Bifurcation in space: Emergence of functional modularity in the neocortex

Xiao-Jing Wang[1,†,*], Junjie Jiang[1,2,†,*], Roxana Zeraati[3,4], Ulises Pereira-Obilinovic[1], Aldo Battista[1], Julien Vezoli[5], Henry Kennedy[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] University of Tübingen, Tübingen 72076, Germany

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

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

† These authors contributed equally to this work.

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

Abstract

How does functional modularity emerge in a cortex composed of repeats of a canonical local circuit? Focusing on distributed working memory, we show that a rigorous description of bifurcation in space describes the emergence of modularity. A connectome-based model of monkey cortex displays bifurcation in space during decision-making and working memory, demonstrating this new concept's generality. In a generative model and multi-regional cortex models of both macaque monkey and mouse, we found an inverted-V-shaped profile of neuronal timescales across the cortical hierarchy during working memory, providing an experimentally testable prediction of modularity. The cortex displays simultaneously many bifurcations in space, so that the corresponding modules could potentially subserve distinct internal mental processes. Therefore, a distributed process subserves the brain's functional specificity. We propose that bifurcation in space, resulting from connectivity and macroscopic gradients of neurobiological properties across the cortex, represents a fundamental principle for understanding the brain's modular organization.

Introduction

Recent technical advances are enabling neuroscientists to image calcium signals or record spiking activities of large populations of single cells in behaving animals [1, 2], opening a new era for the investigation of distributed neural computation across the multi-regional brain [3]. These studies report widespread neural correlates of task-relevant information and observed widespread activity signals that have been interpreted as evidence of a lack of spatial specificity. Therefore, a central challenge in the field is to elucidate the local versus global neural processes underlying behavior. We tackled this challenge using computational modeling of the multi-regional cortex. In psychology, modularity denotes an organization of the mind into distinct component capabilities [4]; in neuroscience, it refers to functional specialization of brain areas. 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 such as face representation [5, 6], language, music or theory of mind. This definition is compatible with distributed neural representation and processing across multiple brain regions, and stands in contrast to the absence of modularity manifested by merely graded variations of engagement across the entire cortical mantle.

According to a central tenet of neuroscience, a canonical local circuit is repeated numerous times throughout the cortical mantle and shared across mammalian species [7]. In line with this view, the cortex is commonly described as a graph where parcellated areas are considered as identical nodes, each with distinct inputs and outputs that determine its function [8]. However, input-output patterns alone do not explain a variety of qualitatively functional abilities in different parts of the cortex, exemplified by the contrast between the primary sensory areas and the prefrontal cortex [9, 10, 11]. For concreteness, consider working memory, our brain's ability to maintain and manipulate information

internally without external stimulation [12, 13]. Working memory represents an excellent case study because it is essential for major cognitive processes and has been extensively investigated. The underlying neural mechanism of this core cognitive function involves persistent neural firing that is self-sustained internally during a time delay between stimulus and response [14, 15, 16]. A large body of literature has documented that working memory representations are distributed over some cortical areas but not others [17, 18]. In particular, with regards to 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) does [19]. These observations suggest a sharp onset of working memory representation along the cortical hierarchy. How can such a functional modularity be reconciled with the uniform canonical architecture of the cortex?

Recent experimental and computational research suggests clues to solve this major puzzle. Heterogeneities in different parts of the cortex [20, 21, 22] have been quantified; they are not random but display macroscopic gradients along low-dimensional axes such as the anatomically defined cortical hierarchy [23]. Such macroscopic gradients have been incorporated into connectome-based models of the multi-regional cortex of macaque monkey [24, 25] and mouse [26] for distributed working memory. In these models, the idea of a canonical local circuit is implemented by the mathematical equations of an excitatory-inhibitory neural network in each parcellated area; variations of the strength of synaptic excitation or/and inhibition are incorporated in the form of the macroscopic gradients. Computational modeling 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 led us to speculate that the sharp transition is "akin to a bifurcation in space" [25].

Motivated by hints from the previous work, the present work is designed to test that hypothesis in order to establish "bifurcation in space" rigorously and explore its experimental tests. The mathematical term "bifurcation" denotes the sudden onset of a qualitatively novel behavior under a graded change of the properties of a dynamical system [27]. The idea of bifurcation *in space* is conceptually novel; while it emerges from an interactive large-scale brain circuit as a collective phenomenon, it nevertheless occurs locally in space. We used a generative model of the cortex [28], that can generate an arbitrary number of areas characterized by the experimentally measured connection statistics of the cortex [29, 30]. This approach enabled us to derive a normal form equation close to the bifurcation [27]. The analytical prediction fits well with numerical simulation results, thereby formally establishing the concept of bifurcation in space.

Similar to the phase transition of ice melting in physics, we observed that the timescale of neural activity approaches infinity at the transition point—a phenomenon known as "critical slowing down" [31, 32]. Consequently, the time constants of neural firing fluctuations are predicted to be maximal for cortical areas near the transition, larger than those in both areas lower in the hierarchy that are devoid of persistent activity and areas higher in the hierarchy that exhibit robust persistent activity. In other words, along the cortical hierarchy, an inverted-V-shaped pattern of time constants is predicted to dominate neural fluctuations during the persistent activity that is associated with working memory. Note that previously we showed critical slowing down for a dynamical system as a whole, while here, we show that it manifests locally at a particular location in a spatially extended system.

Furthermore, we considered a connectome-based model of macaque cortex [25, 24] extended to include MST with hitherto unpublished connectivity data. The model was used to simulate a classic perceptual decision-making task in both the reaction time version

and the fixed duration version that requires working memory as well [33, 34]. We found that a subset of areas (a module) underlies subjective decision, which is differentiable from veridical evidence in error trials; a bifurcation in space unfolds in time, underlying dynamical and distributed decision-making. These results broaden our findings beyond working memory. Both this macaque monkey cortex model and a connectome-based mouse cortex model exhibit an inverted-V-shaped profile of time constants during working memory, identifying a specific model prediction that is testable experimentally across model species.

These results are highly non-trivial. First, the 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, about 66% of all possible inter-areal pathways are present [30], however, bifurcation occurs locally in space. Third, a bifurcation in space does not require fine-tuning of a parameter, unlike ice that melts when the temperature is adjusted to precisely zero degrees Celsius. Fourth, bifurcation in space so defined can apply to any one of the numerous spatially distributed persistent activity states. In other words, a cortical system with fixed parameters displays simultaneously many bifurcations in space, each engaging a subset of cortical areas constituting a specialized module that could potentially subserve distinct internally driven brain functions.

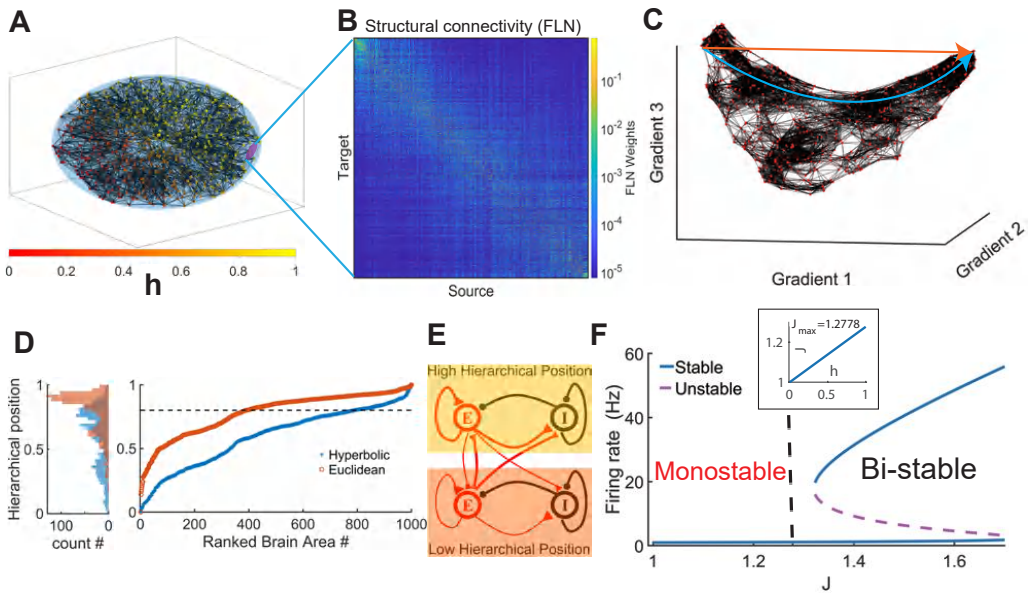

Figure 1: Spatially embedded generative model of a cortex. (A) Network connectivity of 1000 parcellated cortical areas produced by a generative model of the mammalian cortical connectivity [28], which is spatially dependent, directed, and weighted. (B) Connectivity matrix of the model. The connectivity weights are by the fraction of labeled neurons (FLN) in the connectomic analysis of the macaque cortex [30]. (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 three first 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. (5) in SI). In the large-scale system, the local E-to-E and E-to-I weights are scaled with the cortical heterogeneity factor J, which displays a macroscopic gradient as a function of the hierarchical position h (Insert), with the maximal value J (vertical dashed line) below the threshold.

Results

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 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 [28], unlike a purely topographic graph.

This model captures the central aspects of the mammalian neocortical connectivity and neural dynamics but is simple enough to be suitable for mathematical analysis. The model (Fig. 1A) is both generative and random, and thus can be used to produce multiple *realizations* of a mammalian neocortical network model with sample connectivity matrices (see Fig. 1B for one network realization) that share the same statistical distributions empirically observed in the inter-areal connectivity [29, 30, 35, 28]. This network model has three advantages. First, conclusions rendered from this model can be applied to different mammalian species [30, 36, 37, 38]. 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 show that all our results hold up in connectome-based models of both macaque [25] and mouse cortex [26].

We define the hierarchical distance using the diffusion map embedding method [39, 40] applied to our model. Note that the aim was to introduce a macroscopic gradient of

excitation in the abstract model, other methods could achieve the same purpose. 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. 1C). 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. 1D 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. 1D 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. 1A), 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. S1A). 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. 1E and Methods). Consistent with the macroscopic gradient of excitation observed in the cortex [23], the local and long-range excitatory weights are scaled by the hierarchical location

(i.e., $J \propto h$, insert in Fig. 1F). 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 at the threshold of $J > J_{\text{threshold}}$. To focus on collective large-scale dynamics, here we consider the case when the maximal value of J at the top of the hierarchy is smaller than $J_{\text{threshold}}$ (Fig. 1F) so that the observed distributed working memory representation emerges from long-distance area-to-area connection loops, thereby extending the concept of synaptic reverberation [41, 14, 16] to the large-scale multi-regional brain.

Bifurcation in hierarchical space

In contrast to isolated areas, 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. 2A, 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. S2B). We expect this transition zone to shrink to a point in the limit of infinitely large networks. We denote this point in hierarchical space as the bifurcation location (Fig. 2A, Fig. S2B). Interestingly, at the bifurcation point, there is a firing rate gap reminiscent of classical first-order phase transitions in statistical physics [42].

Mapped into our generative model's ellipsoid where connectivity was originally em-

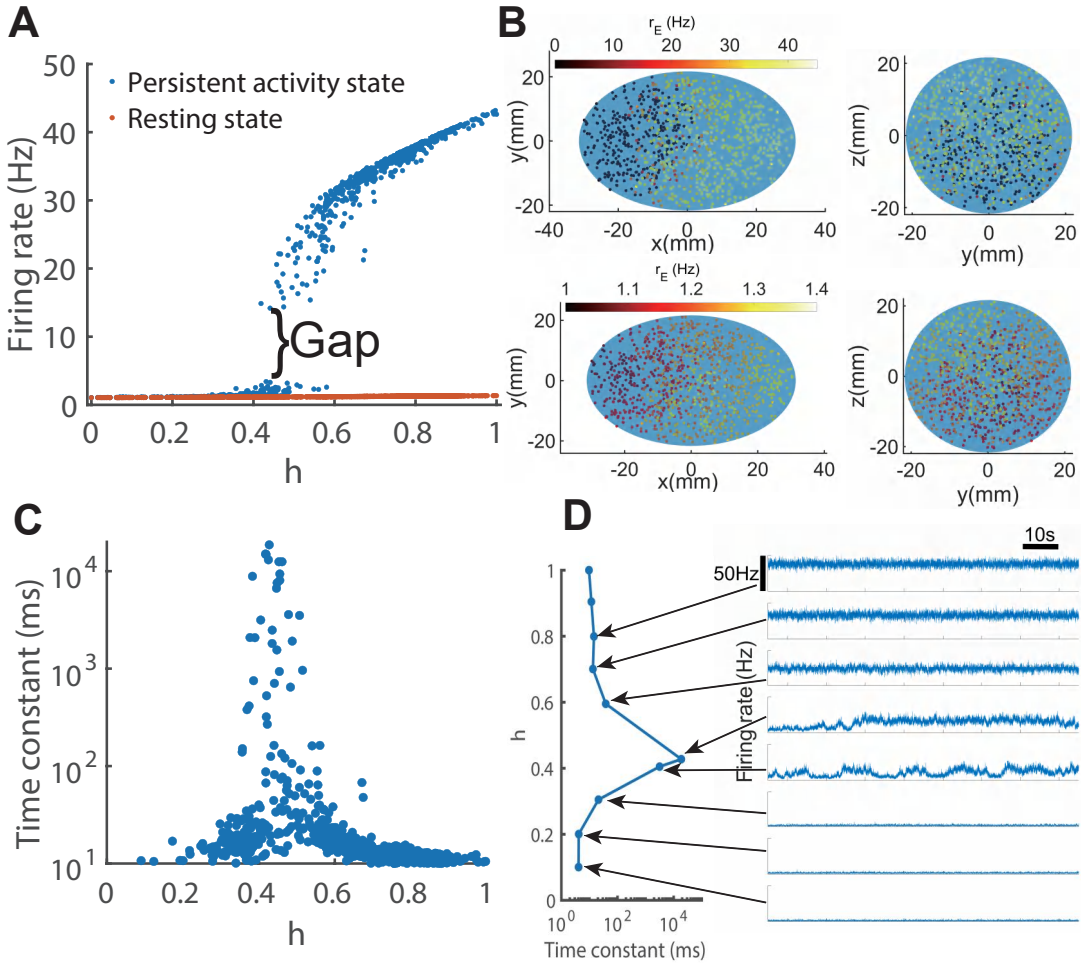

Figure 2: Bifurcation in hierarchical space. (A) Resting (brown) and persistent activity states (blue) are shown with firing rates plotted against areas ranked by hierarchical position. In the persistent activity state, a subset of areas represent working memory and are separated from the rest of the network by a firing rate gap. (B) 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. (C) The time constant of all brain areas for the persistent activity state of panel A with $1,000$ brain areas. (D) 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.

bedded, the firing rate of both active and resting states increases along its major axis (Fig. 2B). In contrast to the resting state, where the firing rate increases smoothly along the hierarchy (Fig. 2B, lower panel), in the persistent activity state, there is a firing rate gap between the rostral and the caudal areas (Fig. 2B, upper panel). Furthermore, the persistent firing rate increases along the minor axis z (Fig. 2B).

