Bifurcation in space:
Emergence of function modularity in the neocortex

Xiao-Jing Wang,[1,*,†], Junjie Jiang,[1,2,†], Ulises Pereira-Obilinovic,[1]

[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.

† These authors contributed equally to this work.

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

Abstract

Abstract: How does functional modularity emerge in a multiregional cortex made with repeats of a canonical local circuit architecture? We investigated this question by focusing on neural coding of working memory, a core cognitive function. Here we report a mechanism dubbed "bifurcation in space", and show that its salient signature is spatially localized "critical slowing down" leading to an inverted V-shaped profile of neuronal time constants along the cortical hierarchy during working memory. The phenomenon is confirmed in connectome-based large-scale models of mouse and monkey cortices, offering an experimentally testable prediction to assess whether working memory representation is modular. Many bifurcations in space could explain the emergence of different activity patterns potentially deployed for distinct cognitive functions, This work demonstrates that a distributed mental representation is compatible with functional specificity as a consequence of macroscopic gradients of neurobiological properties across the cortex, suggesting a general principle for understanding brain's modular organization.

Introduction

Recent technical advances are enabling neuroscientists to image calcium signals or record spiking activities of many single cells in behaving animals [1, 2], opening a new era of investigating neural computation distributed in the multiregional brain [3]. Studies reported widespread neural correlates of task relevant information, observed wide-spread activity signals are sometimes interpreted as evidence of a lack of any spatial specificity. Therefore, a central challenge in the field is to elucidate local versus global neural processes underlying behavior. We tackled this challenge using computational modeling of the multiregional 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 [5]. We propose that modularity is understood in terms of a selective subset of cortical areas, which are not necessarily spatially congruent, engaged in a distinct brain function. This definition is compatible with distributed neural representation and processing across multiple brain regions, but in contrast to the absence of modularity manifest by merely graded variations of engagement across the entire cortical mantle.

We designed this research to address the question of biological mechanism underlying the emergence of functional modularity. At the same time, our work suggests a sensitive test to arbitrate whether there is modularity of a given cognitive function in the first place by neurophysiological experiments.

According to a central tenet of neuroscience, a canonical local circuit is repeated numerous times throughout the cortical mantle and shared across mammalian species [6]. Consistent with this view, the cortex is commonly described as a graph where parcellated areas are considered as identical nodes, each with different inputs and outputs that determine its function [7]. 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 [8, 9, 10]. For the sake of concreteness, consider working memory, our brain's ability to maintain and manipulate information internally in the absence of external stimulation [11, 12]. Working memory represents an excellent case study because it is essential for major cognitive processes and has been extensively investigated. The underlying mechanism of this core cognitive function involves persistent neural firing that is self-sustained internally during a temporal delay between a stimulus and a response [13, 14]. A large body of literature have documented that working memory representations are distributed over some cortical areas but not others [15, 16]. In particular, for working memory of visual motion information, there is evidence that the middle temporal (MT) area does not, but its monosynaptic projection target, the medial superior temporal (MST) area, does show persistent activity during a mnemonic delay [17], suggesting a sharp onset of working memory representation along the cortical hierarchy. How can such functional modularity be reconciled with a 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 [18, 19, 20] have been quantified, they are not random but display macroscopic gradients along low-dimensional axes such as the cortical hierarchy [21, 22]. Such macroscopic gradients have been incorporated into connectome-based models of a multiregional cortex of macaque monkeys [23, 24] and mice [25] for distributed working memory. In such a model, the idea of a canonical local circuit is implemented by the same 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 macroscopic gradients [22]. Computational modeling revealed an abrupt transition, at some stage of the cortical hierarchy, that sep-

arates cortical areas exhibiting information-coding self-sustained persistent activity from those that do not. These findings offer hints that the entire cortex organized according to a canonical microcircuit architecture can nevertheless be divided into subnetworks of areas for modular functions.

The present work was designed to test the hypothesis that such a transition represents a "bifurcation in space". The mathematical term "bifurcation" denotes the sudden onset of a qualitatively novel behavior by virtue of a graded change of a dynamical system's property [26]. The idea of bifurcation in space is conceptually novel because it emerges from an interactive large-scale brain circuit as a collective phenomenon but occurs locally in space. We used a generative model of the cortex [27], which can generate a model with an arbitrary number of areas characterized by the experimentally measured connection statistics [28, 29]. With a large number of parcellated areas in such a model, we were able to "zoom in" on the transition point along the cortical hierarchy. We derived a normal form equation close to the bifurcation [26] in our large-scale cortex model. The analytical prediction fits well with numerical simulation results, thereby firmly establishing the concept of bifurcation in space.

We found that, as a signature of phase transition, the timescale of neural activity diverges to infinity at the transition, called "critically slowing down" [30, 31]. Consequently, the time constant of neural firing fluctuations would be maximal for cortical areas near the transition, larger than those in both lower areas devoid of persistent activity and higher areas showing robust persistent activity. In other words, along the cortical hierarchy, there is an inverted V-shaped pattern of time constants that dominate neural fluctuations during persistent activity associated with working memory. Note that critical slowing-down has been shown previously for a dynamical system as a whole, but here is manifest locally at a particular site of a spatially extended system.

These results are highly non-trivial. The model is set up in such a way that, when disconnected from each other, none of isolated areas is capable of maintaining persistent activity. Consequently, the observed working memory representation must be a collective phenomenon depending on long-distance connection loops. The connectivity is dense, about 67% of all possible inter-areal pathways are present. Yet, bifurcation occurs locally in space. Furthermore, importantly, such bifurcation in space is defined for any one of numerous spatially distributed persistent activity states, each engages a subset of cortical areas (a discrete specialized system) and could potentially serve a distinct function. In other words, there are many bifurcations in space for a single multiregional cortical system with fixed parameters. This finding suggests that the concept of bifurcation in space can account for the modularity of various cognitive functions like decision-making [32]. Finally, we reproduced the salient finding of an inverted-V shaped profile of time constants in the connectome-based models of the macaque monkey cortex [24] and the mouse cortex [25], thereby identifying specific model predictions that are testable experimentally.

