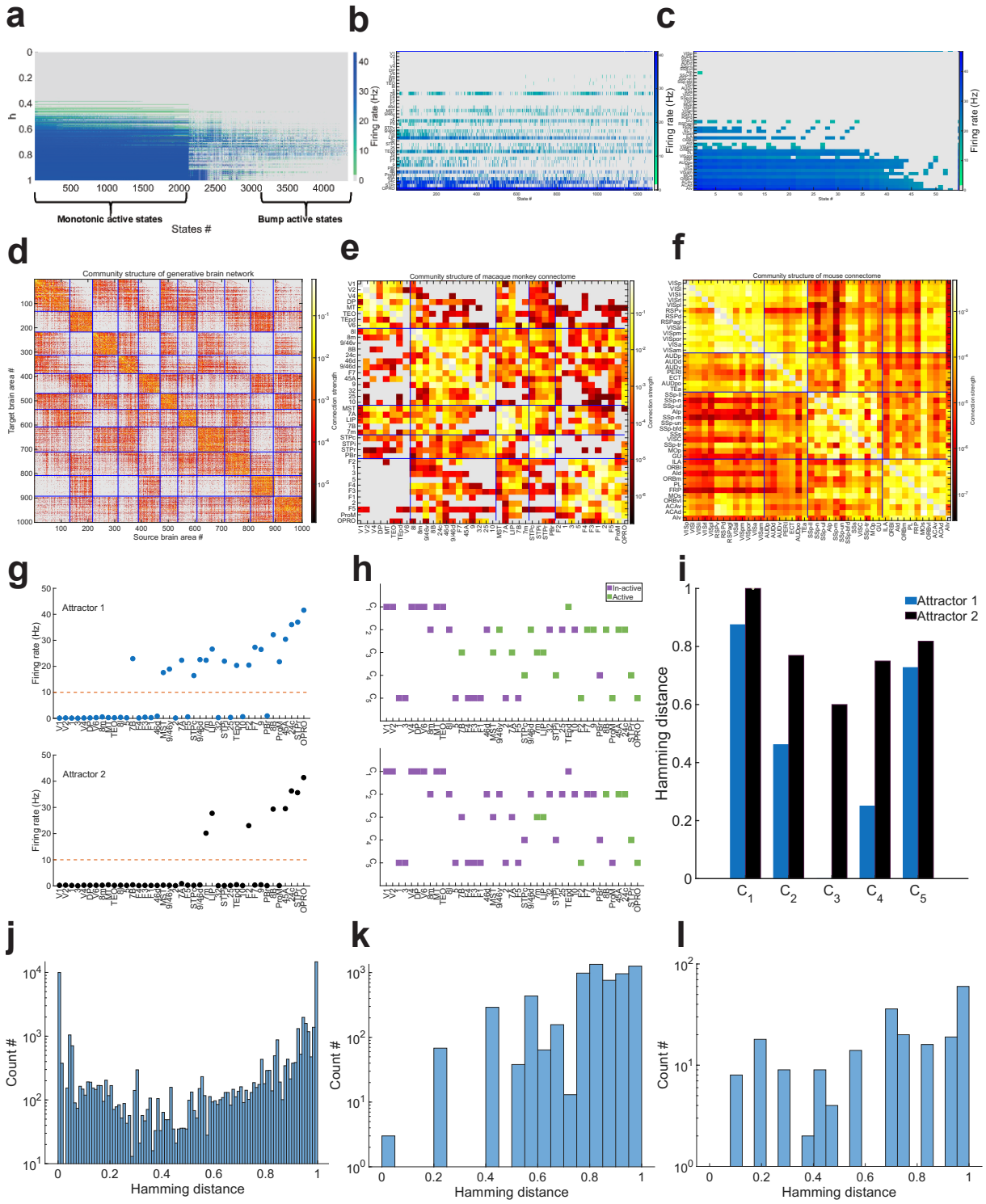
a
b
c
h
Monotonic active states
Bump active states
States #
Firing rate (Hz)
d
e
f
Community structure of generative brain network
Community structure of macaque monkey connectome
Community structure of mouse connectome
Target brain area #
Source brain area #
Connection strength
g
h
i
Attractor 1
Attractor 2
Firing rate (Hz)
In-active
Active
Attractor 1
Attractor 2
Hamming distance
j
k
l
Count #
Hamming distance

Figure 5

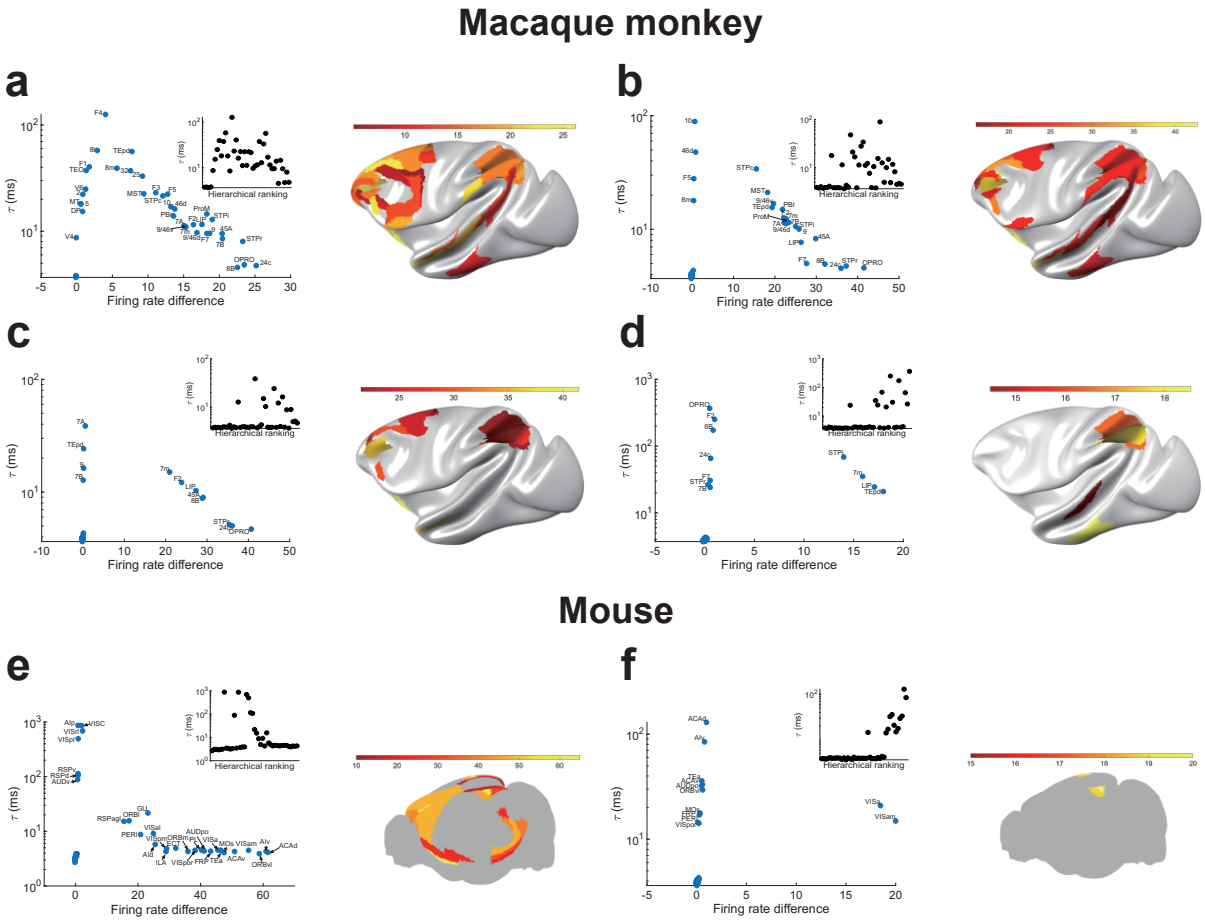
Macaque monkey
a
b
c
d
Mouse
e
f
τ (ms)
Firing rate difference
Hierarchical ranking

Figure 6

Materials and Methods

Anatomical data and numerical parameter set of the connectome-based cortical model of macaque monkey and mouse

The macaque connectome data with 41 cortical areas are shown in Fig.S1a. In particular, we included hitherto unpublished measured connections of the cortical area MST, an addition to the most recent macaque monkey connectome used in Froudist-Walsh et al. [21]. This is important, as MST is an important area of the visual system with the main monosynaptic afferents from the area MT, the former but not the latter exhibit persistent activity during a mnemonic delay [19]. Following the injection of a retrograde tracer into the target cortical area MST, we precisely quantified retrogradely labeled neurons throughout the brain (excluding labeled neurons in the injected target area) and assigned them to supra- and infra-granular layers of all the source areas projecting to the target, as described previously [21, 80]. This allowed us to define two important metrics: i) the fraction of labeled neurons (FLN) representing the normalized strength of the projections from source areas to the target area – MST and ii) the proportion of supragranular labeled neurons (SLN), indexing hierarchical distance between all connected cortical areas relative to the target – MST. SLN is a continuous variable that gradually varies between 1 for long-distance feedforward projections and 0 for long-distance feedback projections, for short-distance projections both feedforward and feedback tend to 0.5. Additionally, the hierarchical position of each brain area can be estimated based on the feedback and feedforward connections among the 41 brain areas [80]. In brief, we used a beta-binomial model, with SLN values of each measured connection of the 41 areas dataset as input, to determine hierarchical levels of all areas based on the model coefficients as detailed in [81]. Importantly, we used numbers of neurons in the model hence generating a weighted

hierarchy. The overall FLN values are shown in Fig. S1a. The resulting hierarchical positions are shown in Fig. S1b. For the cortical model of the macaque monkey model in this paper, all the other parameters are the same as in Mejias et al. [20], except the σ, J_s^{min}, J_s^{max},G, and Z. For Fig 1, Fig 2 a-c, d upper panel, Fig. 6a, Fig. S2 a-e, and Fig. S4, we choose $\sigma = 24pA$, $J_s^{min} = 0.225$, $J_s^{max} = 0.27$, $G = 0.48$, and $Z = 0.82$. For Fig 2 d lower panel, we choose $\sigma = 44pA$, $J_s^{min} = 0.225$, $J_s^{max} = 0.27$, $G = 0.48$, and $Z = 0.82$. For Fig 5b,k, we choose $\sigma = 0pA$, $J_s^{min} = 0.225$, $J_s^{max} = 0.71$, $G = 0.48$, and $Z = 0.25$. For Fig. 5g, upper panel, and Fig 6b, we choose $\sigma = 0pA$ and $\sigma = 16pA$, respectively. The other parameters' are $J_s^{min} = 0.225$, $J_s^{max} = 0.63$, $G = 0.48$, and $Z = 0.42$. For Fig. 5g, lower panel, and Fig 6c, we choose $\sigma = 0pA$ and $\sigma = 18pA$, respectively. The other parameters' are $J_s^{min} = 0.225$, $J_s^{max} = 0.75$, $G = 0.48$, and $Z = 0.27$. For Fig 6d, we choose $\sigma = 10pA$, $J_s^{min} = 0.225$, $J_s^{max} = 0.62$, $G = 0.48$, and $Z = 0.27$. The AMPA noise strength for excitatory populations r_E^A and r_E^B are the same.

The mouse connectome data with 43 cortical areas show in Fig.2, Fig. 5f, and Fig. 6 are same as in Ding et al. [22]. Besides, the model setting is also the same, except for the σ, $g_{E,self}$, μ_{IE}, and β. For Fig. 2e-g, h upper panel, Fig. 6e, and Fig. S2f-g right panel, the parameters set are exactly the same as Ding et al. [22], except the $\sigma = 70pA$. For Fig. 5c,l, we choose $\sigma = 0pA$, $g_{E,self} = 0.45$,$\mu_{IE} = 0.25$, and $\beta = 1.4$. For Fig. 6f, we choose $\sigma = 2pA$, $g_{E,self} = 0.45$,$\mu_{IE} = 0.25$, and $\beta = 1.4$. The AMPA noise strength for excitatory populations r_E^A and r_E^B are the same.

For Fig. 1, in RT task simulations with evidence favoring option A, the RT was read out when $\Delta(t)$ reached a positive threshold for choice A or a negative threshold for choice B. The threshold value was adjusted (to 0.525) for the model to capture RTs comparable to those observed in monkey experiments [79]. The psychometric function is fitted with a Weibull function. Reaction times for correct and error trials are computed separately.

Time-to-threshold and activity maps are computed from the 500 correct or error trials average.