A signature of a "phase transition" in physical systems is critical slowing down, which denotes the phenomenon of fluctuations on all timescales (scale-free) close to a critical point, such as ice-water transition at zero degrees Celsius. To investigate whether our network displays critical slowing down associated with the bifurcation in space, we calculated the autocorrelation function of stochastic neural activity in the mnemonic working memory state, from which a timescale is extracted (see Methods, Section "Auto-correlation function of excitatory firing rate and estimated time scales").

The time scale of neural fluctuations increases from milliseconds to tens of seconds for areas near the transition point (Fig. 2C), 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. S2H), reproducing the result in the resting state from the linear theory of connectome-based large-scale models [43, 44] (Fig. S2H, left panel).

In the model, when the input-output neuronal transfer function's gain (parameter d) decreases, the firing rate gap disappears in the persistent activity state. In this scenario, the system is divided into two-components: activity is roughly constant and low for areas low in the hierarchy, then starts to increase without a discrete jump of firing rate for areas

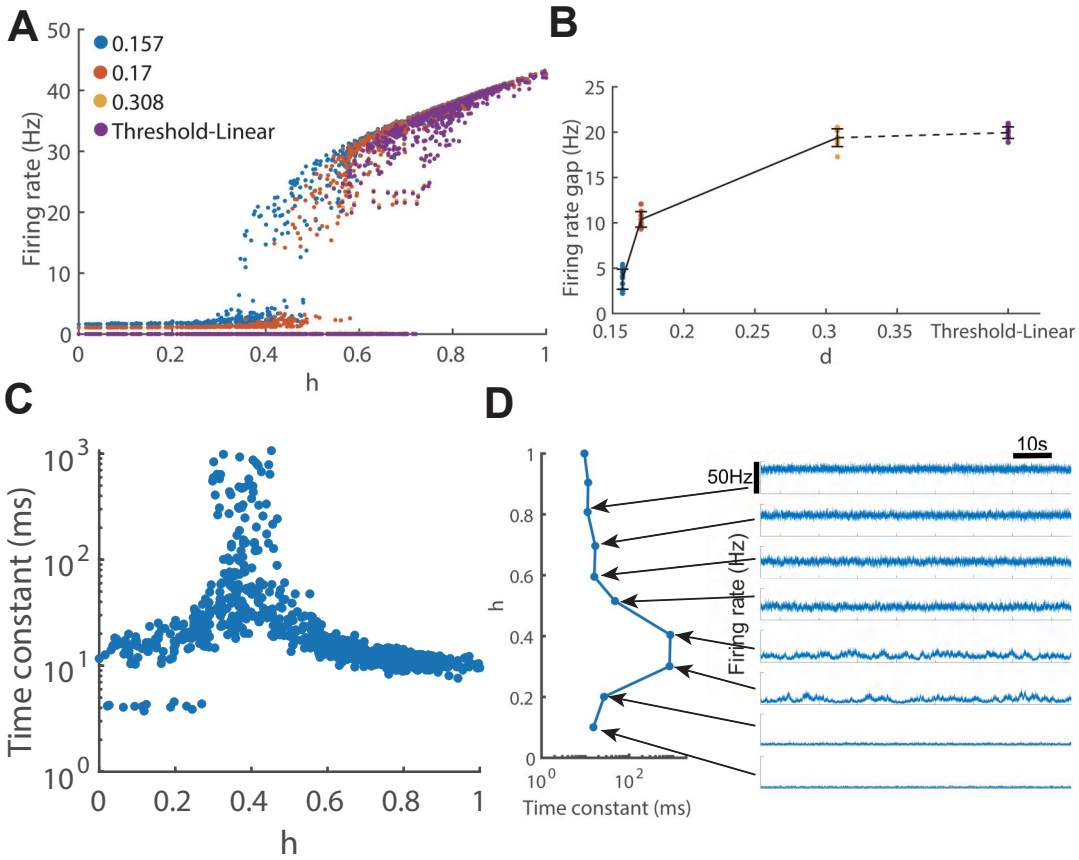

Figure 3: The effects of the input-output transfer function gain parameter d on bifurcation in space. (A) The persistent activity state for different d values. As the gain parameter d increases from 0.157 to ∞, the firing rate gap progressively widens from zero. (B) The maximum firing rate difference among all area-pairs. Means (dots) and error bars correspond to the results of 10 different network realizations. (C-D) Model behavior with the gain parameter $d = 0.157$. (C) The time constant of neural fluctuations in all the cortical areas during the persistent active state. (D) The time constant (left) and firing rate time series (right) of 8 selected brain areas in the active state with the smooth transition.

higher in the hierarchy (see Fig. 3A and Fig. 3B, blue dots). As the input-output gain increases, the firing rate gap becomes progressively large (see Fig. 3A) until the bifurcation in space becomes apparent. The firing rate gap at the transition point increases with the gain d reaching its maximum for the threshold-linear transfer function (see Methods, Eq. (6)) [45] (Fig. 3A-B). With a sufficiently small d value, the transition becomes smooth with virtually no firing rate gap, but the working memory state nevertheless continues to be characterized by an inverted-V-shaped time scale profile (Fig. 3C-D), suggesting that critical slowing down does not require the presence of a firing gap in the persistent activity state.

The normal form analysis of bifurcation in space

To elucidate rigorously the transition observed in Figs. 2-3 as a new form of bifurcation, we undertook mathematical analysis in a simplified instance of threshold-linear input-output transfer function (corresponding to a large gain d in our model). Close to a bifurcation, the model can be approximately reduced to a "normal form" allowing its behavior to be described as we show below [46].

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. 4A). For a given active state, the input current for each area i depends on its input connections (the FLN_{ij} values) and the collective activity state. Indeed, the firing rates are driven by L_E, 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. (7)), the firing rate as a function of h and L_E can be

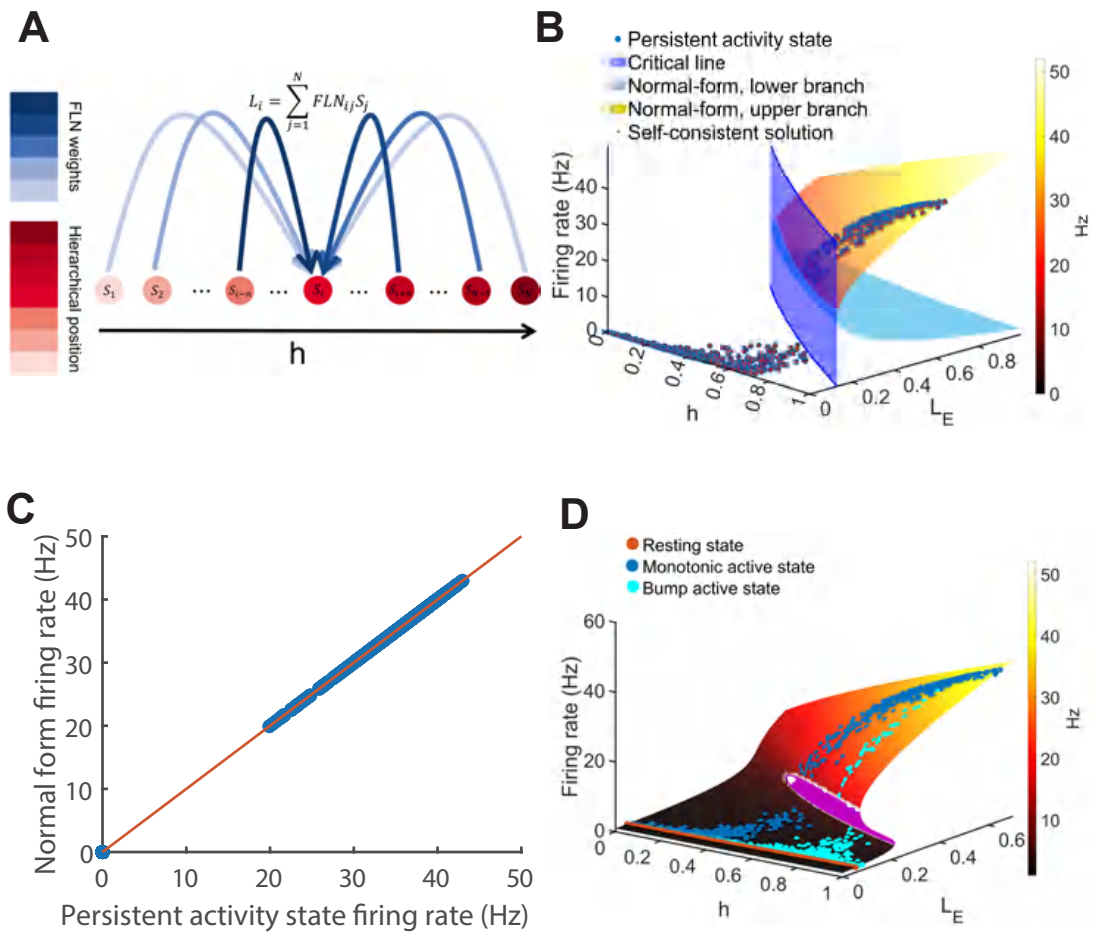

Figure 4: The geometry of distributed attractor states. (A) Illustration of long-range excitatory input currents to an area i^{th}. The gradient of red and blue color corresponds to hierarchical positions of areas and the weight of long-distance connections FLN_{ij} from area j to area i, respectively. (B) Normal form analysis of the neocortex model with threshold-linear transfer function. The bifurcation in hierarchical space occurs at the critical line in the plane of hierarchy h and long-range gating variable L_E. (C) The firing rate from network simulations versus the predicted firing rate from the normal form analysis show perfect agreement. (D) Cumulative firing rates of all the areas of the resting state, monotonic persistent activity states, and bump persistent activity states for the generative model.

readily solved, yielding a bistability surface as a function of h and L_E (see Fig. 4B). At this limit, the normal form of bifurcation in hierarchical space is obtained (see Methods, Section "The bifurcation in hierarchy space normal form", in particular, Eq. (29)). Using that 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. 4C).

As the case of finite gain parameter d, we found that for any steady state of the network, the firing rate of all cortical areas must lay on a surface parameterized by the hierarchy h and the long-range excitatory input current L_E (Fig. 4D, Fig. S4, Fig. S5, and Fig. S6). We refer to this surface as the solution surface (Fig. 4D, Fig. S4, Fig. S5, and Fig. S6). The solution surface does not depend on the network size N. In the resting state, both firing rates and long-range excitatory inputs L_E are low (Fig. 4D). In a persistent activity state (Fig. 2A 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 strongly interconnected areas nearby in the hierarchy [28, 29] (Fig. 4D). In the same active state, the long-range excitatory input currents L_E^i are large for areas above the transition since they receive strong inputs from nearby areas with elevated persistent activity.

What is the geometry of the solution surface in our network? We find that the solution surface folds at a *cusp*, corresponding to a point in the two-dimensional space of h and L_E (Fig. S6). This geometry is reminiscent of the cusp described in classical bifurcation theory for non-linear dynamical systems [47, 46]. For a system that undergoes a cusp bifurcation, the solution surface representing the steady state solution in a two-dimensional parameter space folds at a cusp point [46]. 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 long-range

excitatory input currents L_E within the folded region exhibit bistability (see Methods and Fig. 4D). 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. 2A. 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 is occurring 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 becomes absent as all areas participate in the persistent activity state (Fig. 5A, 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. 5B) or the resting state (Fig. 5C). 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. 5D-F). These results demonstrate that the desired functional modularity depends critically on both the connectomic properties and the macroscopic gradient.