Results

A generative model for the mammalian neocortex.

To understand the fundamental mechanism of bifurcation in the neocortex's space, we build a simplified multi-regional neocortex model. In Physics, a phase transition is studied by "zooming in" very close to a criticality. For example, ice melts at zero degree Celsius; temperature must be precisely tuned to that critical point for understanding the ice-to-water transition. Similarly, in order to rigorously investigate mathematically a bifurcation in space, we need to "zoom in" close to an abrupt transition point along the cortical hierarchy that separates those areas engaged in working memory representation from

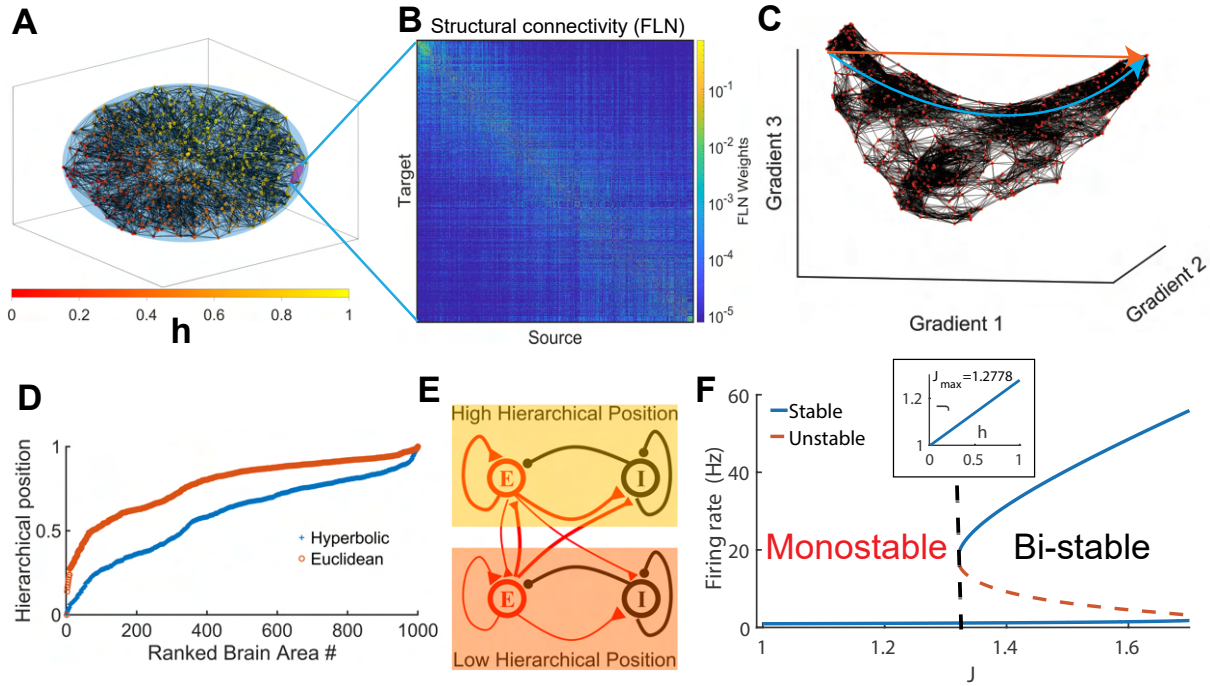

Figure 1: Spatially embedded generative model of a cortex. (A) Network connectivity of 1000 parcellated cortical areas produced by a generative model of the mammalian cortical connectivity [27]. In this model, the network's parcellations are embedded in an ellipsoid and are connected by spatially dependent and weighted connections. (B) Connectivity matrix obtained from the generative model. The connectivity weights are defined similarly to the fraction of labeled neurons (FLN) in the connectomic analysis of the macaque cortex [29]. (C) Three-dimensional diffusion map embedding of the generated 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, hierarchical position of any area can be defined based on Euclidean (brown) or hyperbolic (blue) distance to a starting area at the bottom of the hierarchy (see Methods). (D) Hyperbolic metric shows 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. Long-range projections between cortical areas are excitatory. The strength of local and long-range connections' strength is indicated by the thickness. (F) Bifurcation diagram of an isolated cortical circuit (see equation (1) in SI). In the large-scale system, the local E-to-E and E-to-I weights are scaled with the factor J, which displays a macroscopic gradient as a function of the hierarchy value h (Insert).

those that are not. However, a connectome-based model of the mouse or monkey cortex has a relatively small number of areas thus the distance between any pair of adjacent areas along the hierarchy cannot be reduced as much as desired. To overcome this limitation, we used a generative model of (unlike a purely topographic graph) a spatially embedding mammalian cortex [27].

We aim for this model to capture central aspects of the mammalian neocortical connectivity and neural dynamics but to be simple enough to be suitable for mathematical analysis. The model (see Fig. 1A) is generative and random, thus can be used to produce *realizations* of a mammalian neocortical network model with sample connectivity matrices (see Fig. 1B for one network realization) that share the same statistical distributions observed in the inter-areal connectivity [28, 29, 33, 27]. This network model has three advantages. First, conclusions rendered from this model can be applied to different mammalian cortices [29, 34, 35, 36]. Second, in this model, the number of brain areas can be arbitrarily large, enabling us to examine bifurcation phenomenon close to a transition point. Third, we show that our results are robust by studying the network dynamics over multiple network realizations, i.e., our results depend only the connection statistics but not qualitatively on the specific network realizations. In the last section of this paper, we show that all our results hold in connectome-based models of macaque cortex [24] and mouse cortex [25].