Auto-correlation function of excitatory firing rate and estimated time scales

We calculate the auto-correlation function of each cortical area based on the excitatory firing rate time series. The sample rate and total length of the firing rate time series are $200Hz$ and 80 seconds, which leave out the transient period. First, we calculate the auto-correlation function using the *autocorr* function in Matlab and set the maximum lag as 50 seconds (which is equal to 10000 sample steps). After that, we estimate the time scale of the brain area based on the auto-correlation function. Since the auto-correlation function could have more than one time scale, we fit the auto-correlation using both the single-exponential and double-exponential functions, which shows as follows:

Single-exponential function:

$$ae^{\frac{-\Delta T}{\tau}} + c, \tag{1}$$

Double-exponential function:

$$ae^{\frac{-\Delta T}{\tau_1}} + (1-a)e^{\frac{-\Delta T}{\tau_2}} + c, \tag{2}$$

where ΔT is the time lag of the auto-correlation function, τ is the estimated time constant of a brain area with the single-exponential function, τ_1 and τ_2 are the two estimated time constants of each brain area with double-exponential function. For the double-exponential function, we define a combined time constant $\tau_c = a\tau_1 + (1-a)\tau_2$. However, if $a < 0.07$ or $a > 0.93$, we will choose τ_2 or τ_1 as the final time constant of double-exponential fitting, respectively. Otherwise, we choose τ_c as the final time constant of the double

exponential fitting. We fit the auto-correlation function with the single-exponential and double-exponential functions using the *fit* function in Matlab. For the *fit* function, we set the upper and lower bound for each parameter as $a \in (0, 1)$, $\tau \in (1, \infty)$, $\tau_1 \in (1, \infty)$, $\tau_2 \in (1, \infty)$, $c \in (-1, 1)$, the algorithm of fitting procedure is Levenberg-Marquardt.

In general, We determine each brain area's final time constant based on the fitting's root-mean-square error (RMSE). If the RMSE of the single exponential fitting is larger than two times the RMSE of the double exponential fitting, we choose the double exponential fitting τ_c as the final time constant of the brain area. Otherwise, we choose the single exponential fitting τ as the final time constant of the brain area.

On the other hand, each double exponential fit exhibits a rapid temporal component approximately at 4 Hz, aligning with the temporal dynamics characteristic of AMPA noise of the cortical model of the macaque monkey (Fig. S2d). Consequently, we disregard this rapid component and focus solely on the prolonged temporal scale in Fig. 2d in the main text. However, for some brain areas, such as $V1$, the weight of the rapid time scale is larger than 0.93, which means that this brain area only contains a rapid time scale. For those brain areas, we will use this rapid time scale.

A generative model for the mammalian cortical connectivity

We use the model in [41] to generate multiple cortical network realizations. In this model, we randomly choose the center of N brain areas in a three-dimensional ellipsoid. The ellipsoid is parcellated into N areas through a Voronoi partition. Axon growth starts at a randomly chosen source in the ellipsoid. The direction of growth is determined by the summing force of all the areal centers. Growth length is randomly chosen from an exponential distribution used for modeling the reported distance effects on the connectivity [42]. Since

the axon's source, direction, and length are determined, we can locate the axon's target position in the ellipsoid. We then add a connection from the source area to the target area and repeat the axon growth process $N \times 2.1978 \times 10^4$ times. Through this process, the generative network has a similar in- and out-degree distribution to the actual macaque monkey brain network, measured using retrograde tract-tracing methods and a similar triad distribution. Additionally, based on previous studies [41], a three-dimensional ellipsoid provides a better fit for the connectivity and motif distribution compared to a two-dimensional spheroid. Moreover, the ellipsoid shape is a useful approximation of cortical geometry, reflecting a simple but meaningful structure that effectively captures the hierarchical organization of brain areas, such as the rostro-caudal gradient observed in the cortex.

Diffusion map method for the embedding of connectivity

We analyze the connectivity generated using the diffusion map method [44]. This class of non-linear dimensionality reduction has recently been applied to human and macaque monkey connectomes [45]. Briefly, this method assumes a hypothetical *diffusion process* on the nodes of the symmetric version of the generative network connectivity (that is, the matrix FLN). The symmetry of the network ensures the equilibrium and reversibility of the diffusion process. This diffusion process generates a *diffusion metric space* where the distance between the cortical areas is defined. In this diffusion space, closer areas share more loops, which connect them with stronger connections. On the other hand, areas further apart in diffusion space share fewer loops and weaker connections. When this method is used on connectivity, the cortical network is embedded in a few "principal gradients" of the diffusion process. These principal gradients are the principal components of the normalized graph Laplacian of the diffusion process. This process leads to

embedding the connectivity matrix into a low-dimensional space. Its dimensionality is determined by the selected number of principal gradients (three for Fig. 3c). We applied this method in the symmetric version of the FLN matrix $L = FLN + FLN^T$.

We closely followed the method described in [44]. First, we define the following normalized matrix

$$L_{ij}^{\alpha} = \frac{L_{ij}}{(\sum_{k=1}^{N} L_{ik} \sum_{k'=1}^{N} L_{jk'})^{\alpha}}. \tag{3}$$

Then, we obtain the Markov transition matrix of the hypothetical Markov process on the connectivity as

$$M_{ij} = \frac{L_{ij}^{\alpha}}{\sum_{k'=1}^{N} L_{ik'}}. \tag{4}$$

For the discrete Markov process defined on the connectivity at any time, t, the probability of jumping from the edge j to edge i is given by the matrix M_{ij}^{t}. To define the principal gradients, we perform the eigenvalue decomposition of M_{ij}^{t}, obtaining

$$M_{ij}^{t} = \sum_{l=1}^{N} \lambda_l^t \psi_i^l \phi_j^l, \tag{5}$$

Then, the principal gradients are the set of vectors

$$\left[\lambda_1^t \vec{\psi}^1, \cdots, \lambda_N^t \vec{\psi}^N\right]. \tag{6}$$

where the real eigenvalues are ordered from the largest to the smallest, with $\lambda_1 = 1$, which corresponds to the stationary distribution $\vec{\psi}^1$ for $t \to \infty$. For the dimensionality reduction presented in this work, we use the first three principal gradients $\lambda_2^t \vec{\psi}^2$, $\lambda_3^t \vec{\psi}^3$, and $\lambda_4^t \vec{\psi}^4$.

We examine the Markov process at very short time scales by taking the value $t = 0$. We choose the hyperparameter $\alpha = 0.5$ since the underlying Markov process approximates the Fokker-Planck operator in this case [44].

Constructing cortical hierarchies from the network connectivity

We calculate two classes of hierarchies based on the three-dimensional embedding of the structural connectivity matrix through the diffusion map (see Fig. 3c): Euclidean and Hyperbolic hierarchies. To calculate either class of hierarchy value, we choose the cortical area with the smallest value in the first principal gradient as the first area in the hierarchy (origin area). This choice is arbitrary. We compute the distance in diffusion space between each cortical area and the origin area to determine the hierarchical position. For the Euclidean hierarchy, the hierarchical value of a cortical area i is computed using the normalized Euclidean distance $h^i_{Euc} = dist^{i0}_{Euc}/dist^{max}_{Euc}$. Here, the value $dist^{i0}_{Euc}$ is the Euclidean distance between brain area i and the origin area, and the value $dist^{max}_{Euc}$ is the maximum Euclidean distance of all the brain areas to the origin area. Each brain area's ranked Euclidean hierarchical position is shown as the brown circles in Fig. 3d.

For the hyperbolic distance, we estimate the distance from the origin area along the hyperbolic shape in the embedding space. To do that, we create a nearest-neighbor network for all the brain areas in the embedding space. We link all the brain areas by connecting each brain area with its neighbor within a Euclidean distance threshold $dist^{thr}$. The $dist^{thr}$ is the maximum distance of all the distances between each brain area and its nearest neighbor. The weight of each link in the nearest neighbor network is the Euclidean distance between the two cortical areas. After creating the nearest neighbor network, we estimated the hyperbolic distance between brain area i and the origin area by finding the shortest path between them. The length of the shortest path is the summation of the link

weights (i.e., Euclidean distances) within this shortest path. The shortest path finding is computed using the *dijkstra_path_length* function in the Python package of NetworkX. We define the hyperbolic hierarchical position of brain area i $h^i_{Hyp} = dist^{i0}_{Hyp}/dist^{max}_{Hyp}$, where $dist^{i0}_{Hyp}$ is the Hyperbolic distance between brain area i and the origin area. The value $dist^{max}_{Euc}$ is the maximum Hyperbolic distance of all the brain areas to the origin area. Each brain area's ranked Hyperbolic hierarchical position is shown as the blue dots in Fig. 3d.

Isolated cortical circuit

The simplified nonlinear dynamical model is adopted from [33], which approximated spiking neural network with AMPA, GABA, and NMDA synapses [32]. The dynamical equations that describe the dynamics for a single cortical area are as follows:

$$\begin{aligned}
\tau_E \frac{dS_E}{dt} &= -S_E + \gamma_E \tau_E (1 - S_E) r_E, \\
\tau_I \frac{dS_I}{dt} &= -S_I + \gamma_I \tau_I r_I, \\
\tau_r \frac{dr_E}{dt} &= -r_E + \phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}), \\
\tau_r \frac{dr_I}{dt} &= -r_I + \phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}),
\end{aligned} \tag{7}$$

where S_E and S_I are the gating variables of the NMDA receptor of the excitatory population and the gating variable of the GABAergic receptor of the inhibitory population, respectively, the variables r_E and r_I are the mean firing rates of the excitatory and inhibitory populations, respectively. The functions ϕ_{exc} and ϕ_{inh} are the input-output transfer functions of the excitatory and inhibitory populations. The variable J is the cortical heterogeneity factor, which is proportional to the hierarchical value and differs for each cortical area. Unless specified, parameters are $\tau_E = 60ms$, $\tau_I = 5ms$, $\tau_r = 2ms$, $\gamma_E = 0.76$, $\gamma_I = 1$, $W_{EE} = 276.48pA$, $W_{EI} = 251pA$, $W_{IE} = 129.6pA$, $W_{II} = 54pA$,

$I_{ext,E} = 329.5pA$, $I_{ext,I} = 260pA$. For the input-output transfer function $\phi(I)$, which is a function that transforms the average input current to a cortical circuit into a mean firing rate, we use two different functions:

1. Abbott-Chance function [48]

$$\phi_{exc}(I) = \frac{aI - b}{1 - e^{-d(aI-b)}}. \tag{8}$$

2. Threshold-linear function

$$\phi_{exc}(I) = [aI - b]_+. \tag{9}$$

The notation $[\bullet]_+$ denotes rectification, i.e., $\phi_{exc}(I) = aI - b$ when $aI - b > 0$ and $\phi_{exc}(I) = 0$ when $aI - b \leq 0$.

The parameters for Abbott-Chance and threshold-linear functions are $a = 0.27Hz/pA$, $b = 108Hz$. The parameter d is the gain in the Abbott-Chance function. For a very large gain d, i.e., in the limit when $d \to \infty$, the Abbott-Chance function becomes equal to the threshold-linear function.

For the inhibitory population, the transfer function is threshold-linear

$$\phi_{inh}(I) = [c_1 I - c_0]_+ \tag{10}$$

where the parameters are $c_1 = 0.308Hz/pA$, $c_0 = 77Hz$.

Dynamical model of the generative mammalian neocortex

We connect cortical areas with local neural dynamics described by equations (7-10) using the connectivity from our generative model of the mammalian neocortex. The long-range projections in our model are from excitatory to excitatory populations [6, 21, 20]. Our large-scale model is described as follows

$$
\begin{aligned}
\tau_E \frac{dS_E^i}{dt} &= -S_E^i + \gamma_E \tau_E (1 - S_E^i) r_E^i, \\
\tau_I \frac{dS_I^i}{dt} &= -S_I^i + \gamma_I \tau_I r_I^i, \\
\tau_r \frac{dr_E^i}{dt} &= -r_E^i + \phi_{exc}(J^i(W_{EE} S_E^i + \mu_{EE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{EI} S_I^i + I_{noi}^i + I_{ext,E}^i), \\
\tau_r \frac{dr_I^i}{dt} &= -r_I^i + \phi_{inh}(J^i(W_{IE} S_E^i + \mu_{IE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{II} S_I^i + I_{ext,I}^i), \\
\tau_r \frac{dI_{noi}^i}{dt} &= -I_{noi}^i + \sqrt{\tau_r \sigma_{noi}^2} \xi^i,
\end{aligned} \tag{11}
$$

where the parameters μ_{EE} and μ_{IE} are the long-range coupling strength. Unless specified, $\mu_{EE} = 69.12pA$, $\mu_{IE} = 62.809pA$. The FLN_{ij} is the long-range connection strength from the source brain area j to the target cortical area i, generated as described in the previous section. The parameter J^i corresponds to the excitation gradient. This cortical heterogeneity factor scales excitation for each cortical area i, which is linearly related to the hierarchical position h^i of cortical area i as $J^i = 1 + \eta h^i$ (see inset of Fig. 3f). We assume that all the brain areas have the same external input current $I_{ext,E}^i = I_{ext,E}$ and $I_{ext,I}^i = I_{ext,I}$. The noise term I_{noi} is an Ornstein-Uhlenbeck process, representing the AMPA synaptic noise with a short time constant $\tau_r = 2ms$ [33]. The parameter σ_{noi} is the standard deviation of the noise, and ξ is Gaussian white noise with zero mean and unit variance. Unless specified, all other parameters are the same as in the isolated brain area.