A diversity of bifurcations in space

Interestingly, we discovered numerous persistent activity states in our network. All the states exhibiting distinct spatial distributions coexist in a single realization of the generative network, similar to findings in connectome-based models of macaque monkey [25] and mouse [26]. For most of these attractor states, firing rate increases monotonically across the hierarchy (see Fig. 2A, Fig 6A, Fig. S8A). In addition, surprisingly, a sizable

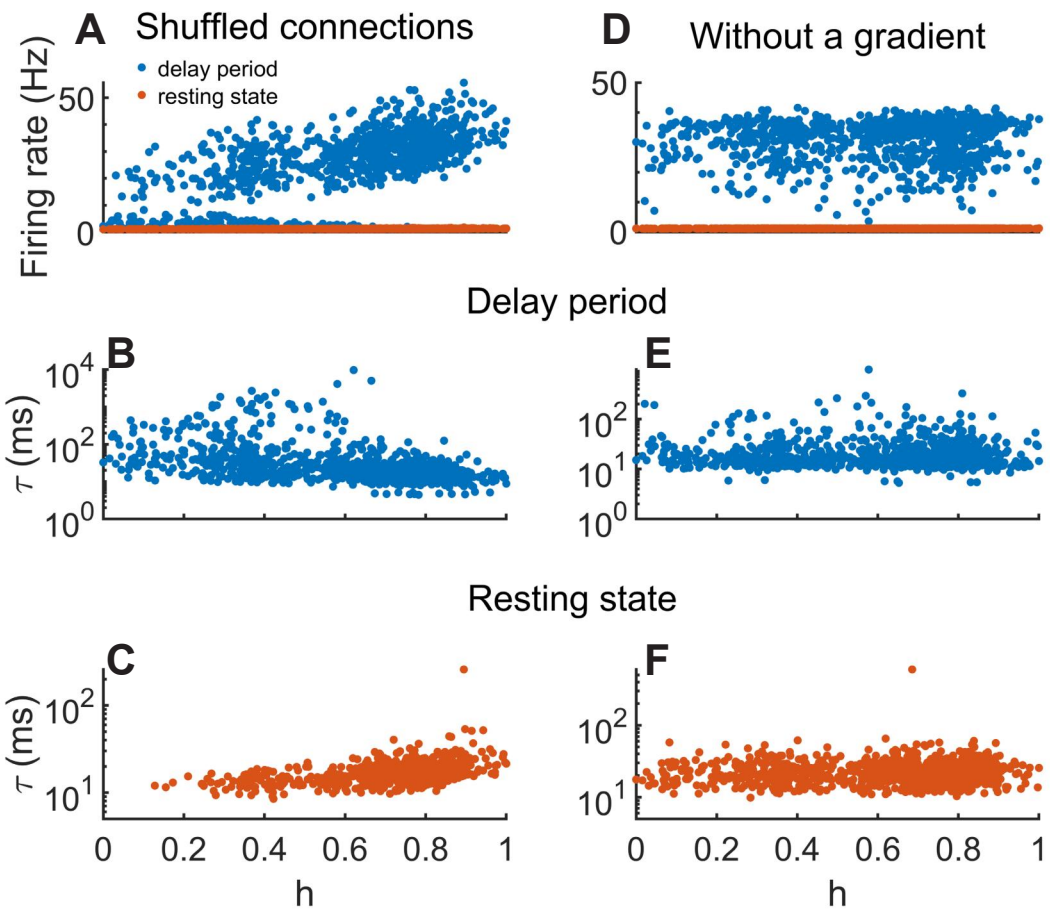

Figure 5: The firing rates and time constants of distributed attractor states when the network connection weights are randomly shuffled (A-C), or the macroscopic gradient of excitation is abolished by randomly shuffling the parameter J (D-F). In either case, the model still displays the bistability of 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.

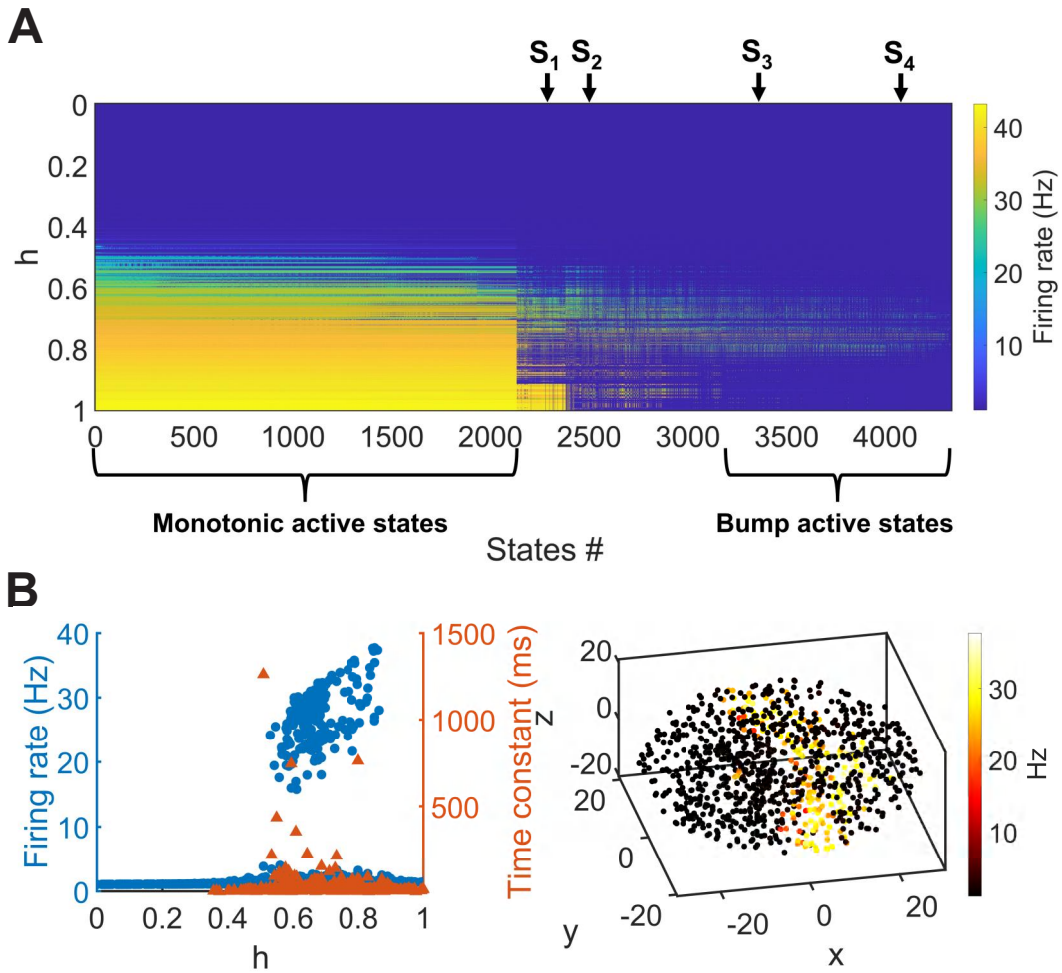

Figure 6: A diversity of spatially distributed attractor states. (A) The firing rate patterns of 4333 active states with cortical areas ranked by hierarchy values. Distributed working memory states display either a monotonic pattern or a bump pattern with elevated activity only in the middle region of 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. All spatially distributed persistent states coexist in a single realization of the generative network model, which is the same as Fig. 2. (B) An example of a bump-shaped persistent activity state indicated by S_2 in (A). Left: firing rate (blue) and time constant (brown) as a function of the hierarchical position; Right: spatial distribution of firing rates in the generative model ellipsoid.

number of attractor states display a localized *bump* of activity in hierarchy space (Fig. 6A, bump active states; Fig. 6B; and SI Fig. S8C). 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. 6B and Fig. S8C, 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. 6B and Fig. S8C). Note that, here, the focus is on the spatial attractors independently from the number of selective neural populations in each area. The question of the total number of such spatial attractors is fundamentally different from the conventional memory capacity analysis of the maximal number of stored stimulus-selective memory items. To illustrate this point, consider a simple and heuristic example of N areas that are not interconnected, each with two (Down and Up) stable states. Since each area is either Down or Up, the total number of spatial attractors is 2^N. The actual number of spatial attractors depends on the inter-areal connectivity as well as local area properties.

With a given network parameter set, all distributed persistent activity states can be plotted on the solution surface, even for different network realizations or networks with different sizes (Fig. 4D). This provides a unifying picture; different states take over different locations on the solution surface reflecting that cortical areas at a given hierarchical position h have different firing rates and long-range excitatory input current L_E values in distinct active states.

In summary, bifurcation in space is defined separately for each set of the active states 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 module) engaged in the corresponding persistent activity state and manifested by critical slowing down at its transition spatial location.

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

As shown above, using a generative model enabled us to establish the concept of bifurcation in space mathematically. We sought to assess this phenomenon in connectome-based multi-regional cortex models, as a necessary step to bridge theory with experimental tests. We expanded a macaque cortex model [25, 24] to 41 parcellated areas (Fig. 8A). The model incorporates hitherto unpublished connectivity data on the medial superior temporal (MST) area, specifically motivated by the physiological observations that neurons in MST, but not their main monosynaptic afferents from the neighboring medial temporal (MT) area, are known to display delay period activity in a working memory task, suggesting a sharp transition [19]. In this model, suitable for stimulus-selective decision-making and working memory, each brain area possesses two excitatory populations encoding stimuli and one inhibitory population [25]. Both long-range and local recurrent excitation strength follows a macroscopic gradient proportional to the hierarchical position of each brain area [43, 48, 23].

To explore whether the novel concept of bifurcation in space is applicable beyond working memory, 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 [33, 34]. 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 [49, 50]. 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 [33], 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) ramped up to a preset level [33, 51, 52], but see the Discussion for more general decision readout from a multi-regional system). 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. This task thus engages decision-making and working memory. The psychometric functions of the model for the RT and FD tasks (Fig. 7A) are comparable to the observed monkey's 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 threshold value (see Methods). The average RT as a function of the coherence is shown in Fig. 7B, with longer RTs in error (open circle) than correct (filled circle) trials, in accordance with the observed behavior.

Fig. 7D-E (Movies 1, 2) shows 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 is larger to the neural population A than

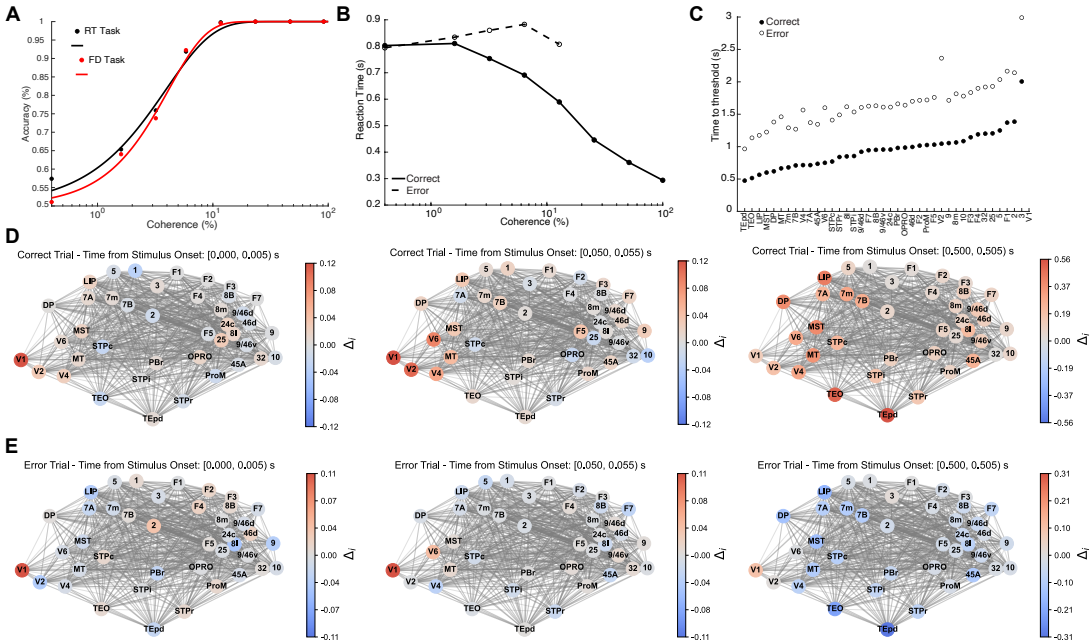

Figure 7: Connectome-based model of macaque monkey cortex capable of perceptual decision making and working memory. (A) Psychometric function (accuracy) for the RT and FD tasks and (B) chronometric function (average reaction times in correct and error trials) of the model, compatible with monkey behavior in motion direction discrimination experiments [52]. (C) Time-to-threshold for all brain areas. (D, E) Activity maps across different brain areas exhibit distinct structures in correct (D) and error (E) trials depending on the time from the stimulus onset (indicated in panel titles). Red indicates areas (indexed by i) following the objective choice ($r^A_{E,i} > r^B_{E,i}$ when input to excitatory population A is larger, $\Delta_i > 0$), and blue indicates a wrong subjective choice ($r^A_{E,i} < r^B_{E,i}$, $\Delta_i < 0$). 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 [30]. Line widths correspond to the relative connectivity strength (FLN) between different brain areas. For the temporal dynamics of the activity maps, see Movie 1 (correct trials) and Movie 2 (error trials).

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; areas TEpd, TEO, and LIP are the first to make the choice in the correct (Fig. 7D, Movie 1) and error (Fig. 7E, Movie 2) trials. Compared to the correct trial, in the error trial, there is a period (Fig. 7E, 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 (Fig. 7E, right panel). MT also reflects the subjective choice, reproducing the observed choice-correlate in MT neurons of behaving monkeys [53], which has been suggested to result from top-down inputs [54] in accordance with our connectome-based large-scale cortex model.