We define the hierarchical distance using the diffusion map embedding method [37, 21] applied to our model. This class of nonlinear dimensionality reduction method embeds the connectivity. In the embedding space, closer areas share a larger number of paths connecting them with stronger connections, while areas further apart share fewer paths and weaker connections (see Methods). Interestingly, the embedding structure of the generated connectivity conforms to a low-dimensional hyperbolic shape (see Fig. 1C).

To define a hierarchical distance, we arbitrarily choose as the start of a hierarchy the cortical area at one of the tips of the hyperbolic shape. We found that for the hyperbolic distance (the distance defined along the hyperbolic shape) the cortical areas display a smooth progression evenly distributed in each hierarchical position (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 for defining the hierarchical position. Strikingly, after remapping each brain area's hyperbolic hierarchical position into the ellipsoid's position (see Fig. 1A), we found the hierarchical position increases along the major axis of the ellipsoid, similar to the hierarchy of the mammalian cortex roughly along the anterior-posterior axis (Fig. S1A).

Based on the hierarchy defined above, we build a simplified yet biologically realistic neocortex model in which macroscopic properties of a canonical circuit vary along the hierarchy. 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 [22], the local and long-range excitatory weights are scaled by the hierarchical position (i.e., $J \propto h$, Fig. 1F). When decoupled, a cortical area has only a resting state at a low firing rate if the strength of synaptic excitation is below a threshold $J_{\text{threshold}} \approx 1.32$ ($J < J_{\text{threshold}}$), exhibits bistability of a resting state and an elevated persistent activity state if $J > J_{\text{threshold}}$. In order 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 $J_{\text{threshold}}$ (insert in 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 [38, 13, 14] to the large-scale multiregional brain.

Bifurcation in the hierarchy space.

The dynamics of the connected network of interacting areas strongly differ from that of isolated areas. The network exhibits a coexistence of a resting state (where all areas of the network exhibit a low firing rate state ∼1Hz) and active states (where some cortical areas display persistent high firing rates while others have 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, the areas that are engaged in the persistent activity state are higher in the hierarchy and separated by a firing rate gap, indicating a transition zone. The size of this transition zone systematically shrinks when we increase the network size (Fig. S2B). In the limit of infinitely large networks, we expect this transition zone shrinks to a point. We denote this point in the hierarchy space 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 [39].

Mapped into our generative model's ellipsoid where the connectivity was originally embedded, the firing rate for 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 (lower panel), in the persistent activity state there is a firing rate gap between the posterior and the anterior areas (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. Do we observe critical slowing down in our network associated with the bifurcation in space? The time scale of our network's fluctuations increases from milliseconds to tens of seconds for areas within the bifurcation region (Fig. 2C), displaying the critical slowing

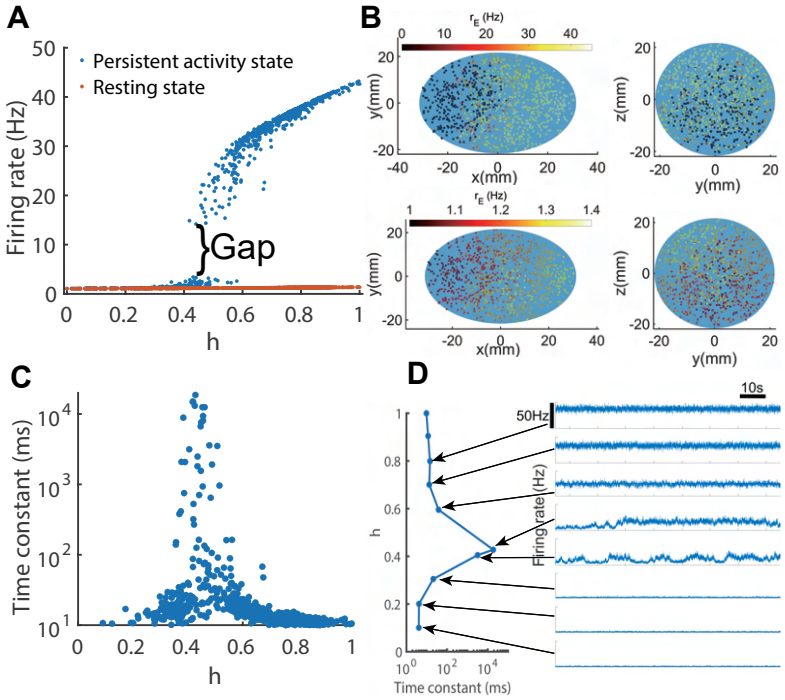

Figure 2: Bifurcation in hierarchical space. (A) Resting (brown) and persistent activity states (blue) are shown with firing rates plotted against areas ranked by the hierarchical position. In the persistent activity state, a subset of areas represent working memory and are separated from the rest of the network by a firing rate gap. (B) The front (left) and side (right) view of the spatial distribution of the persistent activity state (top) and resting (bottom) state in the generative model's ellipsoid. (C) The time constant of all the brain areas at the persistent activity state of panel A with $1,000$ brain areas. (D) The time constant (left) of 10 chosen brain areas and firing rate time series (right) of 8 chosen cortical areas when the network is in the persistent activity state.