Steady states for an isolated cortical area

For solving the steady state of isolated brain area, we set $\frac{dS_E}{dt} = 0$, $\frac{dS_I}{dt} = 0$, $\frac{dr_E}{dt} = 0$ and $\frac{dr_I}{dt} = 0$. After that, we have the steady-state equations as follows:

$$
\begin{aligned}
\frac{-S_E}{\tau_E} + \gamma_E(1 - S_E)r_E &= 0, \\
\frac{-S_I}{\tau_I} + \gamma_I r_I &= 0, \\
-r_E + \phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\
-r_I + \phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0.
\end{aligned} \tag{12}
$$

A meaningful steady state of brain area must have positive firing rates $r_E \geq 0$, $r_I \geq 0$. Thus, we will have the steady state $0 \leq S_E \leq 1$ and $S_I \geq 0$. Therefore, we could reduce our steady-state equation to

$$
\begin{aligned}
\frac{-S_E}{\tau_E} + \gamma_E(1 - S_E)\phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\
\frac{-S_I}{\tau_I} + \gamma_I\phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0.
\end{aligned} \tag{13}
$$

We reorganize the above expression and obtain an expression for S_I given by

$$
\begin{aligned}
S_I = \gamma_I\tau_I(c_1 I_{inh,t} - c_0) &= \alpha c_1 JW_{IE}S_E + \alpha(c_1 I_{ext,I} - c_0), \\
\alpha &= \frac{\gamma_I\tau_I}{1 + \gamma_I\tau_I c_1 W_{II}} = \frac{1}{\frac{1}{\gamma_I\tau_I} + c_1 W_{II}},
\end{aligned} \tag{14}
$$

where we define $I_{inh,t}$ as a total current input to inhibitory population, and $\alpha = 4.6ms$ by using the parameters of Table. 1. Then, we plug-in equation (14) into the steady state S_E equation (13). After this manipulation, the steady state of the single cortical area is determined by the NMDA gating variable S_E as follows:

$$
\begin{aligned}
-S_E + \gamma_E\tau_E(1 - S_E)\phi_{exc}(\alpha_1 S_E + \alpha_2) &= 0, \\
\alpha_1 &= J(W_{EE} - \alpha c_1 W_{EI}W_{IE}), \\
\alpha_2 &= I_{ext,E} - \alpha W_{EI}(c_1 I_{ext,I} - c_0).
\end{aligned} \tag{15}
$$

Where $\alpha_1 = 230.2305JpA$ and $\alpha_2 = 301.1294pA$ by using the parameters of Table. 1. For the threshold-linear transfer function in equation (9), we immediately noticed that $S_E = 0$, $S_I = \alpha(c_1 I_{ext,I} - c_0)$ is one of the steady states solution with $-W_{EI}(\alpha(c_1 I_{ext,I} - c_0)) + I_{ext,E} < 400pA$. This solution corresponds to the resting state and does not depend on the cortical heterogeneity factor J, which means the resting state always exists along the cortical hierarchy with the threshold-linear transfer function.

However, for the other steady states S_E, they obey the following quadratic equation:

$$-a\alpha_1 S_E^2 + (a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})S_E + (a\alpha_2 - b) = 0.$$

By solving the quadratic equation, we obtain two steady-state

$$S_E = \frac{-(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E}) \pm \sqrt{(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})^2 - 4(-a\alpha_1)(a\alpha_2 - b)}}{2(-a\alpha_1)}, \tag{16}$$

therefore, the isolated brain area has a saddle-node bifurcation of S_E, and the bifurcation point at $(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})^2 - 4(-a\alpha_1)(a\alpha_2 - b) = 0$. At the bifurcation point, we have $\alpha_1^* = J^*(W_{EE} - \alpha c_1 W_{EI} W_{IE})$ and $(a(\alpha_1^* - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})^2 - 4(-a\alpha_1^*)(a\alpha_2 - b) = 0$, Thus $\alpha_1^*(\pm) = \frac{((b - a\alpha_2) + \frac{1}{\gamma_E \tau_E}) \pm \sqrt{\frac{4}{\gamma_E \tau_E}(b - a\alpha_2)}}{a}$. If we consider the solution $\alpha_1^*(-)$ then we have that the cortical heterogeneity factor is given by $J^* = \frac{\alpha_1^*(-)}{W_{EE} - \alpha c_1 W_{EI} W_{IE}} = \frac{((b - a\alpha_2) + \frac{1}{\gamma_E \tau_E}) - \sqrt{\frac{4}{\gamma_E \tau_E}(b - a\alpha_2)}}{a(W_{EE} - \alpha c_1 W_{EI} W_{IE})}$. For our parameter setting, $\frac{\alpha_1^*(-)}{W_{EE} - \alpha c_1 W_{EI} W_{IE}} = 7.1592 \times 10^{-4}$, which means that $J^* \ll 1$. However, this is not possible since $J_{min} = 1$. Therefore, the bifurcation cortical heterogeneity factor is given by $J^* = \frac{\alpha_1^*(+)}{W_{EE} - \alpha c_1 W_{EI} W_{IE}} = 1.3483$.

We performed a similar analysis for the Abbott-Chance transfer function in equation (8). By combining equation (8) and equation (15) the steady state is given by

$$-S_E(1 - e^{-d(a\alpha_1 S_E + a\alpha_2 - b)}) + \gamma_E \tau_E (1 - S_E)(a\alpha_1 S_E + a\alpha_2 - b) = 0. \tag{17}$$

The steady state equation (17) is highly nonlinear, and we can not provide an analytic solution. Instead, we solve equation (17) using the Matlab numerical solver *fsolve*. The steady state in equation (17) depends on the gain parameter d of the Abbott-Chance function, and the bi-stable region enlarges when d increases. This is shown by comparing Fig. 3f, Fig. S5e, and Fig. S5f.

Steady states of the dynamical model of the mammalian neocortex

As for the large-scale network model, we write the steady states equation as follows:

$$\begin{aligned}
&\frac{-S_E^i}{\tau_E} + \gamma_E(1 - S_E^i)r_E^i = 0, \\
&\frac{-S_I^i}{\tau_I} + \gamma_I r_I^i = 0, \\
&-r_E^i + \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}\sum_{j=1}^{N} FLN_{ij}S_E^j) - W_{EI}S_I^i + I_{ext,E}^i) = 0, \\
&-r_I^i + \phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}\sum_{j=1}^{N} FLN_{ij}S_E^j) - W_{II}S_I^i + I_{ext,I}^i) = 0. \qquad (18)
\end{aligned}$$

We assume that stable steady states of the large-scale network model are attractor states. Therefore, in the steady states, the long-range excitatory input for the ith brain area $L_E^i = \sum_{j=1}^{N} FLN_{ij}S_E^j$ is a fixed number. The exact value of long-range excitatory inputs L_E^i depends on the connectivity structure of FLN. Based on this assumption, we could rewrite the steady-state equation of the large-scale network model as

$$\begin{aligned}
\frac{-S_E^i}{\tau_E} + \gamma_E(1 - S_E^i)\phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i) &= 0, \\
\frac{-S_I^i}{\tau_I} + \gamma_I\phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}L_E^i) - W_{II}S_I^i + I_{ext,I}^i) &= 0. \qquad (19)
\end{aligned}$$

After some manipulations, we obtain the expression for S_I^i given by

$$S_I^i = \gamma_I \tau_I (c_1 I_{inh,t} - c_0) = \alpha c_1 W_{IE} J^i S_E^i + \alpha c_1 \mu_{IE} J^i L_E^i + \alpha (c_1 I_{ext,I}^i - c_0), \tag{20}$$

where the definition of α is same as in equation (14). Therefore, for the ith cortical area the excitatory gating variable S_E^i obeys the following steady state equation

$$\begin{aligned} &-S_E^i + \gamma_E \tau_E (1 - S_E^i) \phi_{exc}((W_{EE} - W_{EI} \alpha c_1 W_{IE}) J^i S_E^i \\ &+ (\mu_{EE} - W_{EI} \alpha c_1 \mu_{IE}) J^i L_E^i + (I_{ext,E}^i - W_{EI} \alpha (c_1 I_{ext,I}^i - c_0))) = 0. \end{aligned} \tag{21}$$

First, we analyzed the steady state equation (21) when the transfer function is threshold-linear. The above equation (21) can be written as the steady state of the following set of dynamical equations

$$\begin{aligned} \frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) \\ &= -\gamma_E \tau_E \chi_1 J^i (S_E^i)^2 + (\gamma_E \tau_E \chi_1 J^i - \gamma_E \tau_E (\chi_2 J^i L_E^i + \chi_3) - 1) S_E^i \\ &+ \gamma_E \tau_E (\chi_2 J^i L_E^i + \chi_3), \end{aligned} \tag{22}$$

with

$$\begin{aligned} \chi_1 &= a(W_{EE} - W_{EI} \alpha c_1 W_{IE}), \\ \chi_2 &= a(\mu_{EE} - W_{EI} \alpha c_1 \mu_{IE}), \\ \chi_3 &= a(I_{ext,E}^i - W_{EI} \alpha (c_1 I_{ext,I}^i - c_0)) - b, \end{aligned}$$

where $\chi_1 = 62.1622Hz$, $\chi_2 = 12.6106Hz$, $\chi_3 = -19.9985Hz$ by using the parameters of Table. 1, and steady states value of the synaptic variable of the ith cortical area S_E^i obeys the above quadratic equation equal to zero. Importantly, the steady state of S_E^i depends on the hierarchy value through J^i and the long-range excitatory inputs L_E^i.

Since the steady state equation for the synaptic variables of each cortical area S_E^i in equation (22) is given by a quadratic equation, then the steady state can be calculated by using the quadratic formula. This calculation is similar to the steady state calculations for an isolated cortical area above (see equation (16)). However, in our large-scale network model, the quadratic formula of the network model is also dependent on the hierarchy value through J^i and the long-range excitatory inputs L_E^i. Therefore, the bifurcation happening in the hierarchical space is determined by the following expression:

$$(\gamma_E\tau_E\chi_1 J^i - \gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3) - 1)^2 + 4(\gamma_E\tau_E\chi_1 J^i)(\gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3)) \tag{23}$$

$$= \gamma_E(\tau_E\chi_2)^2(J^i)^2(L_E^i)^2 + 2\gamma_E\tau_E\chi_2 J^i(1 + \tau_E\chi_3 + \tau_E\chi_1 J^i)L_E^i$$

$$+ (1 + 2\gamma_E\tau_E(\chi_3 + \tau_E\chi_3^2) + 2\gamma_E\tau_E\chi_1(\tau_E\chi_3 - 1)J^i + \gamma_E(\tau_E\chi_1 J^i)^2) \tag{24}$$

$$= \gamma_E\tau_E^2(\chi_1^2 + 2\chi_1\chi_2 L_E^i + \chi_2^2(L_E^i)^2)(J^i)^2$$

$$+ 2\gamma_E\tau_E(-\chi_1 + \tau_E\chi_1\chi_3 + (\chi_2 + \tau_E\chi_1\chi_3)L_E^i)J^i + (1 + \gamma_E\tau_E^2\chi_3^2 + 2\gamma_E\tau_E\chi_3) \tag{25}$$

$$= 0,$$

where $J^i = 1 + \eta h^i$ and L_E^i are the scaled hierarchy value and long-range excitatory inputs of ith brain area, respectively. The equation (23) is a constrain equation in the two-dimensional space of hierarchical position h and weighted long-range excitatory inputs L_E. Therefore, equation (23) determines where the bifurcation in space is happening in the two-dimensional hierarchy and long-range excitatory inputs space. We refer to this curve as the critical line. For example, the ith brain area with scaled hierarchical value J^i will give a specific quadratic equation (see equation (24)), which determines the bifurcation long-range excitatory inputs L_E^*. Therefore, J^i and L_E^* determine one of the bifurcation points in the two-dimensional space. The ith brain area will be active only when it has long-range excitatory inputs such as $L_E^i > L_E^*$. From another viewpoint, the bifurcation equation could be a quadratic equation of the scaled hierarchical value J^i (eq. 25). For

a ith brain area with long-range excitatory inputs L_E^i, only when it has a hierarchical position $J^i > J^*$ displays non-zero firing rates. The critical line given by equation (23) is shown in Fig. 4e.