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

The FD task allowed us to examine the working memory of a choice across a delay period. The model's performance in the FD task is similar to that of the RT task (Fig. 7A). As in our abstract model, this connectome-based model exhibits the coexistence of a resting state (firing rate around 1.5 Hz) and persistent activity state (more than 8 Hz) encoding working memory (Fig. 8B). The spatial firing rate distribution during working memory states is modular, with only some areas displaying persistent firing (Fig. S10E). Consistent with the 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. 8B). The MST area exhibits sustained activity throughout the delay period, whereas the MT area does not manifest such persistent ac-

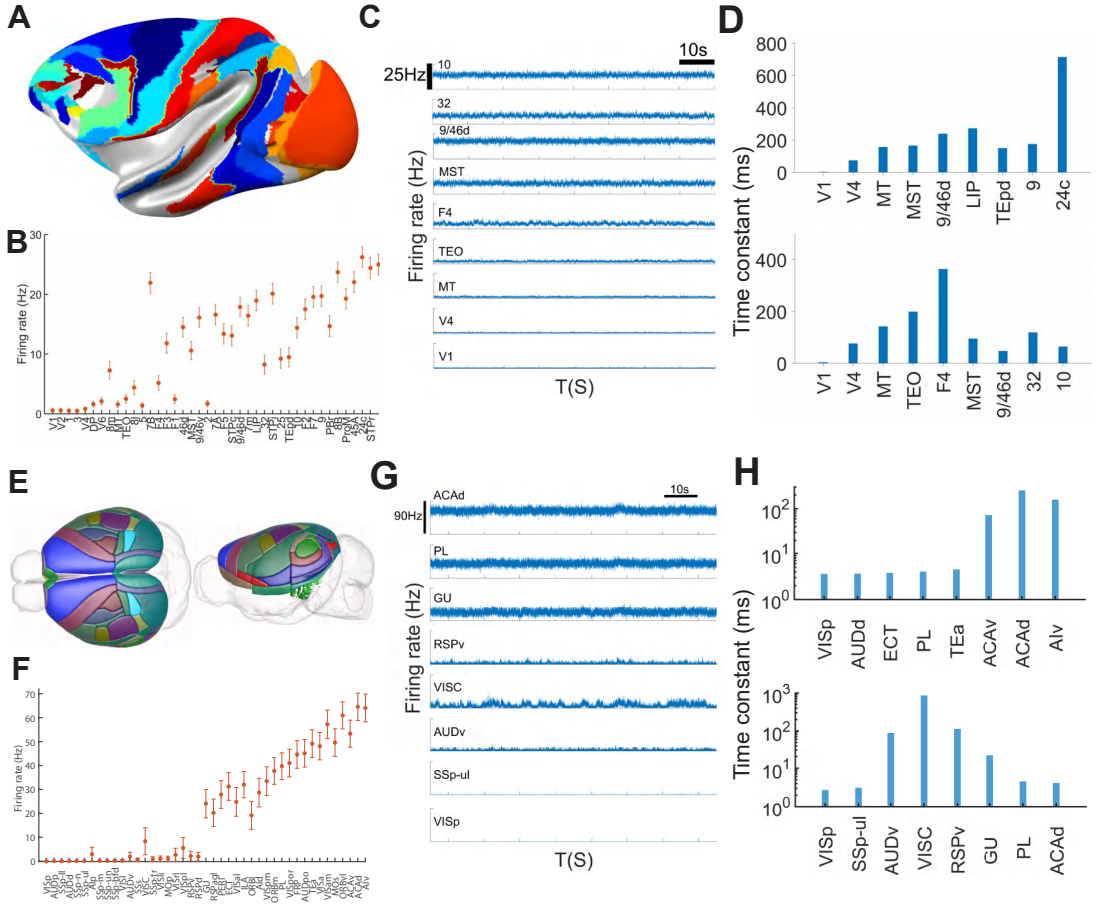

Figure 8: Bifurcation in space of connectome-based cortical models of the macaque monkey (A-D) [25] and mouse (E-H) [26]. (A) Lateral view of macaque neocortex surface with model areas in color. (B) Firing rate of 41 brain areas, ranked by the hierarchical position. (C) Firing rate time series of 9 chosen brain areas when the neocortex model is in the delay period working memory state. (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. (G) Firing rate time series of 8 sample areas in the delay period working memory state. (H) Time constants of neural fluctuations in 8 selected areas in resting (upper panel) and delay period working memory (bottom panel) states.

tivity during this period, as evidenced in Fig. 8B. This observation is in accordance with the experimental findings reported by Mendoza-Halliday et al. (2014) [19]. Furthermore, our analysis identifies a single parameter set that aligns with the experimental data and delineates a broader parameter space within the J_s^{max} and z parameters (Fig. S11A). The standard deviation of the firing rate shows a monotonic increase pattern as a function of the hierarchy (Fig. S10A).

Note that the model parameters suitable for working memory and decision-making are not necessarily identical. In fact, a systematic analysis showed that 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. S11B and 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.

In order to compare the time constant of each brain area during a baseline state with the previous experimental result, we observed that the time constant of the resting state is roughly a monotonic function of the hierarchy (Fig. 8D, upper panel, see also Fig. S10F, left), in accordance with previous modeling [43] and empirical [55] findings. Note that for a nonlinear dynamical system, the time constants are not uniquely defined; they are specific for a given internal state and predicted to differ between the resting and working memory states.

To test for the presence of critical slowing down in the connectome-based large-scale model, we carried out the autocorrelation analysis of stochastic persistent activity time series during the delay period for each area. We found that neuronal fluctuations are fast in the brain areas at high and low hierarchical positions; by contrast, the mnemonic firing of brain areas around the hierarchical bifurcation region shows fluctuations on slower timescales, exemplified by Brodmann area F4, which is part of the premotor cortex

(Fig. 8D, lower panel; see also Fig. S10F, right). The brain areas F4, MT, TEO, and 32 have a time constant larger than 10^2 ms.

We then asked if critical slowing down also occurs in the connectome-based large-scale model of the mouse cortex [26]. The model contains 43 cortical areas in the common coordinate framework v3 atlas [56] (Fig. 8E) with a quantified hierarchy. Synaptic inhibition mediated by parvalbumin-expressing interneurons [57] 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. 8F). 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. 8H, upper panel) and a qualitatively similar inverted-V-shaped profile of time constants in the delay period working memory state (Fig. 8H, lower panel, see also Fig. S10, H right).

Our model of the macaque monkey cortex is presently limited to a subset of areas for which the connectomic data are presently available; the precise hierarchical positions (normalized between 0 and 1) could be somewhat modified in a complete model of all cortical areas. Moreover, the particular area that displays the maximal time constant of mnemonic firing fluctuations may depend on the model parameters. Regardless, the demonstration of the inverted-V-shaped pattern of time constants in the mouse and macaque monkey cortical models offers a strong model prediction that is testable experimentally.

Discussion

Motivated by the experimental advances, theoretical modeling of a multi-regional cortex has come to the fore [58, 3]. Up to now, a focus of research has been on the functional

connectivity of the human cortex in the resting state when subjects are not engaged in tasks [59, 60, 61]. More recently, connectome-based models have been developed for distributed working memory [24, 25], which are now beginning to be applied to perceptual decision-making. The model of [25] with 30 areas was applied to perceptual decisions in the random dots direction discrimination task, which showed different global dynamics in easy versus difficult task conditions but did not address the question of modularity. Our work is distinct from previous work in four ways. First, we rigorously demonstrated the novel concept of bifurcation in space; its general applicability is reflected by the simultaneous presence of different spatial attractors and their corresponding bifurcations in space. Second, our connectome-based model of the macaque cortex reproduces the experimental observation of a sharp transition between the monosynaptically connected MT and MST areas [19], validating this theoretical work. Third, we found an inverted-V profile of time constants during working memory, a surprising and specific prediction that can be tested experimentally. Fourth, we showed 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 [62]. Further studies are needed to elucidate the details of distributed decision-making, in close interplay with experiments such as the recently carried out study in mouse [63]. In particular, how exactly a decision is read out in a multi-regional brain remains an open question.

To investigate how functional modularity emerges in a multi-regional 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 have been rare reports of delay period activity in V1 of behaving monkeys [18], which, however, may correspond to attention signals

needed to perform certain tasks; most experiments failed to find evidence of working memory representation in V1. By contrast, ample evidence shows that a subset of cortical areas are involved in working memory maintenance [17]. In the human literature, fMRI experiments showed that visual working memory content can be decoded from functional MRI measurements in the primary visual cortex [64, 65], but whether sensory areas are critical for working memory storage has been questioned [66]. However, brain imaging data is compatible with modularity because BOLD signals reflect synaptic currents rather than neural spiking activity [67]. In our model, during a mnemonic delay, primary sensory areas receive strong synaptic top-down inputs that target both excitatory and inhibitory cells but do not necessarily give rise to persistent activity.

The present work addresses how functional modularity, *if* present, 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 some critical location in the spatially embedded cortex. The idea of a neural system operating near to criticality has been proposed [68, 69]; open questions include identifying the signatures of criticality and determining whether achieving criticality requires parameter fine-tuning or can instead arise through a self-organized mechanism [70]? A bifurcation in space is robust: parameter changes would merely move the spatial location of bifurcation. Moreover, bifurcation is defined for each of numerous spatially extended persistent activity states that potentially can 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. In other words, there are numerous bifurcations in space in a

given large-scale cortical system.

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 [43, 3]. The difference is explained by the fact that time constants are uniquely defined mathematically only for a linear dynamical system but for a highly nonlinear system, they critically depend on the active state under consideration. 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.

The main results using a generative model of the cortex endowed with experimentally measured connection statistics are confirmed in connectome-based models of macaque and mouse cortices, opening the door to test the predicted inverted-V-shaped profile of time constants in specific areas during working memory. In particular, for working memory of visual motion information, the work in [19] suggests MST as a candidate area close to a criticality. Furthermore, since working memory and decision-making are believed to share a common cortical substrate [71], the inverted-V-shaped timescale profile is likely to hold during decision processes, a proposal in line with the existing evidence that neurons in the caudal parietal cortex display longer integration times underlying accumulation of information than is observed in sensory areas and prefrontal cortex, located lower and higher in hierarchical positions, respectively [72]. Testing this model prediction requires the following considerations. First, 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. 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, at the focus of this work merely for the sake of simplicity. Generally, attractors can display complex temporal dynamics such as chaos rather than steady states [27]. For instance, the neural representation of working memory often involves stochastic oscillations [73, 74] or intermittent dynamics [75]. As discussed elsewhere [16], the attractor paradigm can be consistent with considerable temporal variations of neuronal delay period activity and cell-to-cell heterogeneities. Future research is needed to extend the concept of bifurcation in space beyond steady states.

In this study, we used a model of the cortex, which does not interact with subcortical structures such as the thalamus. Previous work [26] 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, a more detailed investigation of its influence on large-scale dynamics could be pursued in future studies.

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. In contrast to local interactions through diffusion or chemical reactions, interareal cortical interactions involve long-range connections, making it all 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. 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.

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, the NSF Neuronex grant 2015276, and National Institutes of Health grant R01MH062349 (to X.-J.W.); Swartz Foundation postdoctoral 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.). 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, and was actively involved in all details throughout the work; 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 anatomical analysis of the MST connections; 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] Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).

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

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

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

[5] Kanwisher, N. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences* **107**, 11163–11170 (2010).

[6] Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).

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

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

[9] Fuster, J. M. *The Prefrontal Cortex* (Academic Press: New York, 2008), Fourth edn.

[10] Passingham, R. E. & Wise, S. P. *The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight* (Oxford, England: Oxford University Press, 2012).

[11] Wang, X.-J. The prefrontal cortex as a quintessential 'cognitive-type' neural circuit: Working memory and decision making. In Stuss, D. T. & Knight, R. T. (eds.) *Principles of Frontal Lobe Function*, 226–248 (New York: Cambridge University Press, 2013), second edn.

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

[13] D'Esposito, M. & Postle, B. R. The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).

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

[15] Amit, D. J. The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behav. Brain Sci.* **18**, 617–626 (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] von Economo, C. *The Cytoarchitectonics of the Human Cerebral Cortex* (London: Oxford University Press, 1929).

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

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

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

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

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

[26] 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).

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

[28] 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).

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

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

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

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

[33] 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).

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

[35] Wang, X.-J., Pereira, U., Rosa, M. G. & Kennedy, H. Brain connectomes come of age. *Current Opinion in Neurobiology* **65**, 152–161 (2020).

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

[37] 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).

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

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

[40] 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).

[41] Lorente de Nó, R. Vestibulo-ocular reflex arc. *Arch. Neurol. Psych.* **30**, 245–291 (1933).

[42] Landau, L. D. & Lifshitz, E. M. *Statistical Physics*, vol. 5 (Elsevier, 2013).

[43] 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).

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

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

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

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

[48] Demirtaş, M. *et al.* Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194 (2019).

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

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

[51] Huk, A. C. & Shadlen, M. N. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J Neurosci* **25**, 10420–10436 (2005).

[52] 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).

[53] 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).

[54] Wimmer, K. *et al.* Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area mt. *Nature communications* **6**, 6177 (2015).

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

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

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

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

[59] Izhikevich, E. M. & Edelman, G. M. Large-scale model of mammalian thalamocortical systems. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3593–3598 (2008).

[60] Deco, G. & Jirsa, V. K. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J. Neurosci.* **32**, 3366–3375 (2012).

[61] Demirtaş, M. *et al.* Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194 (2019).

[62] Klatzmann, U. *et al.* A dynamic bifurcation mechanism explains cortex-wide neural correlates of conscious access. *Cell Reports* in press (2024).

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

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

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

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

[67] 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).

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

[69] O'Byrne, J. & Jerbi, K. How critical is brain criticality? *Trends in Neurosciences* **45**, 820–837 (2022).

[70] Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality. *Physical Review A* **38**, 364–374 (1988).

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

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

[73] Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).

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

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

[76] Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102**, 7426–7431 (2005).