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 the time scales characteristic of the critical slowing down is only observed for the active state. In contrast, in the resting state, the timescales increase monotonically with the hierarchy (Fig. 2A), recapitulating known results from the linear theory of large-scale models constrained by anatomical connectivity [40, 41] (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 parts: activity is roughly constant and low for areas low in the hierarchy, then starts to increase without a discrete jump of firing rate higher along the hierarchy (see Fig. 3A and figure 3B, blue dots). The larger the input-output gain, the sharper the firing rate progression becomes (see Fig. 3A) until the bifurcation in space comes to light. The firing rate gap at the transition increases with the gain d reaching its maximum for the threshold-linear transfer function (see Methods, equation (2)) [42] (Fig. 3A-B). With a sufficiently small d value, the transition becomes smooth with virtually no firing rate gap, but the working memory state is still characterized by an inverted V-shape time scale profile (Fig. 3C-D). This observation suggests that critical slowing down does not require the presence of a firing gap in the persistent activity state.

The geometry of distributed attractor states.

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 connection matrix and S_E^j is the output synaptic variable from area j (see illustration in Fig. 4A). This input current for each area i varies with the hierarchical position and

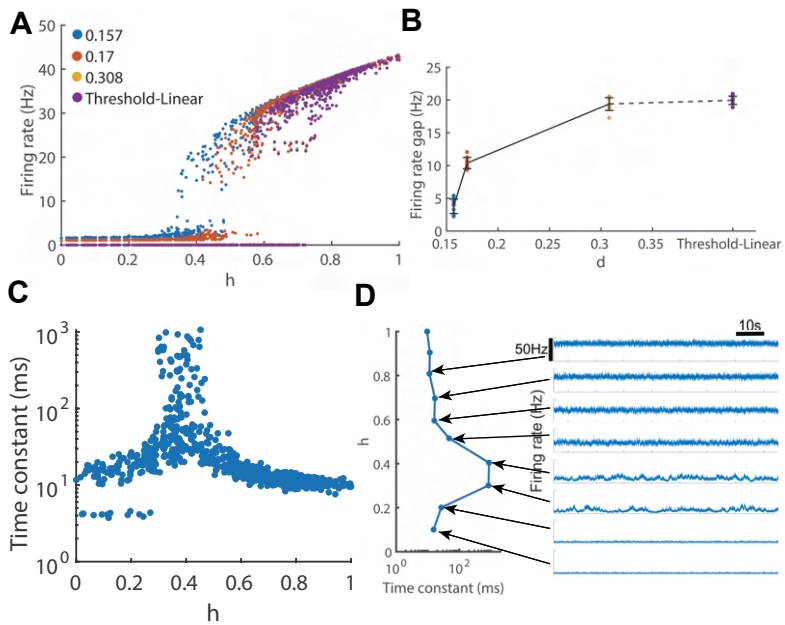

Figure 3: The effects of the input-output transfer function gain parameter d on bifurcation in space. (A) The persistent activity state for different d values. When the gain parameter increases from 0.157 to ∞, the transition along the hierarchy changes from a smooth pattern to another characterized by a jump in the firing rate. (B) The maximum firing rate difference among all the pairs of areas. The mean (dots) and error bars correspond to the results of 10 different network realizations. (C) The time constant of all the cortical areas at the active state. (D) The time constant (left) and firing rate time series (right) of 8 choose brain areas in the active state with the smooth transition. The gain parameter $d = 0.157$ for panels C and D.

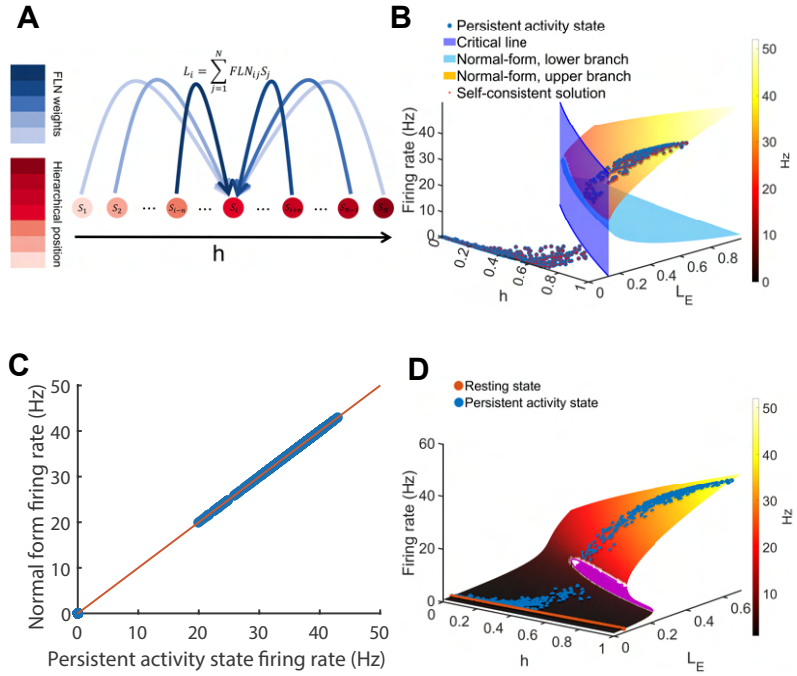

Figure 4: The geometry of distributed attractor states in parameter space. (A) Illustration of long-range excitatory input currents from the network's recurrent interactions to the ith brain area. The gradient of red and blue color corresponds to hierarchical positions of brain areas and the weight of 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 (see Methods). The bifurcation in hierarchical space happens 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, showing perfect agreement between the two. (D) The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the solution surface with $d = 0.17$.

depends on the network's internal state. Therefore, the simulation result shows that the bistability region in the (h,L_E) plane increases with the gain d of the input-output transfer function. In the limit of infinite gain, which is equivalent to a threshold-linear transfer function (see Methods, equation (3)), there is bistability in the entire region of hierarchy h and L_E (see Fig. 4B). In this limit, using mathematical expansions similar to those used in the theory of the normal forms [43], the normal form of bifurcation in hierarchical space is obtained (see Methods, equation (26)). Based on this normal form expression, we solved self-consistent equations for the N firing rates and N L_E variables. The analytical results match perfectly with our numerical simulations (Fig. 4C).