We perform the same analysis for the Abbott-Chance transfer function. The steady-state equation for the large-scale model reads as follows:

$$\begin{aligned}
&\frac{dS_E^i}{dt} = f(S_E^i, L_E^i, J^i) \\
&= -S_E^i(1 - e^{-d(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3)}) \\
&+\gamma_E \tau_E (1 - S_E^i)(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3) = 0.
\end{aligned} \tag{26}$$

We use numerical methods to solve equation (26). Numerically solving the equation (26) will give a steady state surface shown in Fig. 4g, Fig. S10, Fig. S11, and Fig. S12. We refer to this surface as *the* $r(h, L_E)$ *surface*. Any steady-state solution to the network's dynamics will lay on this surface.

Remarkably, our network's $r(h, L_E)$ surface has a very similar geometry to the cusp bifurcation normal form $r(h, L_E)$ surface [49]. The cusp normal form is given by $\frac{dx}{dt} = \beta_1 + \beta_2 x - x^3$, where β_1 and β_2 are two independent parameters [49, 26]. The set of solutions to the steady state equation $\beta_1 + \beta_2 x - x^3 = 0$ gives the cusp normal form $r(h, L_E)$ surface in the (β_1, β_2) parameter space. We refer to this surface as the cusp surface. The cusp surface determines the possible bifurcations that the cusp normal form undergoes [49, 26], and with this, its bifurcation diagram. The cusp bifurcation point is given by $\beta_1 = \beta_2 = 0$. In Fig. S12, for illustration proposes, we overlay β_1 and β_2 axes to highlight the resemblance of our network's $r(h, L_E)$ surface with the cusp surface [49].

Similarly to the cusp surface, in our network's $r(h, L_E)$ surface, the bi-stable region is the region in the J^i and L_E^* parameter space where, for a given active state, brain areas have two stable states: one with low and another with high firing rates. For the $r(h, L_E)$

surface, the bi-stable region increases with the increase of the transfer function gain d, and when $d \to \infty$, the bi-stable region is the maximum. Thus, the $r(h, L_E)$ surface structure depends on the gain parameter d.

The bifurcation in hierarchy space normal form

We derived a reduced equation for the dynamics of our large-scale neocortical network. We refer to this equation as the bifurcation in hierarchy space normal form. Similar to classical normal forms in dynamical systems [49], this is a reduced dynamical equation derived from the network dynamical system, which qualitatively captures the network's nonlinear dynamics close to the bifurcation in hierarchy space. We performed the derivation of this equation analytically for a network with a threshold-linear transfer function.

To calculate this equation, we first calculate the bifurcation points in the network dynamics. Based on the steady-state equation for the threshold-linear transfer function in equation (23), we have the bifurcation point (S_E^*, L_E^*, J^*) fulfill the below equation.

$$\begin{aligned} & f(S_E^*, L_E^*, J^*) = \\ & (\gamma_E \tau_E \chi_1 J^* - \gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3) - 1)^2 \\ & +4(\gamma_E \tau_E \chi_1 J^*)(\gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3)) = 0. \end{aligned} \tag{27}$$

The bifurcation point in our multi-regional network with a threshold-linear transfer function is defined as the point in parameter space where the solutions of the quadratic equation in equation (23) change from complex conjugate to real. This point in parameter space corresponds to the appearance of bi-stability at the single-area level. Areas below the bifurcation point have a single low firing rate stable state. Beyond the bifurcation point, cortical areas have two stable states: one with low and another with high firing rates. To calculate the bifurcation in the hierarchical space normal form, we need to

expand the function f around the bifurcation point (S_E^*, L_E^*, J^*). The expanded function reads as follows:

$$
\begin{aligned}
& f(S_E^i, L_E^i, J^i) = f(S_E^*, L_E^*, J^*) \\
& + \begin{pmatrix} \frac{\partial f}{\partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial J^i}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
& + \frac{1}{2} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix}^T \begin{pmatrix} \frac{\partial^2 f}{\partial (S_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial L_E^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (L_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial L_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial J^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial J^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (J^i)^2}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
& + O(3), \qquad (28)
\end{aligned}
$$

where we have:

$$
\begin{aligned}
\left.\frac{\partial f}{\partial S_E^i}\right|_{S_E^*,L_E^*,J^*} &= -2\gamma_E\tau_E\chi_1 J^* + (\gamma_E\tau_E\chi_1 J^* - \gamma_E\tau_E(\chi_2 J^* L_E^* + \chi_3) - 1),\\
\left.\frac{\partial f}{\partial L_E^i}\right|_{S_E^*,L_E^*,J^*} &= -\gamma_E\tau_E\chi_2 J^* S_E^* + \gamma_E\tau_E\chi_2 J^*,\\
\left.\frac{\partial f}{\partial J^i}\right|_{S_E^*,L_E^*,J^*} &= -\gamma_E\tau_E\chi_1 (S_E^*)^2 + \gamma_E\tau_E(\chi_1 S_E^* - \chi_2 L_E^* S_E^* + \chi_2 L_E^*),\\
\left.\frac{\partial^2 f}{\partial (S_E^i)^2}\right|_{S_E^*,L_E^*,J^*} &= -2\gamma_E\tau_E\chi_1 J^*,\\
\left.\frac{\partial^2 f}{\partial S_E^i \partial L_E^i}\right|_{S_E^*,L_E^*,J^*} &= -\gamma_E\tau_E\chi_2 J^*,\\
\left.\frac{\partial^2 f}{\partial S_E^i \partial J^i}\right|_{S_E^*,L_E^*,J^*} &= \gamma_E\tau_E(-2\chi_1 S_E^* + \chi_1 - \chi_2 L_E^*),\\
\left.\frac{\partial^2 f}{\partial L_E^i \partial S_E^i}\right|_{S_E^*,L_E^*,J^*} &= -\gamma_E\tau_E\chi_2 J^*,\\
\left.\frac{\partial^2 f}{\partial (L_E^i)^2}\right|_{S_E^*,L_E^*,J^*} &= 0,\\
\left.\frac{\partial^2 f}{\partial L_E^i \partial J^i}\right|_{S_E^*,L_E^*,J^*} &= \gamma_E\tau_E\chi_2(1 - S_E^*),\\
\left.\frac{\partial^2 f}{\partial J^i \partial S_E^i}\right|_{S_E^*,L_E^*,J^*} &= \gamma_E\tau_E(-2\chi_1 S_E^* + \chi_1 - \chi_2 L_E^*),\\
\left.\frac{\partial^2 f}{\partial J^i \partial L_E^i}\right|_{S_E^*,L_E^*,J^*} &= \gamma_E\tau_E\chi_2(1 - S_E^*),\\
\left.\frac{\partial^2 f}{\partial (J^i)^2}\right|_{S_E^*,L_E^*,J^*} &= 0.
\end{aligned}
$$

We simplify the expression in equation (28) obtaining

$$
\begin{aligned}
\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) = \zeta_1 (S_E^i)^2 + \zeta_2 S_E^i + \zeta_3, \\
\zeta_1^i &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^i &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} L_E^i + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^i &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} L_E^i + J^i L_E^{i,*} - J^i L_E^i) \\
&\quad -\gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} + \gamma_E \tau_E \chi_2 (J^i L_E^i - J^{i,*} L_E^{i,*})),
\end{aligned} \tag{29}
$$

The above equation (29) corresponds to the bifurcation in hierarchy space normal form. We solve the steady state of the above equation (29) self-consistently and predict the firing rate of the delay period working memory states. The self-consistent equations read as

$$
\begin{aligned}
S_E^i &= \frac{-\zeta_2^{i,sce} - \sqrt{(\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce}}}{2\zeta_1^{i,sce}} \\
\zeta_1^{i,sce} &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^{i,sce} &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^{i,sce} &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + J^i L_E^{i,*} \\
&\quad -J^i \sum_{j=1}^{N} FLN_{ij} S_E^j) - \gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} \\
&\quad +\gamma_E \tau_E \chi_2 (J^i \sum_{j=1}^{N} FLN_{ij} S_E^j - J^{i,*} L_E^{i,*})), \\
\Delta^i &= (\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce} \geq 0, \\
S_E^i &\geq 0.
\end{aligned} \tag{30}
$$

To solve the self-consistent equations, first, we solved equation (25) numerically and obtained $J^{i,*}$ and $L_E^{i,*}$ for the bifurcation point of the ith brain area. Second, we determine the excitatory gating variable $S_E^{i,*}$ by inserting $J^{i,*}$ and $L_E^{i,*}$ into equation (22) and solving numerically the equation (22). Third, we insert (S_E^*, L_E^*, J^*) into the self-consistent equations in equation (30). Lastly, we solve the self-consistent equations in equation (30) and then predict the firing rate pattern of activity during an active state as shown in Fig. 4e. The normal form in equation (29) can be further reduced to a more general expression, as follows

$$\begin{aligned}\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) \\ &= a_1(S_E^i)^2 + (a_2 J^i + a_3 L_E^i + a_4)S_E^i + (a_5 J^i + a_6 L_E^i + a_7 J^i L_E^i + a_8), \\ & a_1, \ldots, a_8 \in R, \end{aligned} \tag{31}$$

where $a_1, \ldots, a_8$ are parameters calculated re-arranging terms in equation (29). The parameter values are $a_1 = -3.3$, $a_2 = 0.08$, $a_3 = -0.67$, $a_4 = 3.11$, $a_5 = 0.1$, $a_6 = 0.3$, $a_7 = 0.3127$, $a_8 = -1.017$ for a state like Fig. 4e. Remarkably, the bifurcation in the hierarchy space normal form has a similar mathematical form (up to a translation) as the saddle-node bifurcation [49]. However, unlike the saddle-node normal form, cortical areas are coupled through the constant and linear coefficients. These coefficients depend on the hierarchy through J^i and the long-range excitatory input current L_E^i. These coefficients represent the network's effect. Therefore, the above equation shows that the bifurcation in space depends on both macroscopic gradients of neuronal properties and the neocortical network structure.

Numerical methods for search of spatial attractor states

In a network with between 1000-10000 brain areas, it is infeasible for our computational resources to find all the possible active states. Therefore, we try to determine as many unique active states as possible by using as many different initial conditions. In practice, we ranked all the 1000 cortical areas by their hierarchical position and then divided them into 20 groups along the hierarchical position. Therefore, each group has 50 cortical regions contiguous in hierarchical position. To reduce the variation of the initial condition, we set the initial condition for each brain area within the same group to be the same.

We obtain the steady state by re-writing equations (19) as self-consistent equations for the variables S_E^i and S_I^i and iterating these equations for finding the steady state solutions for the neural dynamics. The iterated equations read

$$\begin{aligned} S_E^i = & \frac{\tau_E \gamma_E \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i)}{1 + \tau_E \gamma_E \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i)}, \\ S_I^i = & \tau_I \gamma_I \phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}L_E^i) - W_{II}S_I^i + I_{ext,I}^i), \end{aligned} \tag{32}$$

where the long-range excitatory inputs for the ith brain area is given by $L_E^i = \sum_{j=1}^N FLN_{ij}S_E^j$. In any given initial condition, we use only two different initial values for all areas within a group: $S_E = 1$ or $S_E = 0$. Each group may take different values of S_E. Using the above self-consistent equations, we search for the steady states from $2^{20} = 1,048,576$ different initial conditions.

We iterate the equation (32) until the mean absolute difference between two consecutive iterations is smaller than 10^{-10}, or the iteration number is larger than 10000. However, in practice, no initial condition has more than 10000 iterations for 1000 brain areas. After generating an active state, we determine whether it is unique by computing the absolute difference between it and all the unique active states we obtained previously.

Once the sum of the absolute difference of all the brain areas is larger than 0.05, we retain it as a new unique state. Through this process, we obtained 4333 (one of the states in Fig. 5a is the resting state) distinct active states after trying $1,048,576$ initial states. For all the distinct active states, we checked the local stability of the state by calculating all the eigenvalues of the Jacobian matrix. We found that the real part of all eigenvalues in all the active states is negative. Therefore, all the states are locally stable.