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

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

Materials and Methods

A generative model for the mammalian cortical connectivity

We use the model in [28] 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 [29]. 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 [28], 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 [39]. This class of non-linear dimensionality reduction has recently been applied to human and macaque monkey connectomes [40]. 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. 1C). We applied this method in the symmetric version of the FLN matrix $L = FLN + FLN^T$.

We closely followed the method described in [39]. First, we define the following normalized matrix

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

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

$$M_{ij} = \frac{L^{\alpha}_{ij}}{\sum_{k'=1}^{N} L_{ik'}}. \tag{2}$$

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^t_{ij}. To define the principal gradients, we perform the eigenvalue decomposition of M^t_{ij}, obtaining

$$M_{ij}^t = \sum_{l=1}^{N} \lambda_l^t \psi_i^l \phi_j^l, \tag{3}$$

Then, the principal gradients are the set of vectors

$$\left[\lambda_1^t \vec{\psi}^1, \cdots, \lambda_N^t \vec{\psi}^N\right]. \tag{4}$$

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 [76].

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. 1C): 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 $h_{Euc}^i = dist_{Euc}^{i0}/dist_{Euc}^{max}$. Here, the value $dist_{Euc}^{i0}$ is the Euclidean distance between brain area i and the origin area, and the value $dist_{Euc}^{max}$ 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. 1D.

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. 1D.

Isolated cortical circuit

The simplified nonlinear dynamical model is adopted from [50], which approximated spiking neural network with AMPA, GABA, and NMDA synapses [49]. The dynamical equations that describe the dynamics for a single cortical area are described as follows:

$$
\begin{aligned}
\tau_E \frac{dS_E}{dt} &= -S_E + \gamma_E \tau_E (1 - S_E) r_E, \\
\tau_I \frac{dS_I}{dt} &= -S_I + \gamma_I \tau_I r_I, \\
\tau_r \frac{dr_E}{dt} &= -r_E + \phi_{exc}(J W_{EE} S_E - W_{EI} S_I + I_{ext,E}), \\
\tau_r \frac{dr_I}{dt} &= -r_I + \phi_{inh}(J W_{IE} S_E - W_{II} S_I + I_{ext,I}),
\end{aligned} \tag{5}
$$

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 [45]

$$
\phi_{exc}(I) = \frac{aI - b}{1 - e^{-d(aI-b)}}. \tag{6}
$$

2. Threshold-linear function

$$
\phi_{exc}(I) = [aI - b]_+. \tag{7}
$$

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

$$\phi_{inh}(I) = [c_1 I - c_0]_+ \tag{8}$$

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

Dynamical model of the mammalian neocortex

We connect cortical areas with local neural dynamics described by equations (5-8) using the connectivity from our generative model of the mammalian neocortex. The long-range projections in our model are from excitatory to excitatory populations [43, 24, 25]. Our large-scale model is described as follows

$$\begin{aligned}
\tau_E \frac{dS_E^i}{dt} &= -S_E^i + \gamma_E \tau_E (1 - S_E^i) r_E^i, \\
\tau_I \frac{dS_I^i}{dt} &= -S_I^i + \gamma_I \tau_I r_I^i, \\
\tau_r \frac{dr_E^i}{dt} &= -r_E^i + \phi_{exc}(J^i(W_{EE} S_E^i + \mu_{EE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{EI} S_I^i + I_{noi}^i + I_{ext,E}^i), \\
\tau_r \frac{dr_I^i}{dt} &= -r_I^i + \phi_{inh}(J^i(W_{IE} S_E^i + \mu_{IE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{II} S_I^i + I_{ext,I}^i), \\
\tau_r \frac{dI_{noi}^i}{dt} &= -I_{noi}^i + \sqrt{\tau_r \sigma_{noi}^2} \xi^i,
\end{aligned} \tag{9}$$

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 insert of Fig. 1F). We assume that all the brain areas have the same external input current $I^i_{ext,E} = I_{ext,E}$ and $I^i_{ext,I} = I_{ext,I}$. The noise term I_{noi} is an Ornstein-Uhlenbeck process, representing the AMPA synaptic noise with a short time constant $\tau_r = 2ms$ [50]. 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:

$$\begin{aligned}
\frac{-S_E}{\tau_E} + \gamma_E(1 - S_E)r_E &= 0, \\
\frac{-S_I}{\tau_I} + \gamma_I r_I &= 0, \\
-r_E + \phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\
-r_I + \phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0.
\end{aligned} \tag{10}$$

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 $S_E \geq 0$. Therefore, we could reduce

our steady-state equation to

$$\begin{aligned} \frac{-S_E}{\tau_E} + \gamma_E(1-S_E)\phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\ \frac{-S_I}{\tau_I} + \gamma_I\phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0. \end{aligned} \tag{11}$$

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

$$\begin{aligned} S_I = \gamma_I\tau_I(c_1 I_{inh,t} - c_0) &= \alpha c_1 JW_{IE}S_E + \alpha(c_1 I_{ext,I} - c_0), \\ \alpha &= \frac{\gamma_I\tau_I}{1+\gamma_I\tau_I c_1 W_{II}} = \frac{1}{\frac{1}{\gamma_I\tau_I} + c_1 W_{II}}, \end{aligned} \tag{12}$$

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 (12) into the steady state S_E equation (11). After this manipulation, the steady state of the single cortical area is determined by the NMDA gating variable S_E as follows:

$$\begin{aligned} -S_E + \gamma_E\tau_E(1-S_E)\phi_{exc}(\alpha_1 S_E + \alpha_2) &= 0, \\ \alpha_1 &= J(W_{EE} - \alpha c_1 W_{EI}W_{IE}), \\ \alpha_2 &= I_{ext,E} - \alpha W_{EI}(c_1 I_{ext,I} - c_0). \end{aligned} \tag{13}$$

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 (7), 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

$$S_E = \frac{-(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E}) \pm \sqrt{(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})^2 - 4(-a\alpha_1)(a\alpha_2 - b)}}{2(-a\alpha_1)}, \tag{14}$$

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 (6). By combining equation (6) and equation (13) the steady state is given by

$$-S_E(1 - e^{-d(a\alpha_1 S_E + a\alpha_2 - b)}) + \gamma_E \tau_E (1 - S_E)(a\alpha_1 S_E + a\alpha_2 - b) = 0. \tag{15}$$

The steady states equation (15) is highly nonlinear, and we can not provide an analytic solution. Instead, we solve equation (15) using the Matlab numerical solver *fsolve*. The steady state in equation (15) 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. 1F, Fig. S1E, and Fig. S1F.

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:

$$
\begin{aligned}
&\frac{-S_E^i}{\tau_E} + \gamma_E(1 - S_E^i)r_E^i = 0, \\
&\frac{-S_I^i}{\tau_I} + \gamma_I r_I^i = 0, \\
&-r_E^i + \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}\sum_{j=1}^{N} FLN_{ij}S_E^j) - W_{EI}S_I^i + I_{ext,E}^i) = 0, \\
&-r_I^i + \phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}\sum_{j=1}^{N} FLN_{ij}S_E^j) - W_{II}S_I^i + I_{ext,I}^i) = 0. \qquad (16)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{-S_E^i}{\tau_E} + \gamma_E(1 - S_E^i)\phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i) &= 0, \\
\frac{-S_I^i}{\tau_I} + \gamma_I\phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}L_E^i) - W_{II}S_I^i + I_{ext,I}^i) &= 0. \qquad (17)
\end{aligned}
$$

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

$$
S_I^i = \gamma_I\tau_I(c_1 I_{inh,t} - c_0) = \alpha c_1 W_{IE}J^i S_E^i + \alpha c_1 \mu_{IE} J^i L_E^i + \alpha(c_1 I_{ext,I}^i - c_0), \qquad (18)
$$

where the definition of α is same as in equation (12). Therefore, for the ith cortical area

the excitatory gating variable S_E^i obeys the following steady state equation

$$-S_E^i + \gamma_E \tau_E (1 - S_E^i)\phi_{exc}((W_{EE} - W_{EI}\alpha c_1 W_{IE})J^i S_E^i$$
$$+(\mu_{EE} - W_{EI}\alpha c_1 \mu_{IE})J^i L_E^i + (I_{ext,E}^i - W_{EI}\alpha(c_1 I_{ext,I}^i - c_0))) = 0. \quad (19)$$

First, we analyzed the steady state equation (19) when the transfer function is threshold-linear. The above equation (19) can be written as the steady state of the following set of dynamical equations

$$\begin{aligned} \frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) \quad (20) \\ &= -\gamma_E \tau_E \chi_1 J^i (S_E^i)^2 + (\gamma_E \tau_E \chi_1 J^i - \gamma_E \tau_E (\chi_2 J^i L_E^i + \chi_3) - 1) S_E^i \\ &+ \gamma_E \tau_E (\chi_2 J^i L_E^i + \chi_3), \end{aligned}$$

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 (20) 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 (14)). 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:

$$
\begin{aligned}
&(\gamma_E\tau_E\chi_1 J^i - \gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3) - 1)^2 + 4(\gamma_E\tau_E\chi_1 J^i)(\gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3)) && (21)\\
&= \gamma_E(\tau_E\chi_2)^2(J^i)^2(L_E^i)^2 + 2\gamma_E\tau_E\chi_2 J^i(1 + \tau_E\chi_3 + \tau_E\chi_1 J^i)L_E^i \\
&+ (1 + 2\gamma_E\tau_E(\chi_3 + \tau_E\chi_3^2) + 2\gamma_E\tau_E\chi_1(\tau_E\chi_3 - 1)J^i + \gamma_E(\tau_E\chi_1 J^i)^2) && (22)\\
&= \gamma_E\tau_E^2(\chi_1^2 + 2\chi_1\chi_2 L_E^i + \chi_2^2(L_E^i)^2)(J^i)^2 \\
&+ 2\gamma_E\tau_E(-\chi_1 + \tau_E\chi_1\chi_3 + (\chi_2 + \tau_E\chi_1\chi_3)L_E^i)J^i + (1 + \gamma_E\tau_E^2\chi_3^2 + 2\gamma_E\tau_E\chi_3) && (23)\\
&= 0,
\end{aligned}
$$

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 (21) is a constrain equation in the two-dimensional space of hierarchical position h and long-range excitatory inputs L_E. Therefore, equation (21) 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 (22)), 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. 23). 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 (21) is shown in Fig. 4B.

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

$$\begin{aligned}
&\frac{dS_E^i}{dt} = f(S_E^i, L_E^i, J^i) \\
&= -S_E^i(1 - e^{-d(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3)}) \\
&+\gamma_E \tau_E (1 - S_E^i)(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3) = 0. \qquad (24)
\end{aligned}$$

We use numerical methods to solve equation (24). Numerically solving the equation (24) will give a steady state surface shown in Fig. 4D, Fig. S5, Fig. S6, and Fig. S4. We refer to this surface as *the solution surface*. Any steady-state solution to the network's dynamics will lay on this surface.

Remarkably, our network's solution surface has a very similar geometry to the cusp bifurcation normal form solution surface [46]. 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 [46, 27]. The set of solutions to the steady state equation $\beta_1 + \beta_2 x - x^3 = 0$ gives the cusp normal form solution 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 [46, 27], and with this, its bifurcation diagram. The cusp bifurcation point is given by $\beta_1 = \beta_2 = 0$. In Fig. S6, for illustration proposes, we overlay β_1 and β_2 axes to highlight the resemblance of our network's solution surface with the cusp surface [46].

Similarly to the cusp surface, in our network's solution 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 solution 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 solution 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 [46], 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 (21), we have the bifurcation point (S_E^*, L_E^*, J^*) fulfill the below equation.

$$\begin{aligned} &f(S_E^*, L_E^*, J^*) = \\ &(\gamma_E \tau_E \chi_1 J^* - \gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3) - 1)^2 \\ &+4(\gamma_E \tau_E \chi_1 J^*)(\gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3)) = 0. \end{aligned} \tag{25}$$

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 (21) 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:

$$
\begin{aligned}
&f(S_E^i, L_E^i, J^i) = f(S_E^*, L_E^*, J^*) \\
&+ \begin{pmatrix} \frac{\partial f}{\partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial J^i}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
&+ \frac{1}{2} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix}^T \begin{pmatrix} \frac{\partial^2 f}{\partial (S_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial L_E^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (L_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial L_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial J^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial J^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (J^i)^2}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
&+ O(3), \qquad (26)
\end{aligned}
$$

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 (26) obtaining

$$
\begin{aligned}
\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) = \zeta_1 (S_E^i)^2 + \zeta_2 S_E^i + \zeta_3, \\
\zeta_1^i &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^i &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} L_E^i + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^i &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} L_E^i + J^i L_E^{i,*} - J^i L_E^i) \\
&\quad -\gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} + \gamma_E \tau_E \chi_2 (J^i L_E^i - J^{i,*} L_E^{i,*})),
\end{aligned} \tag{27}
$$

The above equation (27) corresponds to the bifurcation in hierarchy space normal form. We solve the steady state of the above equation (27) self-consistently and predict the firing rate of the delay period working memory states. The self-consistent equations read as

$$
\begin{aligned}
S_E^i &= \frac{-\zeta_2^{i,sce} - \sqrt{(\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce}}}{2\zeta_1^{i,sce}} \\
\zeta_1^{i,sce} &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^{i,sce} &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^{i,sce} &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + J^i L_E^{i,*} \\
&\quad -J^i \sum_{j=1}^{N} FLN_{ij} S_E^j) - \gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} \\
&\quad +\gamma_E \tau_E \chi_2 (J^i \sum_{j=1}^{N} FLN_{ij} S_E^j - J^{i,*} L_E^{i,*})), \\
\Delta^i &= (\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce} \geq 0, \\
S_E^i &\geq 0.
\end{aligned} \tag{28}
$$

To solve the self-consistent equations, first, we solved equation (23) 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 (20) and solving numerically the equation (20). Third, we insert (S_E^*, L_E^*, J^*) into the self-consistent equations in equation (28). Lastly, we solve the self-consistent equations in equation (28) and then predict the firing rate pattern of activity during an active state as shown in Fig. 4B. The normal form in equation (27) can be further reduced to a more general expression as reads below

$$\begin{aligned}\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) \\ &= a_1(S_E^i)^2 + (a_2 J^i + a_3 L_E^i + a_4)S_E^i + (a_5 J^i + a_6 L_E^i + a_7 J^i L_E^i + a_8), \\ &a_1, \ldots, a_8 \in R, \end{aligned} \tag{29}$$

where $a_1, \ldots, a_8$ are parameters calculated re-arrenging terms in equation (27). 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. 4B. Remarkably, the bifurcation in the hierarchy space normal form has a similar mathematical form (up to a translation) as the saddle-node bifurcation [46]. 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 unfeasible 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 (17) 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

$$\begin{aligned} S_E^i = & \frac{\tau_E \gamma_E \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i)}{1 + \tau_E \gamma_E \phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i)}, \\ S_I^i = & \tau_I \gamma_I \phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}L_E^i) - W_{II}S_I^i + I_{ext,I}^i), \end{aligned} \tag{30}$$

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 (30) 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. 6A 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.