As the case for finite gain parameter d, we found that for any steady state of our 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, and Fig. S5A-B). We refer to this surface as the solution surface (Fig. 4D, Fig. S4 and Fig. S5A-B). Importantly, the solution surface does not depend on the network size N. In the resting state, both the firing rates and the long-range excitatory inputs L_E are low (Fig. 4D). In a persistent activity state (Fig. 2A with $d = 0.17$), the more active an area is, the larger its output synaptic gating variable is. Therefore, those areas below the transition receive weak input currents from strongly interconnected areas nearby in the hierarchy [27, 28] (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 the solution surface folds at a *cusp*, which corresponds to a point in the two-dimensional space of h and L_E (Fig. S5A-B). This geometry is reminiscent of a cusp described in the classic bifurcation theory for non-linear dynamical systems [44, 43]. 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 [43]. From this point to the folded region in parameter space, the system displays a transition 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. 2D). However, it is important to highlight that bifurcation in space of our model is conceptually different from the conventional cusp. 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), which are shown in the cloud of dots in Fig. 2 A. Recall that when decoupled from each other, none of the isolated areas shows elevated persistent activity. Therefore, this is a collective behavior emerging from the complex area-to-area interactions, yet with the transition occurring at a particular location of the hierarchy.

A diversity of distributed attractor states of persistent activity.

Interestingly, in our network, we found a large number of persistent activity states each with a distinct spatial distribution, similarly in connectome-based models of macaque monkey [24] and mouse [25]. For most of these attractor states the firing rate increases monotonically with the hierarchy (see Fig. 2A, 5A, Fig. S7A). However, surprisingly, a sizable number of attractor states display a localized *bump* of activity in the hierarchy space (Fig. 5A, bump active states; Fig. 5B; and SI Fig. S7C). Strikingly, unlike the monotonic active states where persistent firing roughly follows the posterior-anterior axis of the model ellipsoid, bump active states display in general, scattered spatial patterns of working memory activity (Fig. 5B and Fig. S7C, right panel). The time constant of each area's neuronal fluctuations in those bump states is maximal and exceptionally long at

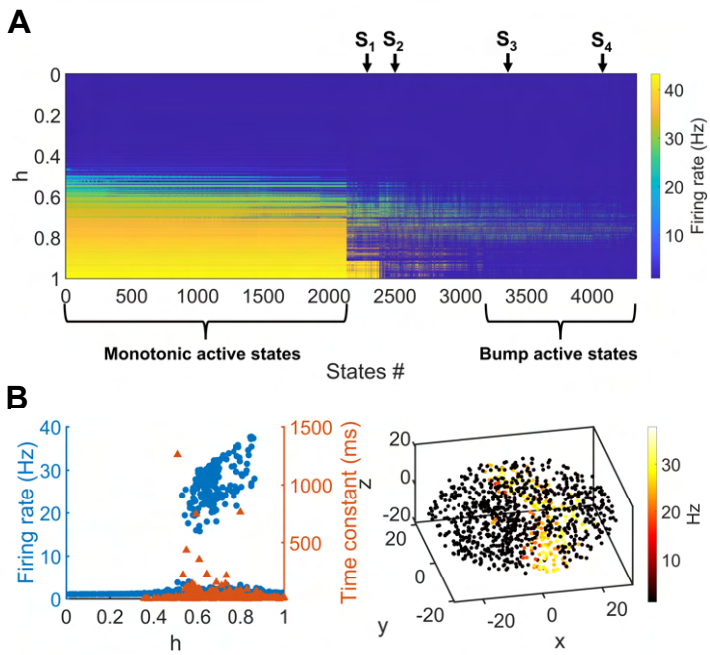

Figure 5: A diversity of distributed working memory states. (A) The firing rate pattern of 4333 active states with cortical areas ranked by hierarchy values. There are two major classes of distributed working memory states: monotonic and bump active states. The x-axis corresponds to the rank of all persistent activity states according to the number of high firing rate areas with firing rates larger than $10Hz$. (B) An example of bump-shaped persistent activity state indicated by S_2 in (A). Left: firing rate (blue) and time constant (brown) as a function of the hierarchy; Right: spatial distribution of firing rates in the generative model ellipsoid. Three other persistent activity states S_1, S_3, and S_4 are shown in Fig S7 A, B and C, respectively.

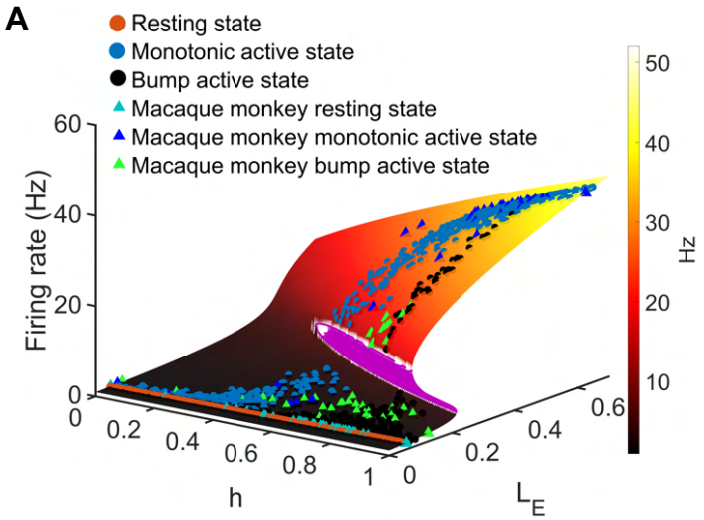

Figure 6: All attractor states of distributed working memory are shown on the cusp surface. (A) The firing rate of all the brain areas of the resting, monotonic persistent activity, and bump persistent activity states for the generative and biologically realistic models.

one of the edges of the bump, for a particular noise strength, as depicted in Fig. 5B and Fig. S7C.