Since we lack a set of self-consistent equations for the monkey and mouse connectome models, we have adopted an alternative search strategy: specifically, we simulate the connectome model under different initial conditions. In practice, we first ranked all 41 cortical areas of the monkey and 43 cortical areas of the mouse based on their hierarchical positions. We then divided these areas into 21 groups (for the monkey) and 22 groups (for the mouse), respectively, following their hierarchical order. Consequently, the brain area with the highest hierarchical position in both the monkey and the mouse constitutes a group of its own. For each group, we tested two initial conditions: one high and one low. As a result, we ran the monkey connectome model with $2^{21} = 2,097,152$ distinct initial conditions, and the mouse connectome model with $2^{22} = 4,194,304$ distinct initial conditions. The connectome model simulations were performed using the Euler method, with a total simulation time of 35 seconds and a time step of 10^{-4} seconds. We retained only the final states where the difference in the gating variable between two consecutive steps was smaller than 10^{-7}. After obtaining an active state, we verified its uniqueness by calculating the absolute difference between this state and all previously identified unique active states. If the sum of the absolute differences across all brain areas exceeded 0.01, the state was retained as a new unique active state. Through this process, we identified 1275 distinct active states for the monkey connectome model (Fig. 5b) and 54 distinct active states for the mouse connectome model (Fig. 5c).

Table 1. Parameters for numerical simulations of the generative model		
Parameter	Description	Value
W_{EE}, W_{EI}	local excitatory coupling to E and I population	276.48pA, 251pA
W_{IE}, W_{II}	local inhibitory coupling to E and I population	129.6pA, 54pA
μ_{EE}, μ_{IE}	Long-range excitatory coupling to E and I population	69.12pA, 62.809pA
τ_E, τ_I, τ_{AMPA}	Main E synaptic time constants, I synaptic time constants, AMPA receptor time constants	60ms, 5ms, 2ms
γ_E, γ_I	E and I synaptic rise constants	0.76, 1
$I_{ext,E}$, $I_{ext,I}$	External background inputs	329.5pA, 260pA
a, b	E population f-I curve	0.27Hz/pA, 108Hz
d	E population f-I curve	0.17 (Fig. 3 Fig. 4f,g Fig. 5a,j Fig. S5 Fig. S6 Fig. S8 Fig. S9 Fig. S10 Fig. S11b Fig. S14), 0.157 (Fig. 4b Fig. S7 Fig. S11 Fig. S12a Fig. S13b)
$c1$, $c0$	I population f-I curve	0.308Hz/pA, 77Hz
h	Normalized hierarchical position	[0, 1]
η	Scaling factor of hierarchical position	0.2778
σ_{noi}	Standard deviation of noise	24pA (Fig. 3i, Fig. 2j, Fig. 4b, Fig. S6h, Fig. S15d upper),8pA (Fig. 5B), 6pA (Fig. S9a), 10pA (Fig. S9b), 16pA (Fig. S9c), 44pA (Fig. S15d lower)), 22pA (Fig. S8a-c), 26pA (Fig. S8d-f)

Supplementary figures

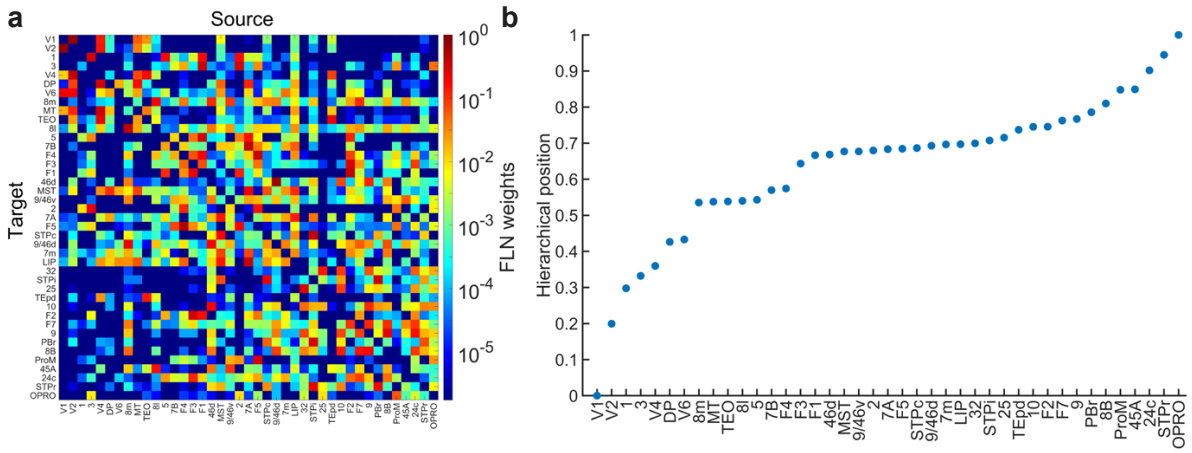

Figure S1: **The FLN matrix and hierarchical position of each brain area of connectome data of the macaque monkey.** (a) The FLN matrix of macaque connectome with 41 brain areas through retrograde tracing [10]. (b) The hierarchical position is estimated based on the 41 brain areas' percentage of supragranular labeled neurons (SLN); the actual method can be found in [80].

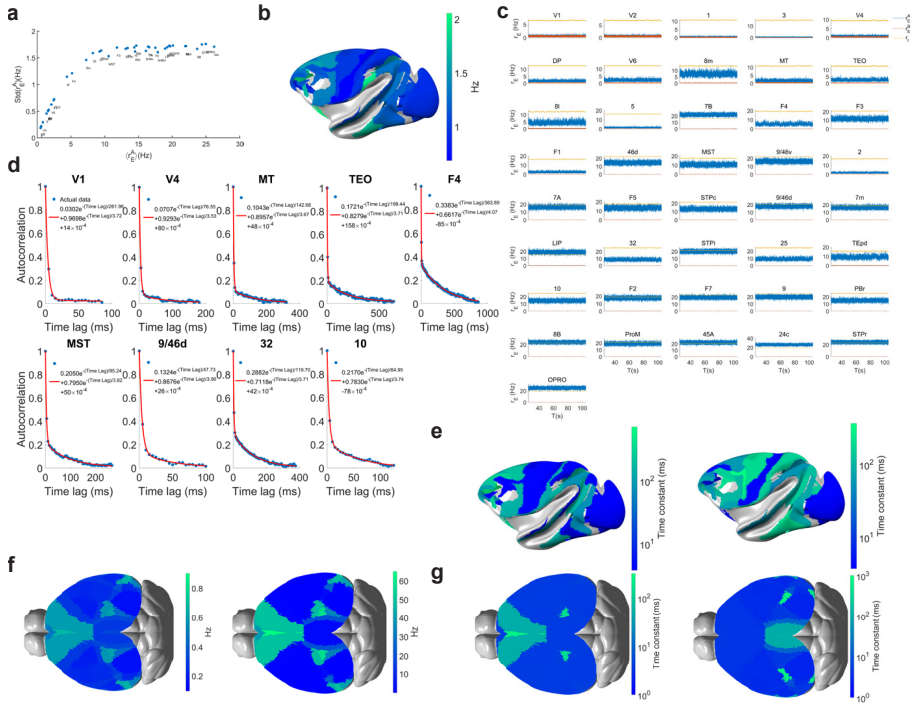

Figure S2: **The supplementary figure of bifurcation in the hierarchical space of connectome-based cortical models of macaque monkey and mouse.** (a) The average firing rate versus the standard deviation of each brain area of the macaque neocortex with 41 brain areas. (b) Spatial activity map of the resting state of the macaque neocortex model with 41 brain areas with the model in [20]. (c) The firing rate of excitatory population A (blue), B (brown), and inhibitory population (yellow) of the 41 brain areas. (d) The autocorrelation function and the related double exponential fitting function of the nine chosen brain areas are in Fig. 2d in the main text. (e) The spatial time constant map of 41 brain areas for resting state (left) and delay period working memory state (right) corresponds to the states of panel b and Fig.6a right panel. (f) Spatial activity map of resting state (left) and the delay period working memory state (right) of the large-scale mouse brain model with 43 brain areas with the model in [22]. (g) The spatial time constant map of 43 brain areas for resting state (left) and delay period working memory state (right) corresponds to the states of panel f.

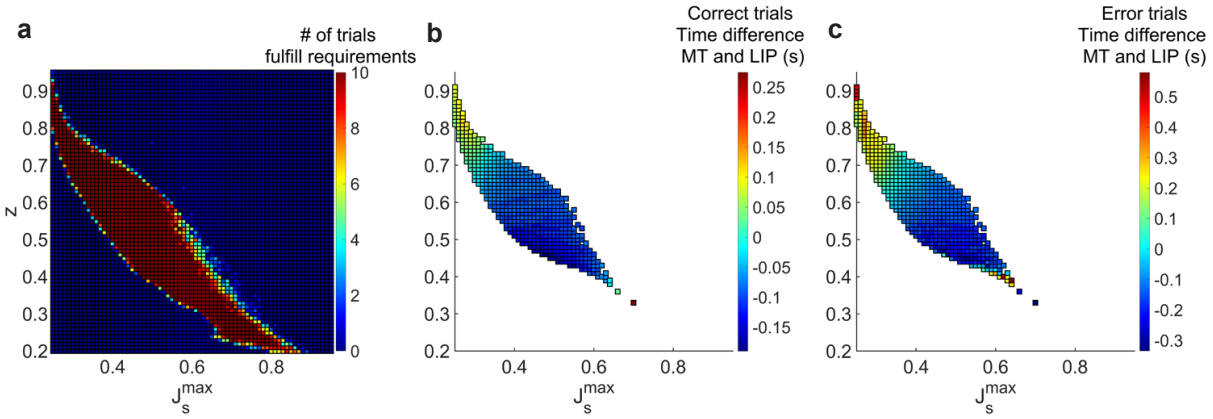

Figure S3: **The parameter region in J_s^{max} and z space that fulfill all the requirements of working memory and decision-making of connectome-based cortical models of the macaque monkey with** 41 **brain areas.** We updated Mejias' model [20] with 41 brain areas and it's corresponding hierarchical value. After keeping all the other parameters the same as the original model [20], we search the parameter set of J_s^{max} and z that fulfill requirements (the resting state should be stable for all the brain areas, the persistent activity shows in some brain area to encoding working memory, and the brain area MST should have the ability to have persistent activity but the brain area MT should not have to be consistent with experiment [19]) (a) shows the number of trials that fulfill all the requirements of the parameter set of J_s^{max} and z. We did 10 trials for each parameter set. If one parameter set with all those trials fulfills all the requirements, it has a value of 10. If no trial fulfills all the requirements, it has a value of 0. (b) shows the mean difference in time to threshold = 0.55 between MT and LIP for correct trials during decision-making across parameter sets in A. We simulated 5000 trials for each pair of J_s^{max} and z, identifying parameters showing both correct and some error trials. The distribution highlights parameter sets with the highest positive difference between MT and LIP during correct trials. (c) shows the mean difference in time to threshold = 0.55 between MT and LIP for error trials using the same set of parameters as panel b. The parameters $J_s^{max} = 0.27$ and $z = 0.82$ were chosen for further analysis based on their optimal performance in correct trials and maintaining J_s^{max} below the critical bistability value of 0.4655. For these values, 200 correct and error trials were generated for additional analyses.

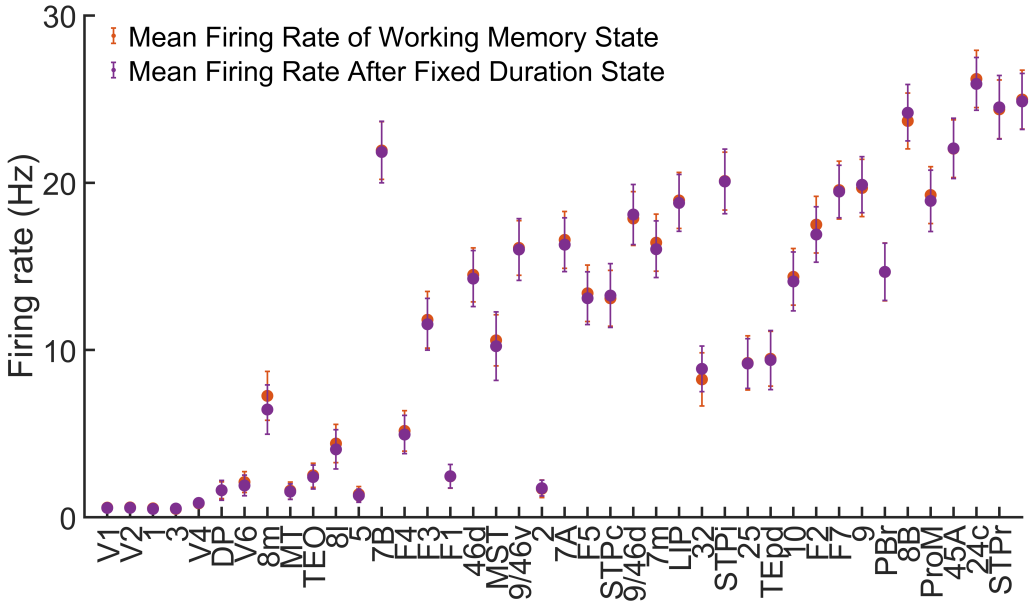

Figure S4: **The delay-period activity for the fixed-duration decision-making task and the working memory state is depicted.** The dots indicate the average firing rate, while the error bars represent the standard deviation of the firing rate.