Table 1. Parameters for Numerical Simulations of 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. 2 Fig. 4D Fig. 5 Fig. 6), 0.157 (Fig. 3 Fig. S5 Fig. S6A Fig. S7B)
$c1$, $c0$	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. 2C Fig. 2D upper Fig. 3C Fig. 3D upper Fig. 8D upper), 29pA (Fig. S2H left), 8pA (Fig. 5B), 6pA (Fig. S8A), 10pA (Fig. S8B), 16pA (Fig. S8C), 44pA (Fig. 8D lower), 19pA (Fig. 7)

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

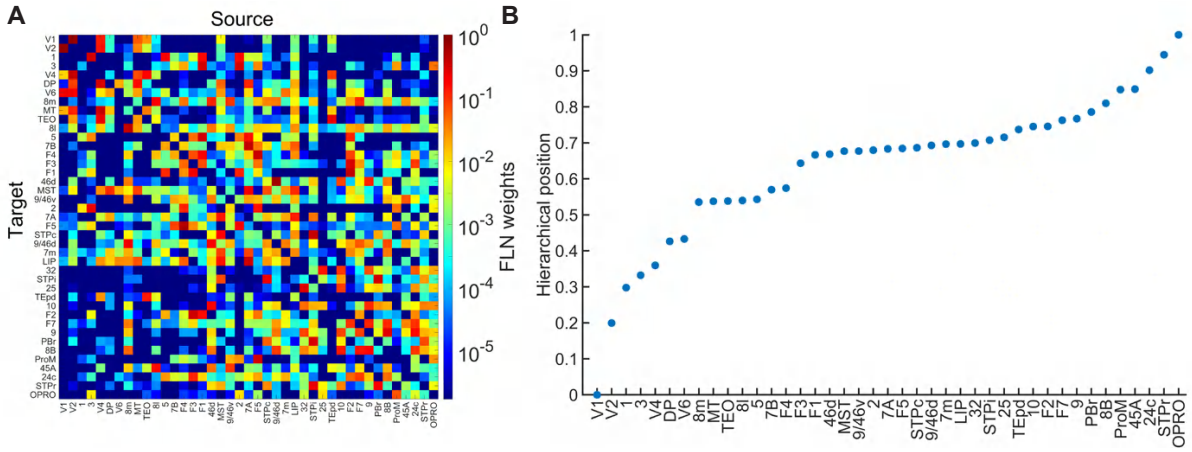

Figure 9: **The FLN matrix and hierarchical position of each brain area of connectome data of the macaque monkey.** (A) The FLN matrix of macaque connectome with 41 brain areas through retrograde tracing [30]. (B) The hierarchical position is estimated based on the 41 brain areas' percentage of supragranular labeled neurons (SLN); the actual method can be found in [77].

The macaque connectome data with 41 cortical areas are shown in Fig.9A. 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. [24]. 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 [24, 77]. 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 [77]. 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 [78]. Importantly, we used numbers of neurons in the model hence generating a weighted hierarchy. The overall FLN values are shown in Fig. 9A. The resulting hierarchical positions are shown in Fig. 9B. For the cortical model of the macaque monkey model in this paper, all the other parameters are the same as in Mejias et al. [25], except the $J_s^{min} = 0.225$, $J_s^{max} = 0.27$, $G = 0.48$, and $Z = 0.82$. The AMPA noise strength for excitatory populations r_E^A and r_E^B are the same.

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 [52]. 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. Model parameters for all panels: $J_s^{max} = 0.27$ and $z = 0.82$.

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:

$$ae^{\frac{-\Delta T}{\tau}} + c, \tag{31}$$

Double-exponential function:

$$ae^{\frac{-\Delta T}{\tau_1}} + (1-a)e^{\frac{-\Delta T}{\tau_2}} + c, \tag{32}$$

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. S10D). Consequently, we disregard this rapid component and focus solely on the prolonged temporal scale in Fig. 8D 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.

Supplementary figures

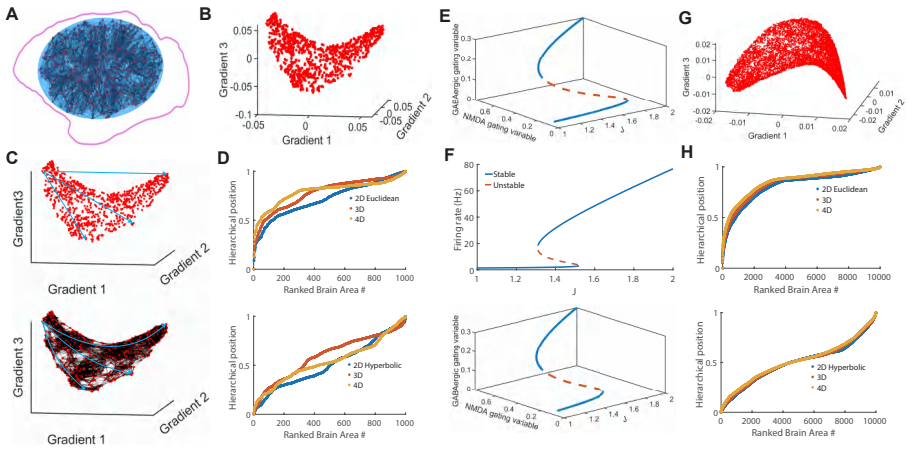

Figure S1: **The supplementary figure of generated connectivity, hierarchy, and local circuit model.** (A) The generated super neocortex network with 1000 brain areas is consistent with the macaque neocortex network statistic. (B) The three-dimensional embedding hyperbolic shapes of the generated super 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. 1 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 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.

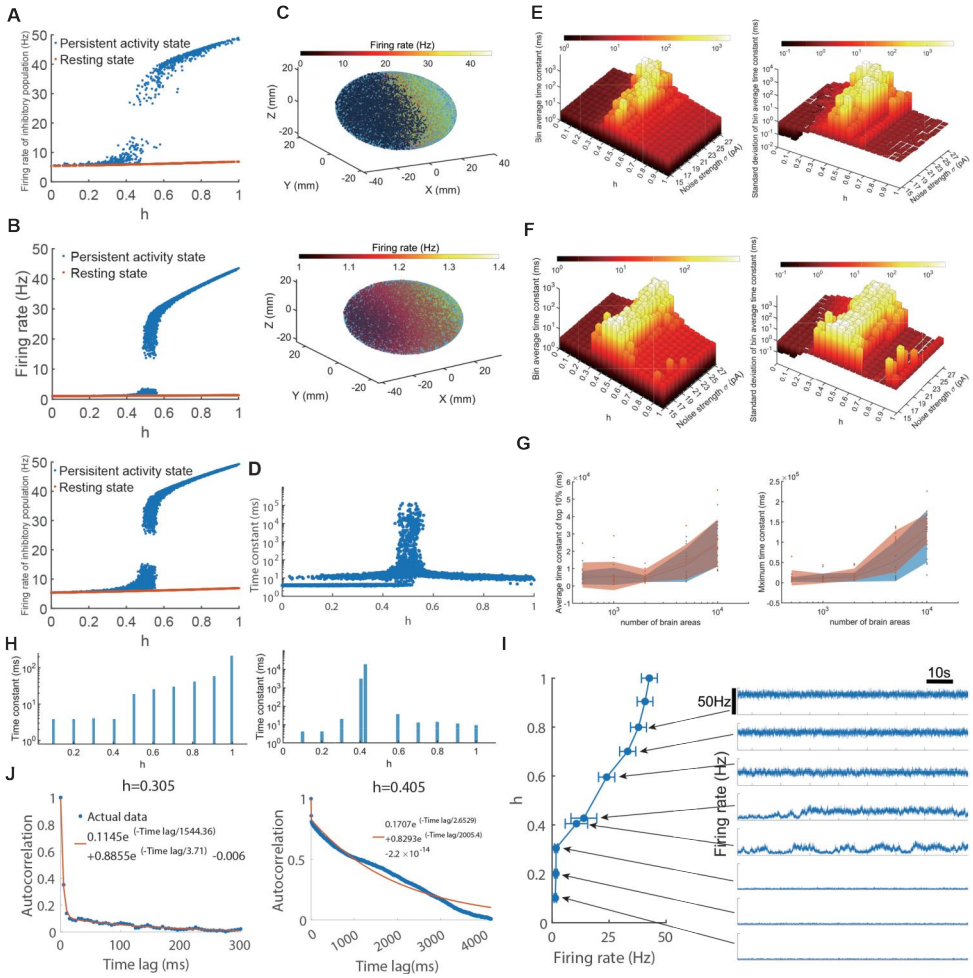

Figure S2: **The supplementary figure of bifurcation in hierarchical space.** (A) Firing rate of all the inhibitory populations for the same parameter set of panel A of Fig. 2 of the main text. (B) The firing rate of the excitatory (upper) and inhibitory (bottom) population of both active (blue) and 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 S2's caption. In this panel, we use the bin averaged time constant of the same set of panel A with bin size equal to 0.05 hierarchical position interval. We averaged from 5 ensemble realization 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 realization and 10 different super neocortex network for each time constant bin. (G) The average time constant of 10% largest time constant (left) and 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 network. 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 A of Fig. 2. (I) The average and standard deviation of the firing rate of 10 chosen cortical brain areas in the persistent activity state and its 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.

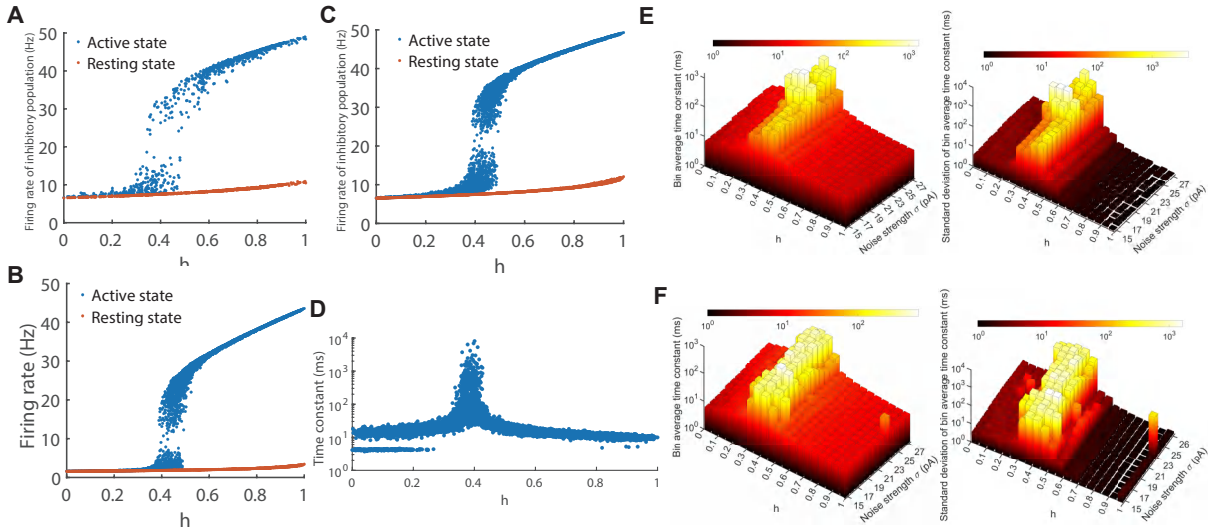

Figure S3: **The delay activity state and time constant for brain networks with** 10000 **brain areas and** $d = 0.157$**.** (A) Firing rate of all the inhibitory populations for the same set of panel A of Fig.3 of the main text with $d = 0.157$. (B) The firing rate of the excitatory population of both 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 A with bin size equal to 0.05 hierarchical position interval. We averaged from 5 ensemble realization 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 realization and 10 different super neocortex network for each time constant bin.