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 of different sizes (see Fig. 6A). This provides a unifying picture; different states take different parts of the solution surface, since cortical areas at the same hierarchical position h have different firing rates and long-range excitatory input current L_E values in distinct internal states.

In summary, bifurcation in space is defined separately for each of the internal states of distributed persistent activity. In other words, a single network has many bifurcations in space, each marked by a subset of areas engaged in the corresponding persistent activity state and critical slowing down at its own transition spatial location.

Bifurcation in space in connectome-based cortex models of monkey and mouse.

Do models of the multiregional cortex constrained by anatomical data also display critical slowing down? To address this question, we considered a connectome-based model of macaque cortex [24] with 40 cortical regions (Fig. 7A). For the sake of simplicity but suitable for stimulus-selective working memory, each brain area has two excitatory populations encoding stimuli and one inhibitory population [24]. The strength of long-range and local excitation follows a macroscopic gradient proportional to the spine count per pyramidal neuron [40, 22]. As in our abstract model, this connectome-based model exhibits the coexistence of a resting state (firing rate around $1Hz$) and persistent activity (more than $10Hz$) encoding working memory (Fig. 7B). The spatial firing rate distribution during working memory states is modular, with only a few areas displaying persistent firing (Fig. S8E). Consistent with the macaque monkey physiological experimental observations [15], association cortical areas but not early sensory areas are engaged in stimulus-selective persistent firing during working memory states (Fig. 7C). The standard deviation of firing rate shows a non-monotonic pattern as a function of the hierarchy (Fig. S8A).

We carried out the autocorrelation analysis of stochastic persistent activity time series from each area, and found that neuronal fluctuations are fast in the brain areas at the high and low hierarchical positions; by contrast, mnemonic firing of brain areas around the hierarchical bifurcation region shows fluctuations on slower timescales, exemplified by Brodmann area 5, which is part of the posterior parietal cortex (Fig. 7D, lower panel; see also Fig. S8F, right). The brain area 5, 2, and $F1$ have a time constant around 10^3 ms.

To compare with the previous experimental result about the time constant of each brain area during a baseline state, we checked the time constant of the resting state, which is roughly a monotonic function of the hierarchy (Fig. 7D, upper panel, see also

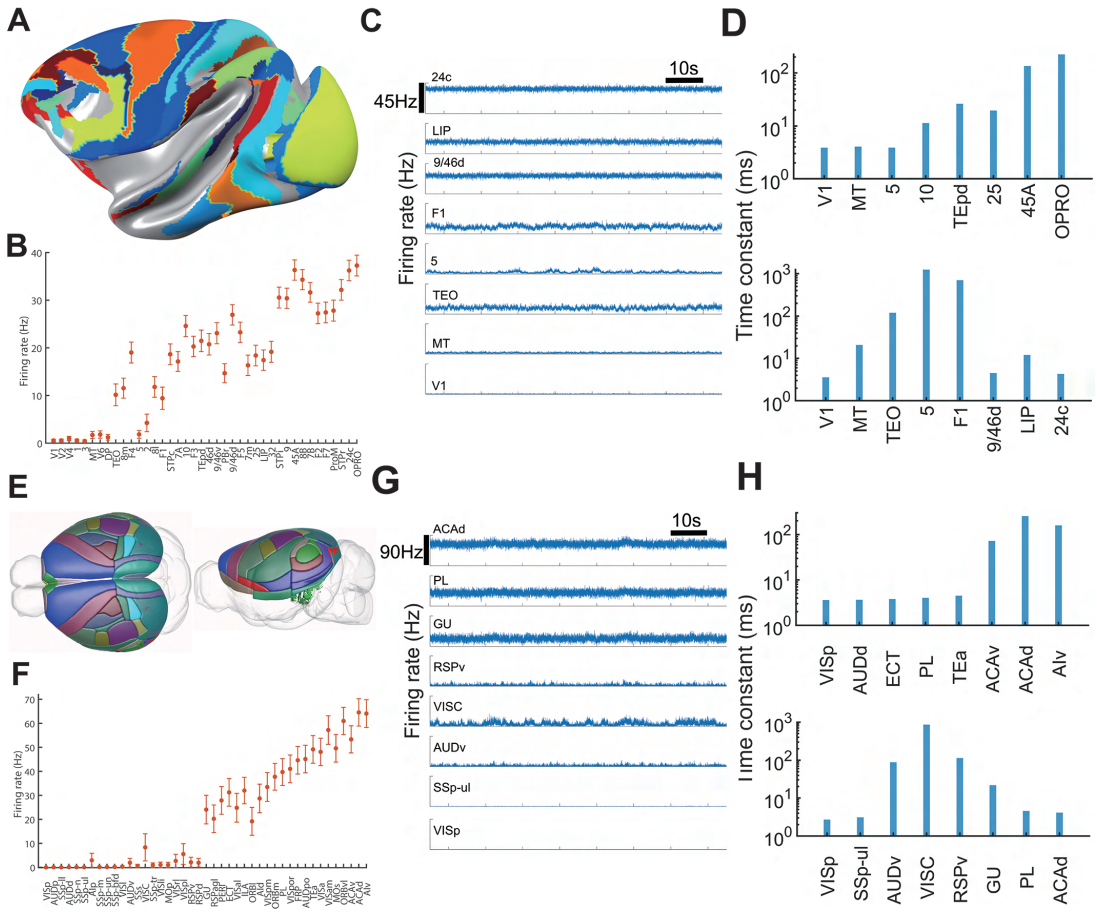

Figure 7: Bifurcation in space of connectome-based cortical models of macaque monkey (A-D) [24] and mouse (E-H) [25]. (A) Lateral view of macaque neocortex surface with model areas in color. (B) Firing rate of 40 brain areas, ranked by the hierarchical position. (C) Firing rate time series of 8 chosen brain areas when neocortex model is in a delay period working memory state. (D) The bar figure of time constants of 8 selected brain areas for resting (upper panel) and delay period working memory (lower panel) state. (E) Superior and lateral view of mouse cortex surface with model areas in color. (F) The firing rate of 43 brain areas with noise (brown) is ranked by the hierarchical position. (G) Firing rate time series of 8 chosen brain areas when large-scale mouse model in a working memory state. (H) The bar figure of time constants of 8 selected brain areas in resting (upper panel) and delay period working memory (bottom panel) state.