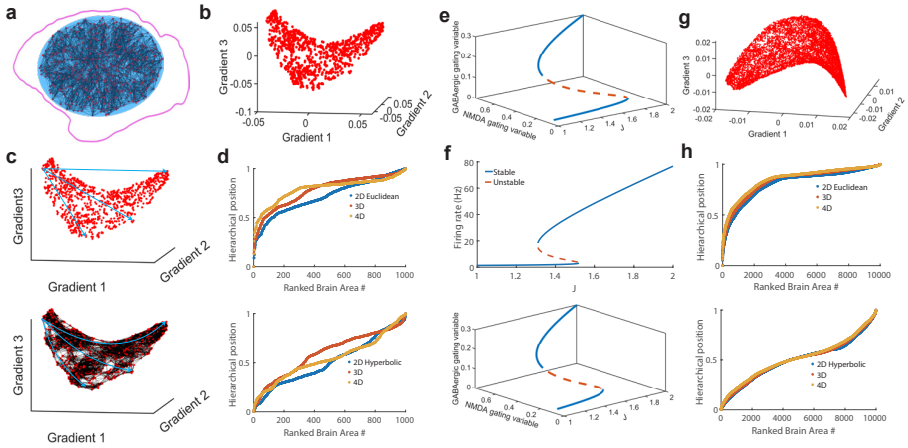

Figure S5: **The supplementary figure of generated connectivity, hierarchy, and local circuit model.** (a) The generated super neocortex network with 1000 brain areas is consistent with the macaque neocortex network statistic. (b) The three-dimensional embedding hyperbolic shapes of the generated super neocortex network of panel a. (c) illustrate Euclidean (upper) and hyperbolic (bottom) distance in the three-dimensional embedding of the super neocortex network. (d) The Euclidean (upper) and hyperbolic (bottom) hierarchical position of all the brain areas. The blue, brown, and yellow lines correspond to the generated super neocortex model's two-dimensional, three-dimensional, and four-dimensional embedding. (e) The NMDA and GABAergic gating variables' steady states vary with the hierarchical position of an isolated brain area with simplified dynamics. This isolated brain area also has the same parameter settings as Fig. 3 f in the main text. (f) The upper part displays the bifurcation diagram of an isolated brain area with simplified dynamics, with a gain parameter of $d = 0.157$. In the lower part, the steady states of the NMDA and GABAergic gating variables in this isolated brain area vary with its hierarchical position. (g) The three-dimensional embedding hyperbolic shape of the generated super neocortex with 10000 brain areas. (h) The Euclidean (upper) and hyperbolic (bottom) hierarchical position of all the brain areas for the 10000 brain area network. The blue, brown, and yellow lines correspond to the two-dimensional, three-dimensional, and four-dimensional embedding of the generated super neocortex model with 10000 brain areas.

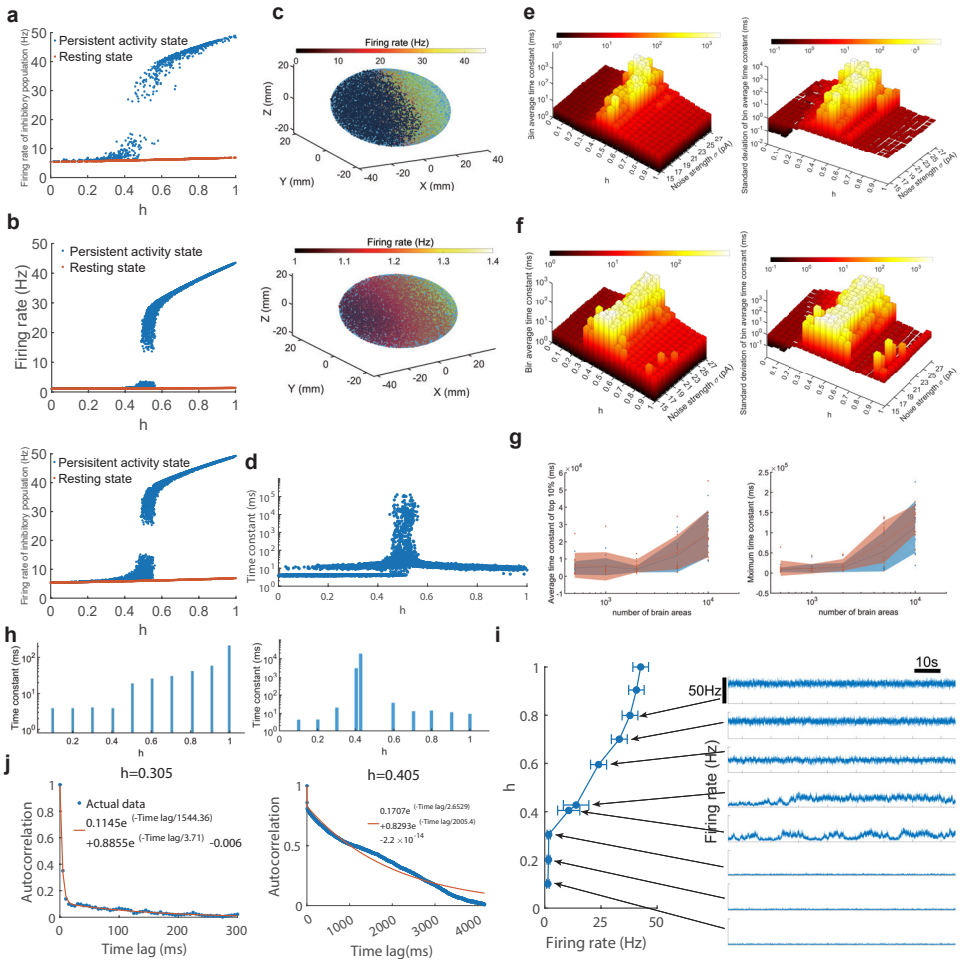

Figure S6: **The supplementary figure of bifurcation in hierarchical space.** (a) Firing rate of all the inhibitory populations for the same parameter set of panel a of Fig. 3g of the main text. (b) The firing rate of the excitatory (upper) and inhibitory (bottom) population of both active (blue) and resting (brown) state of all the brain areas in the generated super-neocortex network with $10,000$ brain areas. (c) The spatial distribution of monotonic active state persistent firing rate (upper) and resting state firing rate (bottom) in the actual ellipsoid space with $10,000$ brain areas. (d) The time constant of all the brain areas at the monotonic active state of panel b with $10,000$ brain areas and noise strength $\sigma = 24pA$. (e) The distribution of bin average (left) and standard deviation of bin-averaged time constant (right) along the hierarchical position change with the noise strength, which increases from $15pA$ to $27pA$, of 5 noise ensemble. Continue to the next page.

Continue with figure S6's caption. In this panel, we use the bin-averaged time constant of the same set of panel a with a bin size equal to 0.05 hierarchical position interval. We averaged from 5 ensemble realizations for each time constant bin. (f) The distribution of bin average (left) and standard deviation of bin-averaged time constant (right) along the hierarchical position change with the noise strength, which increases from $15pA$ to $27pA$, of 5 noise ensemble and 10 network ensemble. The bin size is the same as in panel E, but the average included 10 in different super neocortex networks. We averaged from 5 noise ensemble realizations and 10 different super-neocortex networks for each time constant bin. (g) The average time constant of 10% largest time constant (left) and the maximum time constant (right) change with the network size. The 10 dots at a specific number of brain areas mean the 10 different super neocortex networks. The shaded region represents within one standard deviation. (h) The time constant of 10 chosen cortical areas in the persistent activity state (right) and 10 selected areas in the resting state (left). The states correspond to the states in panel g of Fig. 3. (i) The average and standard deviation of the firing rate of 10 chosen cortical brain areas in the persistent activity state and their corresponding firing rate time series. (j) The autocorrelation and double exponential fitting of two selected brain areas' $h = 0.305$ and $h = 0.405$. The brain areas $h = 0.305$ and $h = 0.405$ are at the bottom and top of the active states' inverted V shape of the time constant.

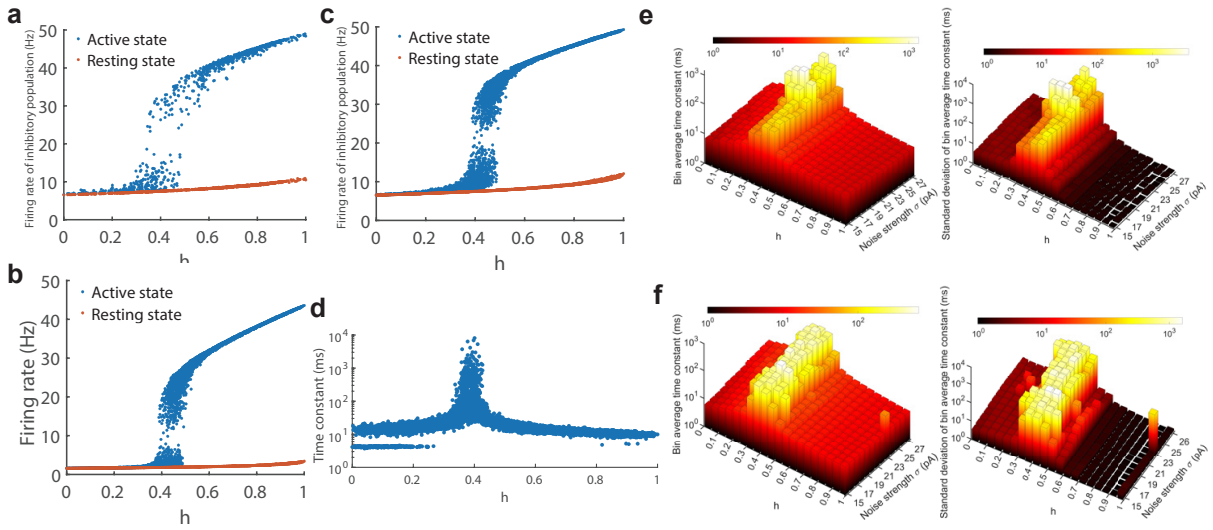

Figure S7: **The delay activity state and time constant for brain networks with** 10000 **brain areas and** $d = 0.157$**.** (a) Firing rate of all the inhibitory populations for the same set of panel b of Fig.4 of the main text with $d = 0.157$. (b) The firing rate of the excitatory population of both active (blue) and resting (brown) state of all the brain areas in the generated super neocortex network with $10,000$ brain areas and the same parameter setting as panel a. (c) The firing rate of the inhibitory population of both the active and resting state of all the brain areas with the same setting as panel b. (d) The time constant of all the brain areas at the monotonic active state of panel b with $10,000$ brain areas and noise strength $\sigma = 24pA$. (e) The distribution of bin average (left) and standard deviation of bin-averaged time constant (right) along the hierarchical position change with the noise strength, which increases from $15pA$ to $27pA$, of 5 noise ensemble. In this panel, we use the bin-averaged time constant of the same set of panel with a bin size equal to 0.05 hierarchical position interval. We averaged from 5 ensemble realizations for each time constant bin. The other parameters are the same as in panel d. (f) The distribution of bin average (left) and standard deviation of bin-averaged time constant (right) along the hierarchical position change with the noise strength, which increases from $15pA$ to $27pA$, of 5 noise ensemble and 10 network ensemble. The bin size is the same as in panel e, but the average included 10 in different super neocortex networks. We averaged from 5 noise ensemble realizations and 10 different super-neocortex networks for each time constant bin.