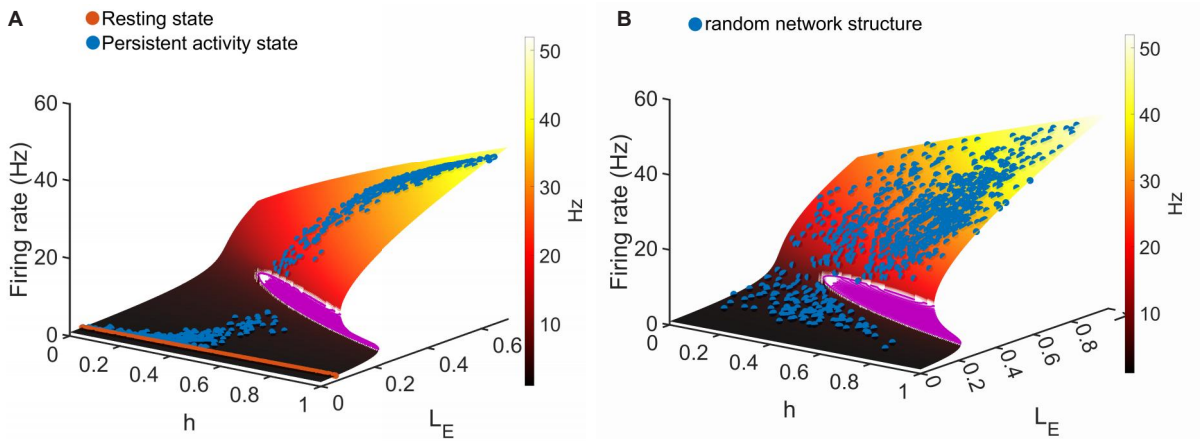

Figure S4: **The geometry of distributed attractor states in parameter space.** (A) The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the solution surface with $d = 0.17$ (corresponding to Fig. 2A). (B) The mapping of the delay active state for randomly shuffled network connections of the generative network to the solution surface (corresponding to Fig. 5A).

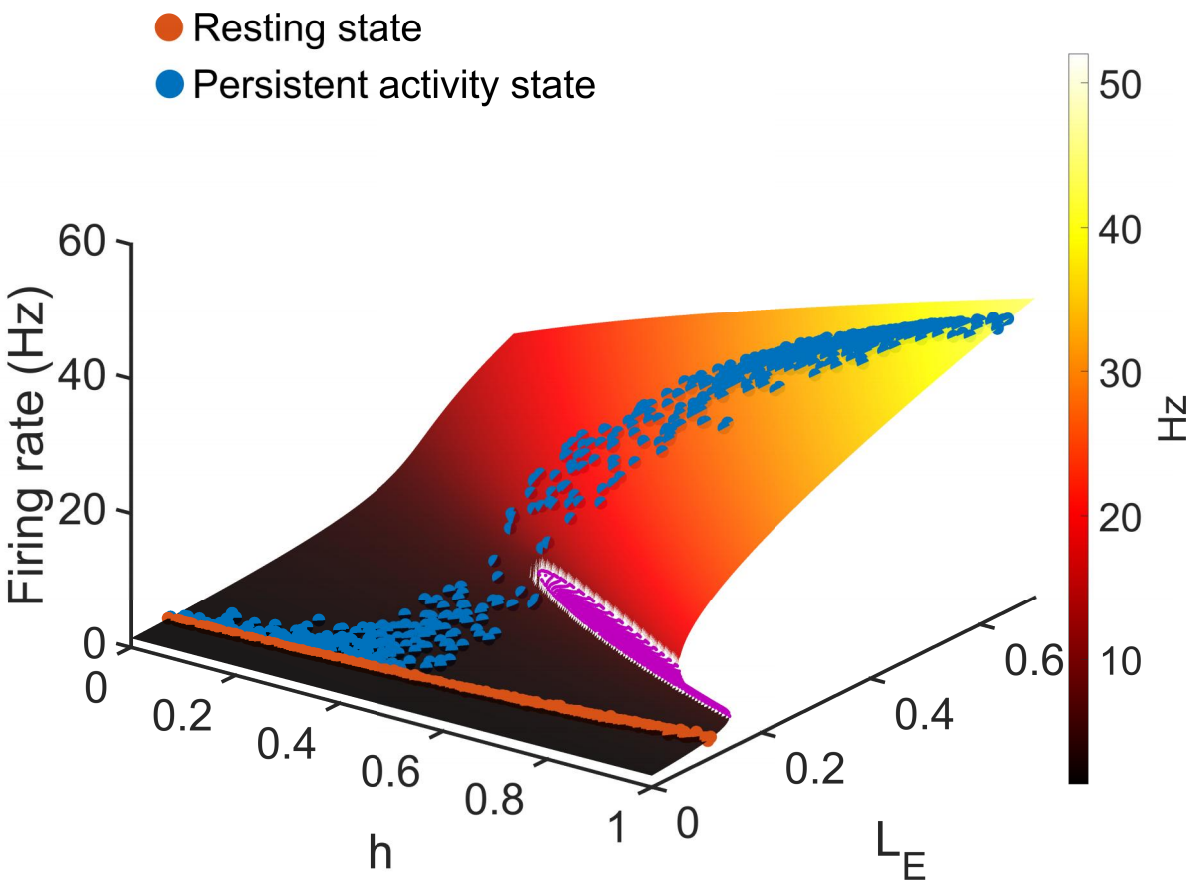

Figure S5: **The solution surface of $d = 0.157$.** The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the solution surface ($d = 0.157$). In this case, the active state continuously transitions without a firing rate gap.

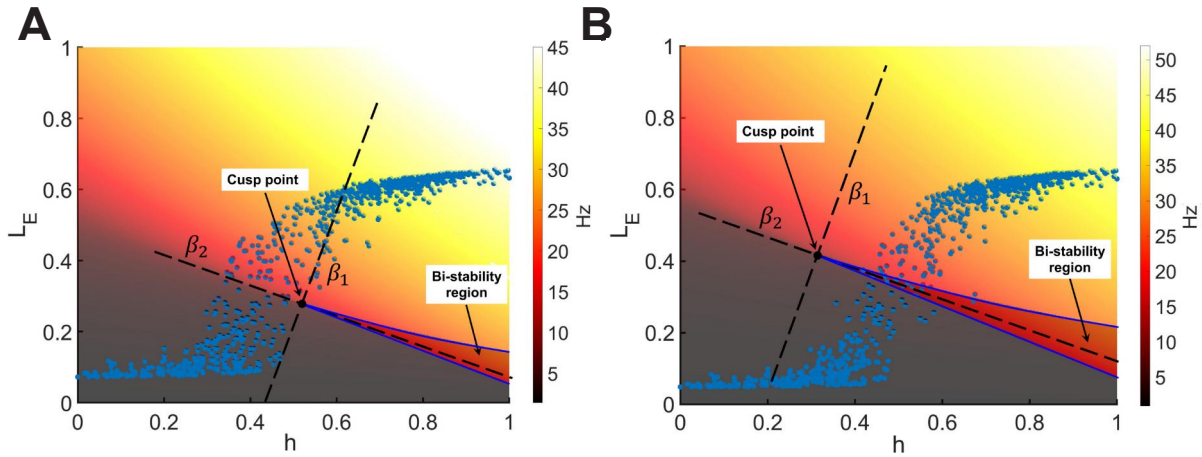

Figure S6: **Cusp geometry determines the bifurcation in hierarchical space.** (A and B) The geometry of the solution surface and the cusp point determine the bifurcation in hierarchical space. The blue dots correspond to the persistent activity state of the neocortex model with $d = 0.157$ (panel A) and $d = 0.17$ (panel B), respectively. For comparison proposes, 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 solution 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 solution surface is folded for hierarchy values h and long-range excitatory input current L_E in the ranges $0.6-1$ and $0.1-0.2$, respectively.

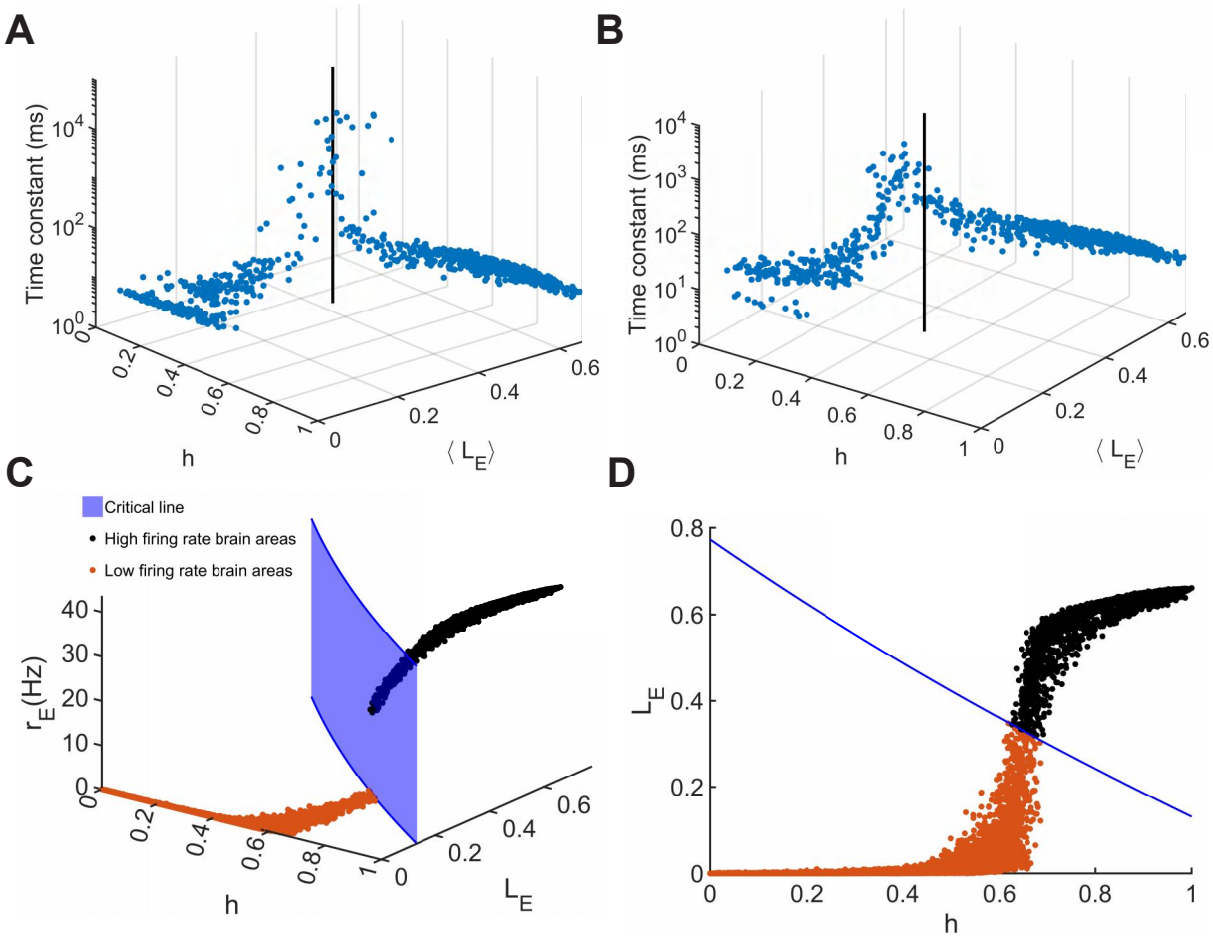

Figure S7: **Time constant around cusp point and geometry of solution surface of network with 10000 brain areas with threshold-linear transfer function.** (A-B) The time constant of all the brain areas arrange in the h and $\langle L_E \rangle$ space for $d = 0.17$ and $d = 0.157$ with 1000 brain areas, 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 C of Fig. 2 and 3, 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.

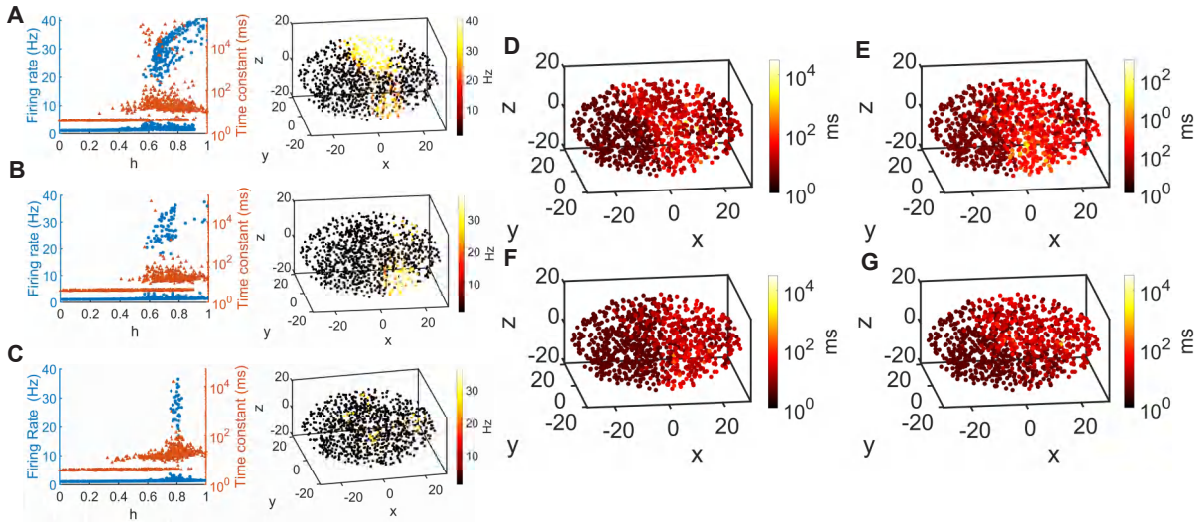

Figure S8: **The supplementary figure of the diversity of distributed working memory states.** (A-C) The firing rate (blue) and time constant (brown) of active state $S1$, $S3$, and $S4$ (left) for each cortical area, respectively. The spatial distribution of active state $S1$, $S3$, and $S4$ (right) firing rate in the generative model ellipsoid, respectively. (D-E) In the generative model ellipsoid, the spatial distribution of active state $S1$, $S2$, $S3$, and $S4$ time constants, respectively.