Fig. S8F, left), which is consistent with previous modeling [40] and experimental [45] results.

We then asked if critical slowing down also occurs in the connectome-based large-scale mouse cortex model [25]. The model contains 43 cortex areas in the common coordinate framework v3 atlas [46] (Fig. 7E) with a quantified hierarchy. There is a macroscopic gradient of synaptic inhibition mediated by parvalbumin-expressing interneurons [47] that decreases along the hierarchy. The mouse model also exhibits the coexistence of a resting state and persistent activity states appropriate for working memory function (Fig. 7F). Once again, we performed autocorrelation analysis of mnemonic persistent activity and found a qualitatively similar inverted-V shaped profile of time constant (Fig. 7H, lower panel, see also Fig. S8, H right).

It is worth noting that our model of the macaque monkey cortex is presently limited to a subset of areas for which the connectomic data are available, the precise hierarchical positions (normalized between 0 and 1) could be slightly modified in a complete model of all cortical areas. Moreover, exactly which area displays maximal time constant of mnemonic firing fluctuations, may depend 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

In this work, we chose working memory of central importance to cognition and behavioral flexibility to investigate how functional modularity emerges in a multiregional cortex. One may question whether there is indeed modularity for working memory, given that visual working memory content can be decoded from functional MRI measurements in

the primary visual cortex [48, 49]. However, the interpretation of fMRI results remains controversial [50]. By contrast, there is ample evidence at the single cell level that a subset of cortical areas are involved in working memory maintenance [15]. Regardless, the debate does not directly bear on the present work which addresses the question of how functional modularity, *if* it is present, may emerge under the assumption that the cortex is made of repeats of a canonical circuit.

We found that the mechanism is mathematically described as a novel form of bifurcation, that occurs at some critical location in the spatially embedded cortex. The idea of a neural system to operate near a criticality has been proposed [51, 52], among open questions are what are the signatures of criticality, whether fine-tuning of parameters is required or can be realized through a self-organized mechanism [53]. A bifurcation in space is robust: parameter changes would merely move the spatial location of bifurcation. Moreover, it 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 stores sensory information, another maintains a behavioral rule that guides sensorimotor mapping etc. Each of these is modularly organized in the sense of selectively engaging a subset of areas but not the others, mathematically described by its own bifurcation in space. In other words, there are many bifurcations in space in a given large-scale cortical system.

An observable manifestation of bifurcation in space is critical slowing down near the transition. Consequently, along the cortical hierarchy, there is an inverted V-shaped pattern of time constants that dominate neural fluctuations during working memory. This is in contrast to our previous report that during a resting state, the dynamical timescale roughly increases along the cortical hierarchy [40, 3]. The difference is explained by the fact that time constants are uniquely defined mathematically only for a linear dynamical

system; they depend on the internal state of a highly nonlinear system. This 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 an abstract model of cortex endowed with experimentally measured connection statistics are confirmed in connectome-based models of the macaque cortex and mouse cortex, opening the door to test the predicted inverted V-shaped profile of time constants at specific areas during working memory. In particular, for working memory of visual motion information, the work in [17] 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 [54], the inverted V-shaped timescale profile is likely to hold during a decision process, a proposal in line with the existing evidence that neurons in the posterior parietal cortex display longer integration times underlying accumulation of information than both sensory areas and the prefrontal cortex, located lower and higher hierarchical positions, respectively [55]. Testing this model prediction requires the following considerations. First, time constant estimates may vary in different behavioral epochs and tasks, here we focus on an internal state independently of external inputs during a mnemonic delay. Second, there is heterogeneity of time constants across single cells within an area, therefore sufficient statistics is needed for a cross-area comparison. Third, a mnemonic delay period needs to be much longer than to-be-assessed autocorrelation times. Fourth, critical slowing down is manifest near a criticality; the number of cortical areas is limited and it remains to be seen experimentally how close one can get to a bifurcation locus in the cortical system.

This work focuses on spatial patterns of modular neural representations mathematically described as attractor states. The concept is not limited to simple patterns of higher

versus low firing rates which are discussed in this work merely for the sake of simplicity. As a matter of fact, monkey and mouse experiments showed that in a recorded cortical area, during working memory some neurons increase firing while others reduce firing such that the totality of neural population activity is roughly the same as in the resting state [56, 14]. The principle of bifurcation in space is applicable to more complex spatial patterns of neural activity. Moreover, attractors can display complex temporal dynamics such as chaos rather than steady states [26]. For instance, neural representation of working memory often involves stochastic oscillations [57, 58]. As discussed elsewhere [14], the attractor paradigm can well be consistent with considerable temporal variations of neuronal delay period activity as well as cell-to-cell heterogeneities. Future research is needed to extend the concept of bifurcation in space beyond steady states.

In general, bifurcation in space could underlie a sudden appearance of new behavior in a region of 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 which makes it all the more remarkable that criticality can occur locally in a multiregional cortex. We rigorously established the concept of bifurcation in space using the normal form theory of bifurcation. A *recurrent* deep neural network (a hierarchical cortex with many feedback loops) and macroscopic gradients are sufficient to give rise to various spatial distributed persistent activity states, thus several functionally modular networks in our model. Research along these lines should broadly help us explain the emergence of novel brain capabilities that are instantiated in certain parts of the brain merely as a result of quantitative changes of properties, providing a mechanistic foundation for functional modularity.