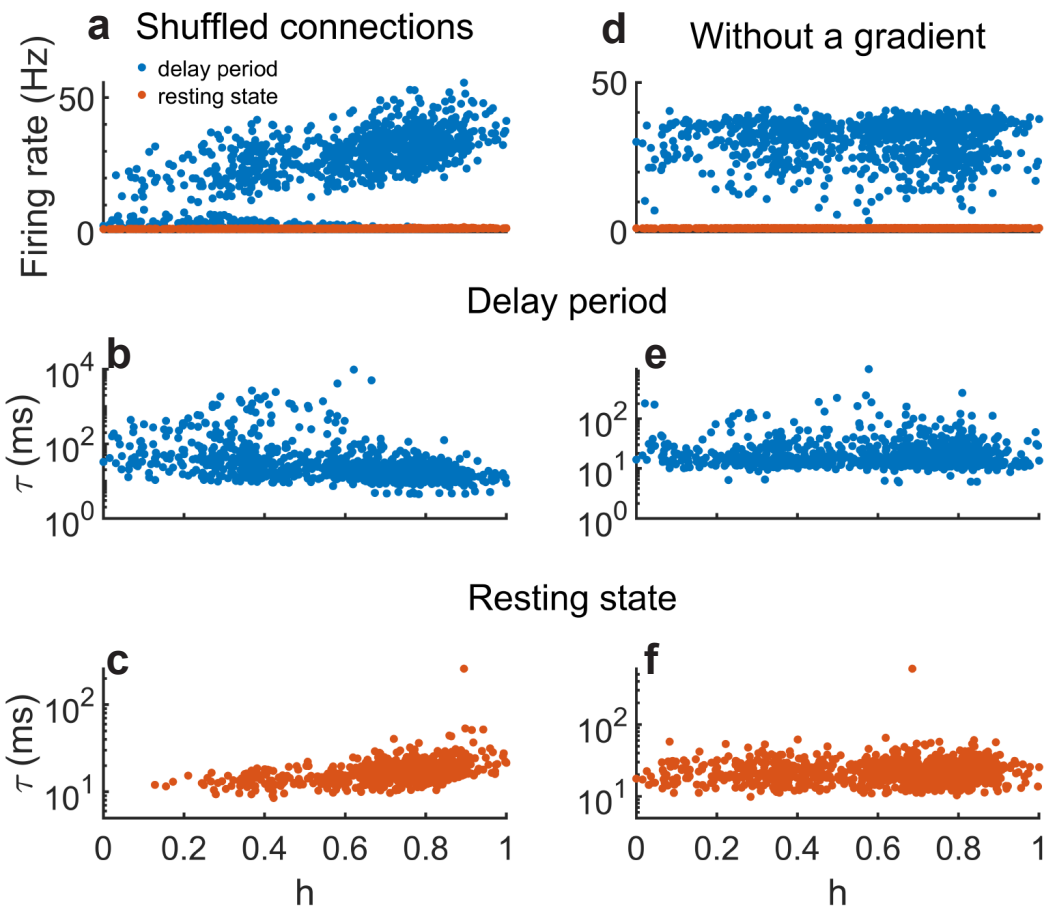

Figure S8: The firing rates and time constants of distributed attractor states when the network connection weights are randomly shuffled (a-c), or the macroscopic gradient of excitation is abolished by randomly shuffling the parameter J (d-f). In either case, the model still displays the bistability of a resting state (orange) and an elevated persistent activity state (blue). However, modularity disappears because all the areas participate in the persistent activity state, and there is no inverted-V shaped profile of time constants in the active state.

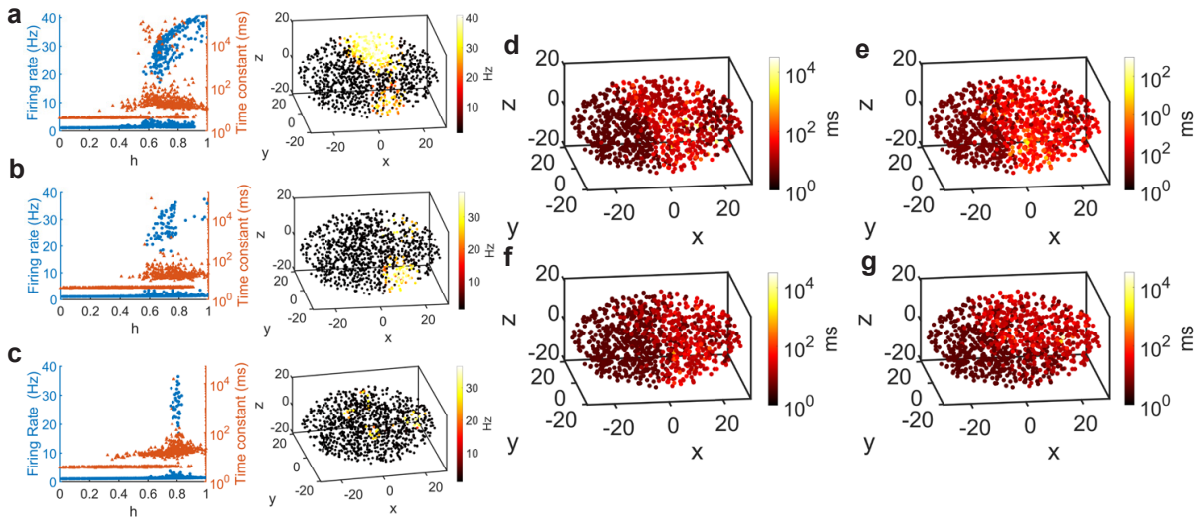

Figure S9: **The supplementary figure of the diversity of distributed working memory states.** (a-c) The firing rate (blue) and time constant (brown) of active state $S1$, $S3 = 3454$, and $S4 = 4068$ (left) for each cortical area, respectively. The state $S1$, $S2$, and $S4$ correspond to the 2300th, 3454th, and 4068th states of the 4333 active states show in Fig. 5a. The spatial distribution of active states $S1$, $S3$, and $S4$ (right) firing rate in the generative model ellipsoid, respectively. (d-g) In the generative model ellipsoid, the spatial distribution of active state $S1$, $S2$, $S3$, and $S4$ time constants, respectively. $S2$ shows in Fig. 4f, and is the 2489th state of the 4333 active states in Fig. 5a.

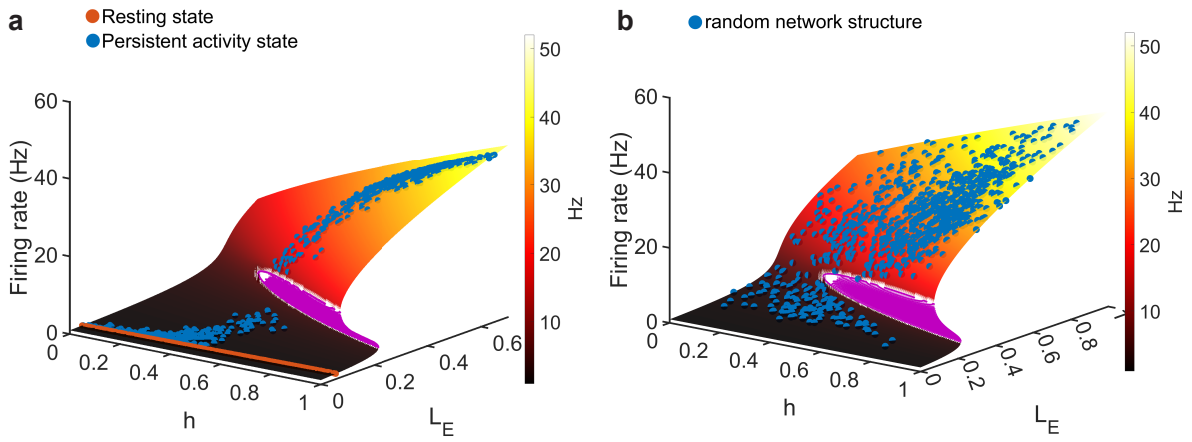

Figure S10: **The geometry of distributed attractor states in parameter space.** (a) The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the $r(h, L_E)$ surface with $d = 0.17$ (corresponding to Fig. 3g). (b) The mapping of the delay active state for randomly shuffled network connections of the generative network to the $r(h, L_E)$ surface (corresponding to Fig. S8a).

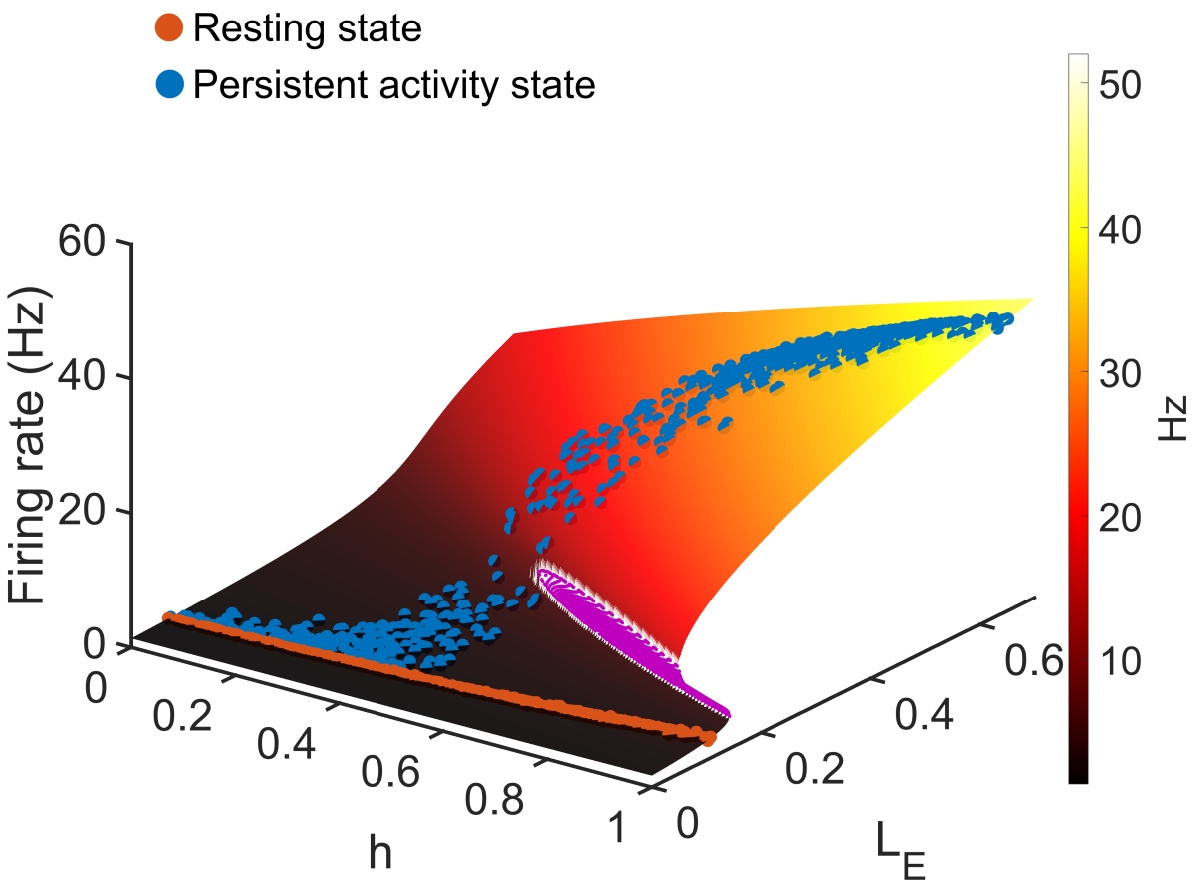

Figure S11: **The $r(h, L_E)$ surface of $d = 0.157$.** The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the $r(h, L_E)$ surface ($d = 0.157$). In this case, the active state continuously transitions without a firing rate gap.

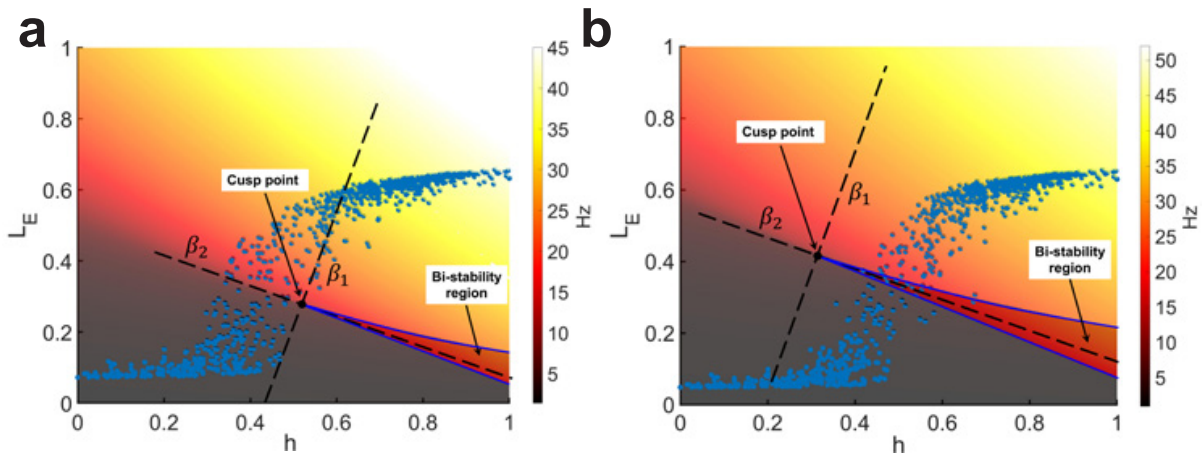

Figure S12: **Cusp geometry determines the bifurcation in hierarchical space.** (a and b) The geometry of the $r(h, L_E)$ surface and the cusp point determine the bifurcation in hierarchical space. The blue dots correspond to the persistent activity state of the neocortex model with $d = 0.157$ (panel a) and $d = 0.17$ (panel b), respectively. For comparison purposes, the axes β_1 and β_2, which correspond to the two control parameters in the cusp bifurcation normal form (see Methods), overlay on the $r(h, L_E)$ surface. From the figure, we know that for areas low in the hierarchy, the firing rate increases smoothly with h and L_E. Beyond this cusp point, the $r(h, L_E)$ surface is folded for hierarchy values h and weighted long-range excitatory input current L_E in the ranges $0.6 - 1$ and $0.1 - 0.2$, respectively.