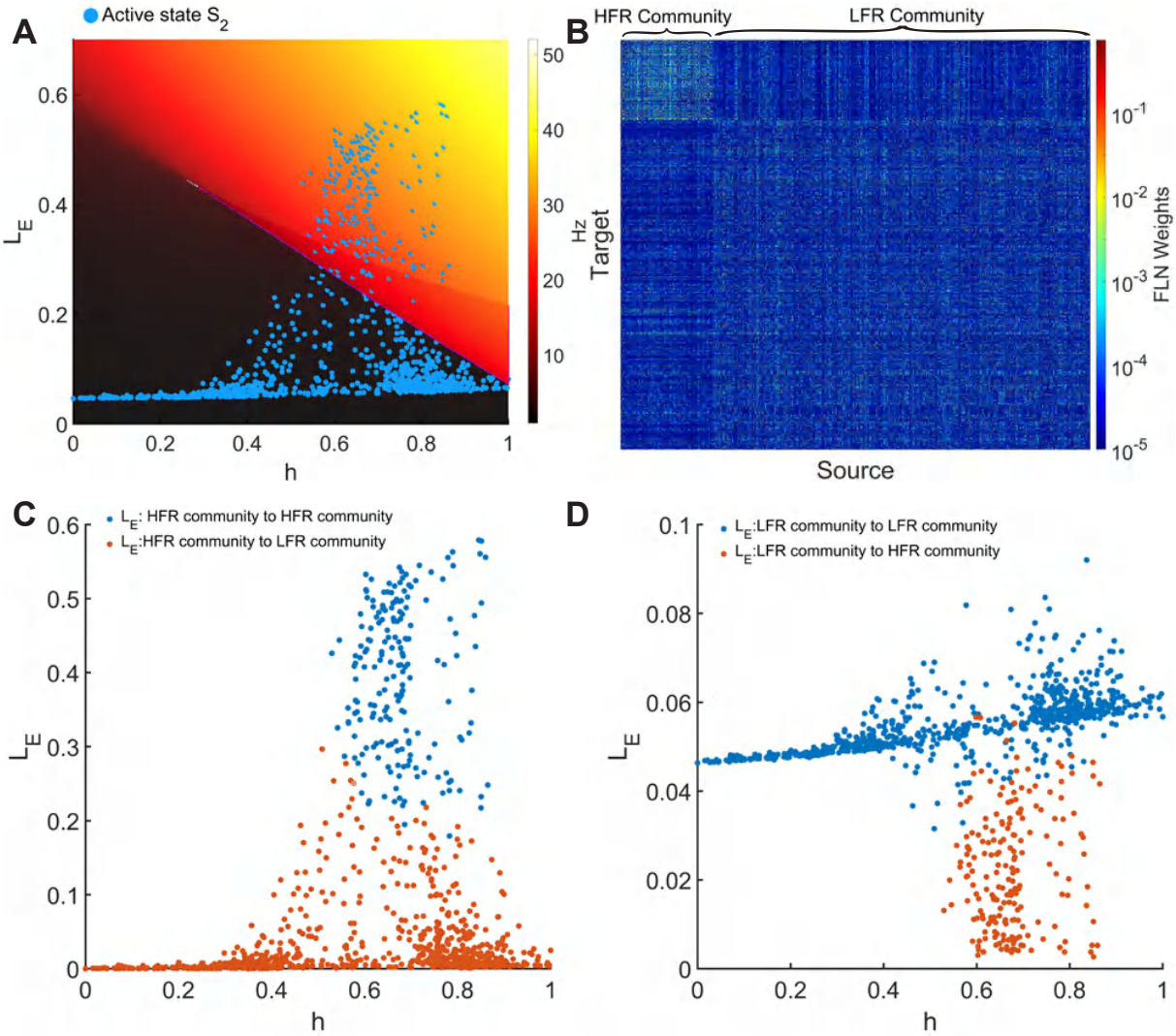

Figure S9: **The supplementary figure of L_E of the bump active state S_2.** In the bump state S_2, the long-range excitatory input (L_E) is widely distributed within the hierarchical bump region. Brain areas receiving a sufficiently large L_E will exhibit a high firing rate, whereas those with a small 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 S8'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 solution surface for the active state S_2 is illustrated in Fig. 6B. (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.

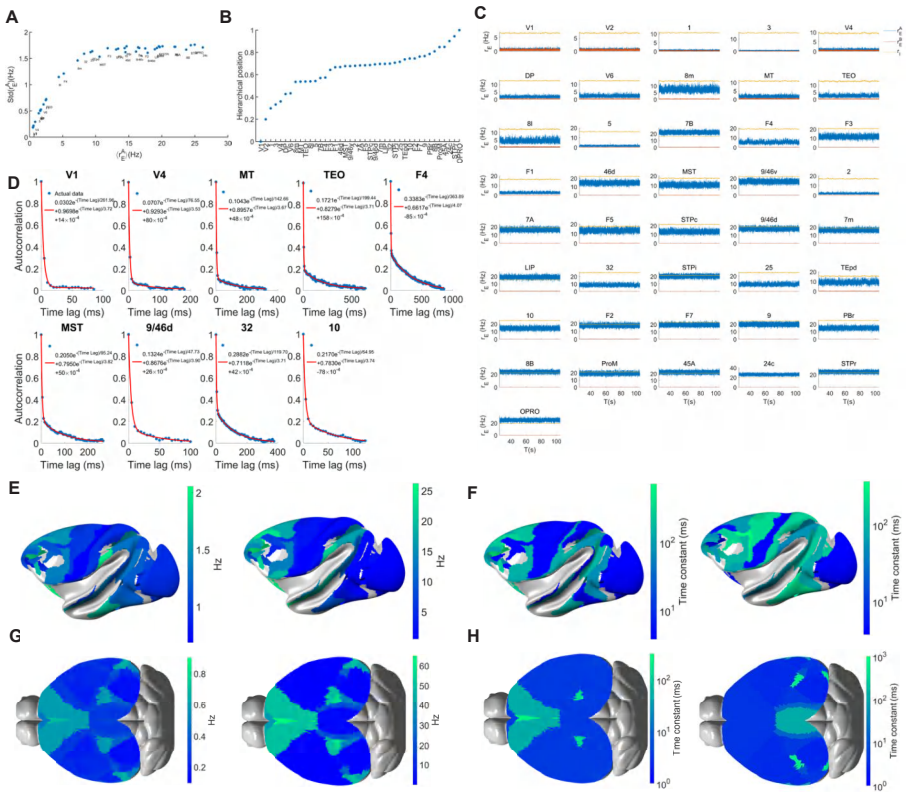

Figure S10: **The supplementary figure of bifurcation in the hierarchical space of connectome-based cortical models of macaque monkey and mouse.** (A) The average firing rate versus the standard deviation of each brain area of the macaque neocortex with 41 brain areas. (B) The normalized hierarchical position of the 41 brain areas. (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. 8D in the main text. (E) Spatial activity map of the resting state (left) and the monotonic delay period working memory state (right) of the macaque neocortex model with 41 brain areas with the model in [25]. (F) 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 E. (G) 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 [26]. (H) 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 G.

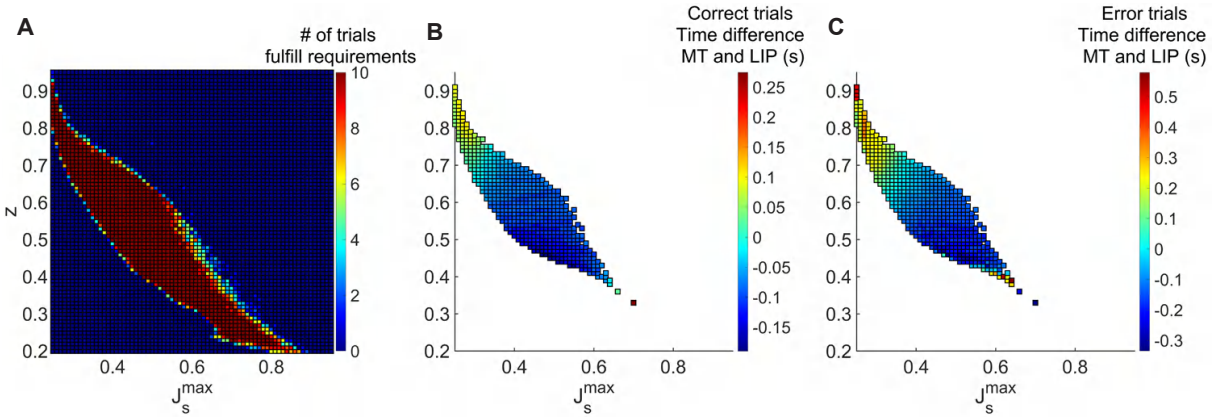

Figure S11: **The parameter region in J_s^{max} and z space that fulfill all the requirements of working memory and decision-making of connectome-based cortical models of the macaque monkey with** 41 **brain areas.** We updated Mejias' model [25] with 41 brain areas and it's corresponding hierarchical value. After keeping all the other parameters the same as the original model [25], 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.

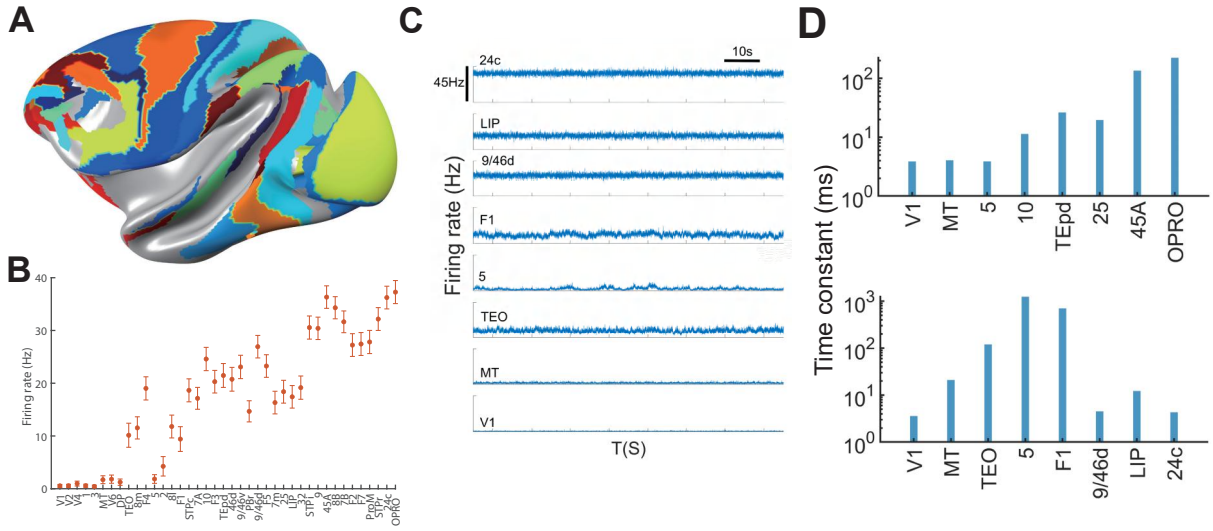

Figure S12: **Bifurcation in the space of the connectome-based cortical models of the macaque monkey (A-D)** **[25]** (A) Lateral view of macaque neocortex surface with 40 model areas in color. (B) Firing rate of 40 brain areas, ranked by the hierarchical position. The brain areas' hierarchical position is the same as in Froudist-Walsh et al. [24]. (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.

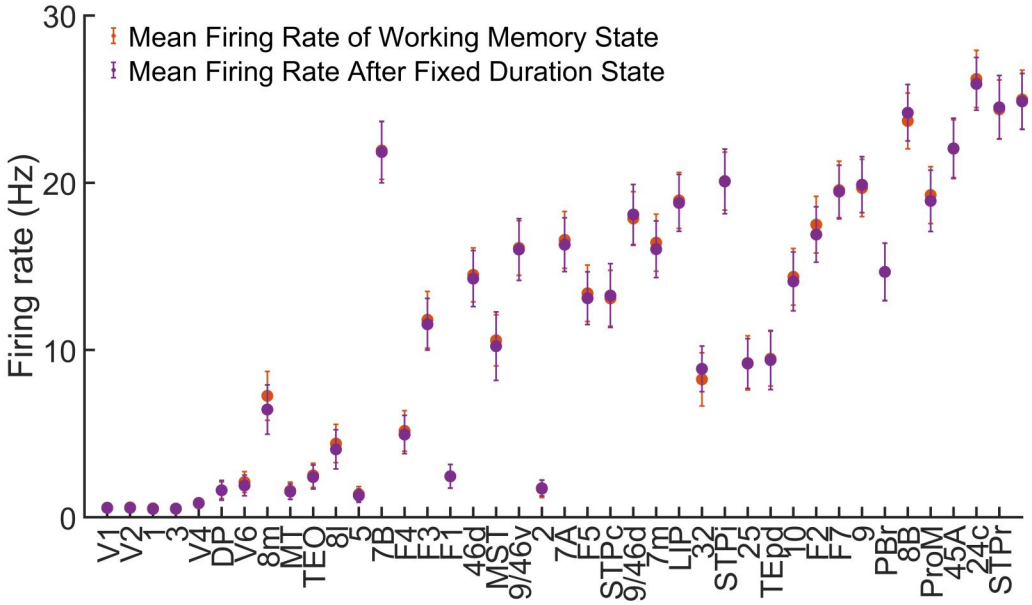

Figure S13: **The delay-period activity for the fixed-duration decision-making task and the working memory state is depicted.** The dots indicate the average firing rate, while the error bars represent the standard deviation of the firing rate.