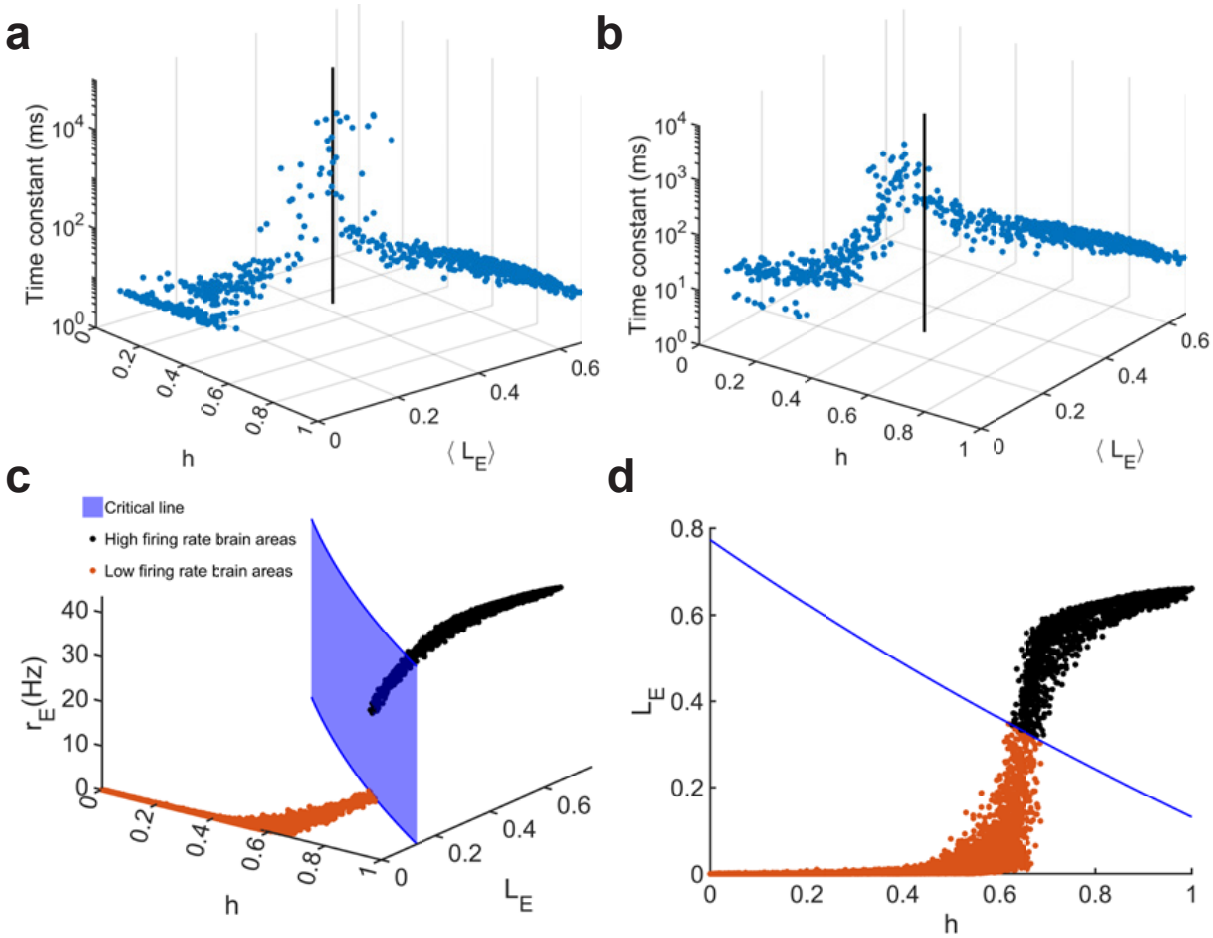

Figure S13: **Time constant around cusp point and geometry of $r(h, L_E)$ surface of network with 10000 brain areas with threshold-linear transfer function.** (a-b) The time constant of all the brain areas arranged in the h and $\langle L_E \rangle$ space for $d = 0.17$ and $d = 0.157$ with 1000 brain areas, respectively. The $\langle L_E \rangle$ is the average long-range gating variable of 80 seconds. The noise strength of panels a and b is the same as that of panel i of Fig. 3 and panel b of 4, respectively. The black line marked out the h and L_E value of the estimated "cusp point." (c) The critical line and firing rate distribution in h and L_E space for the threshold-linear transfer function with 10000 brain areas. (d) the top view of panel c.

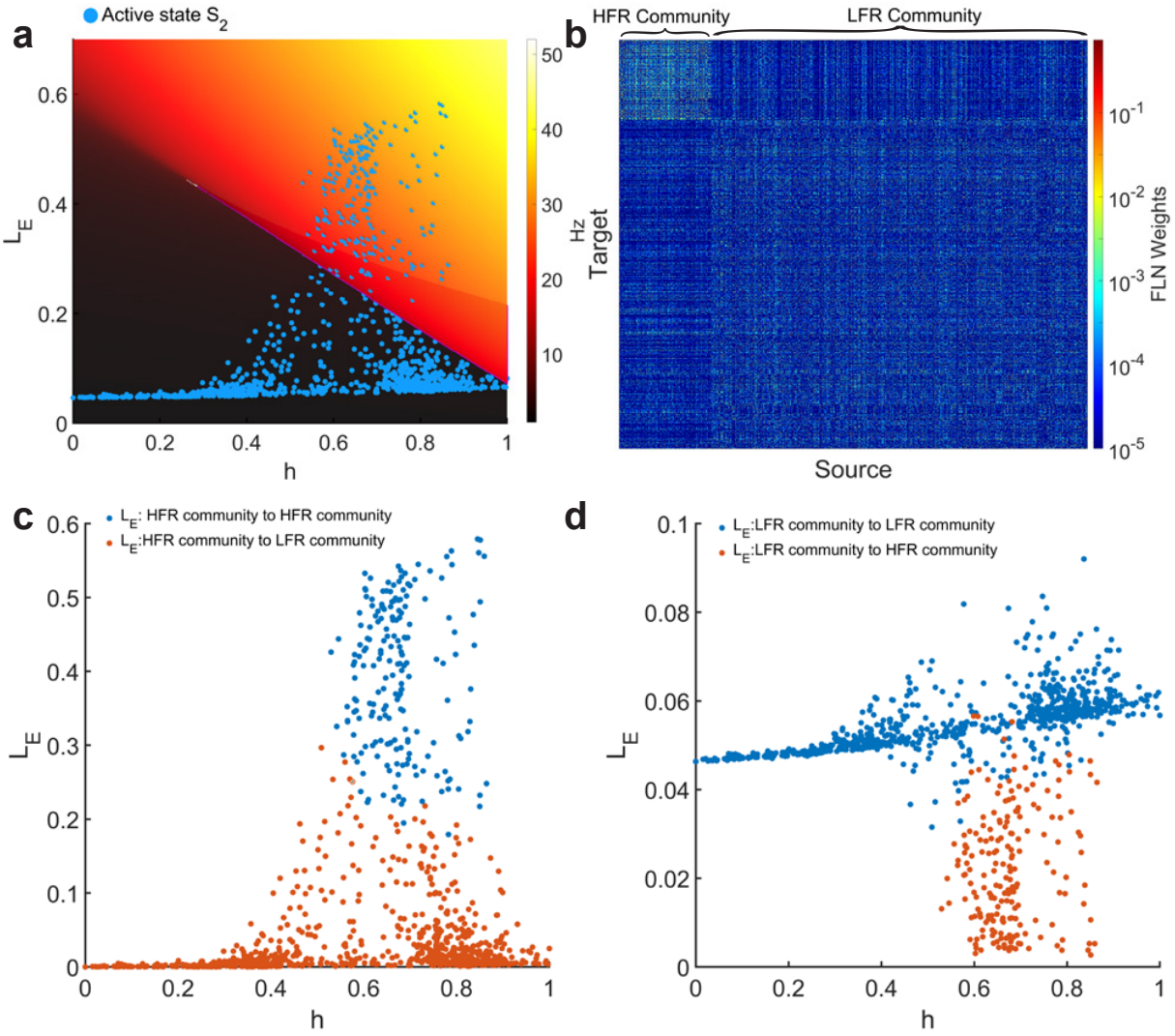

Figure S14: **The supplementary figure of L_E of the bump active state S_2.** In the bump state S_2, the weighted long-range excitatory input (L_E) is widely distributed within the hierarchical bump region. Brain areas receiving a sufficiently large L_E will exhibit a high firing rate, whereas those with a small L_E will only achieve a low firing rate state. The high firing rate (HFR) community also has a greater average connection strength than the low firing rate (LFR) community. Due to the community structure of the connections, the long-range excitatory input within the HFR community exceeds the input from the HFR community to the LFR community.

Continue with figure S14's caption. Conversely, the long-range excitatory input within the LFR community is greater than the input from the LFR community to the HFR community. This pattern is a result of the normalization of input FLN weights. (a) A top-down view of all brain areas on the $r(h, L_E)$ surface for the active state S_2 is illustrated in Fig. 4f. (b) The FLN matrix is organized by the HFR and LFR communities. Brain areas in S_2 with a firing rate exceeding 10 Hz are included in the HFR community, while the remaining areas are included in the LFR community. (c) The long-range excitatory input from the HFR community to brain areas within the HFR community is represented by blue dots, and the input from the HFR community to brain areas within the LFR community is depicted by brown dots. (d) Similarly, the long-range excitatory input from the LFR community to brain areas within the LFR community is shown by blue dots. In contrast, the input from the LFR community to brain areas within the HFR community is represented by brown dots.

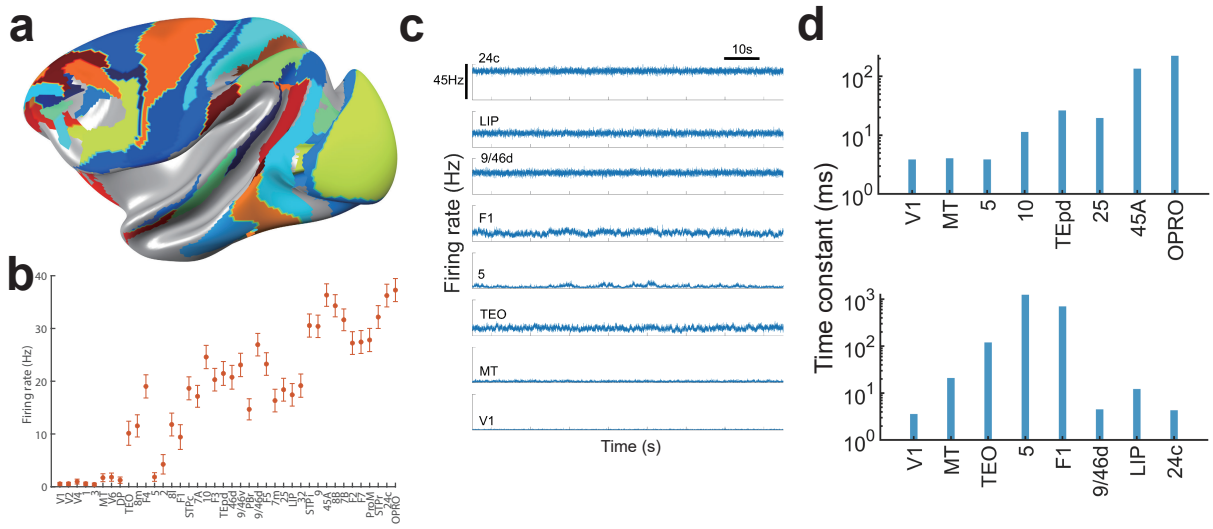

Figure S15: **Bifurcation in the space of the connectome-based cortical models of the macaque monkey (a-d)** **[20]** (a) Lateral view of macaque neocortex surface with 40 model areas in color. (b) Firing rate of 40 brain areas, ranked by the hierarchical position. The brain areas' hierarchical position is the same as in Froudist-Walsh et al. [21]. (c) Firing rate time series of 8 chosen brain areas when neocortex model is in a delay period working memory state for the model in (b). (d) The bar figure of time constants of 8 selected brain areas for resting (upper panel) and delay period working memory (lower panel) state.