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

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

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

[9] 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).

[10] 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.

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

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

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

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

[15] 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).

[16] 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).

[17] 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).

[18] von Economo, C. *The Cytoarchitectonics of the Human Cerebral Cortex* (London: Oxford University Press, 1929).

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

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

[21] 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).

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

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

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

[25] Ding, X., Froudist-Walsh, S., Jaramillo, J., Jiang, J. & Wang, X.-J. Predicting distributed working memory activity in a large-scale mouse brain: the importance of the cell type-specific connectome. *bioRxiv* doi: https://doi.org/10.1101/2022.12.05.519094 (2022).

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

[27] 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).

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

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

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

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

[32] Klatzmann, U. *et al.* A connectome-based model of conscious access in monkey cortex. *bioRxiv* doi: https://doi.org/10.1101/2022.02.20.481230 (2022).

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

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

[35] 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).

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

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

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

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

[40] 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).

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

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

[43] Kuznetsov, Y. A., Kuznetsov, I. A. & Kuznetsov, Y. *Elements of applied bifurcation theory*, vol. 112 (Springer, 1998).

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

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

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

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

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

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

[50] Xu, Y. Reevaluating the sensory account of visual working memory storage. *Trends Cogn. Sci.* **21**, 794–815 (2017).

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

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

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

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

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

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

[57] 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).

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

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

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

Acknowledgments

We thank Jorge Mejias and Xingyu Ding for help with cortical model codes of the macaque and mouse, respectively. **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.).

Author contributions: X.-J.W. conceived the project; X.-J.W., J.J., and U.P.-O. designed the research; J.J. and U.P.-O. performed the research with supervision and inputs from X.-J.W.; X.-J.W., J.J., and U.P.-O. wrote the paper. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All data and code will be available after published.

Materials and Methods

Generative model for the mammalian cortical connectivity.

We use the model in [27] to generate multiple cortical network realizations. Briefly, in this model, we start by randomly choosing the center of N brain areas in a three-dimensional ellipsoid. After that, the ellipsoid is parcellated into N areas through a Voronoi partition. Then, axon growth starts by randomly choosing a source in the ellipsoid. The direction of the growth is determined by the summing force of all the areal centers. The growth length is randomly chosen from an exponential distribution for modeling the previously reported distance effects on the connectivity [28]. Since the axon's source, direction, and length are determined, we can find the axon's target position in the ellipsoid. After that, we 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 generated network has not only a similar in- and out-degree distribution with the actual macaque monkey brain network, which is measured using retrograde tract-tracing methods, but also a similar triad distribution. In this model, an ellipsoid better fits the connectivity data than a two-dimensional spheroid [27].

Diffusion map method for connectivity embedding.

We analyze the generated connectivity using the diffusion map method [37]. This is a class of non-linear dimensionality reduction that has been recently applied to human, and macaque monkey connectomes [21]. Briefly, this method assumes a hypothetical *diffusion processes* on the nodes of the symmetric version of the generated network connectivity (i.e., the FLN matrix). This diffusion process generates a *diffusion metric space* where the distance between cortical areas can be defined. In this diffusion space, closer areas

share a larger number of loops connecting 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 the 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 $FLN + FLN^T$.

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): Euclidian and Hyperbolic hierarchies. To calculate either class of hierarchy value, we first choose the cortical area with the smallest value in the first principal gradient as the first area in the hierarchy or origin area. This choice is arbitrary. To determine the hierarchical position, we compute the distance in diffusion space between each cortical area with the origin area. For the Euclidean hierarchy, the hierarchical value of a cortical area i is computed using the normalized Euclidean distance $h^i_{Euc} = dist^{i0}_{Euc}/dist^{max}_{Euc}$. Here the value $dist^{i0}_{Euc}$ is the Euclidean distance between brain area i and the origin area, and the value $dist^{max}_{Euc}$ is the maximum Euclidean distance of all the brain areas to origin area. The ranked Euclidean hierarchical position of each brain area is shown as the brown circles in Fig. 1D.

For the hyperbolic distance, we estimate the distance with 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 weights of the links (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 [59], which approximated spiking neural network with AMPA, GABA, and NMDA synapses [60]. The dynamical equations that describe the dynamics for a single cortical area are described as follows:

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

where, S_E and S_I are the gating variable of NMDA receptor of the excitatory population and the gating variable of 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 excitation 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 [42]

$$\phi_{exc}(I) = \frac{aI - b}{1 - e^{-d(aI-b)}}. \tag{2}$$

2. Threshold-linear function

$$\phi_{exc}(I) = [aI - b]_+. \tag{3}$$

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

$$\phi_{inh}(I) = [c_1 I - c_0]_+ \tag{4}$$

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 (1-4) using the connectivity from our generative model of the mammalian neocortex. The long-range projections in our model are from excitatory to excitatory populations [40, 23, 24]. Our large-scale model is described as follows

$$
\begin{aligned}
\tau_E \frac{dS_E^i}{dt} &= -S_E^i + \gamma_E \tau_E (1 - S_E^i) r_E^i, \\
\tau_I \frac{dS_I^i}{dt} &= -S_I^i + \gamma_I \tau_I r_I^i, \\
\tau_r \frac{dr_E^i}{dt} &= -r_E^i + \phi_{exc}(J^i(W_{EE} S_E^i + \mu_{EE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{EI} S_I^i + I_{noi}^i + I_{ext,E}^i), \\
\tau_r \frac{dr_I^i}{dt} &= -r_I^i + \phi_{inh}(J^i(W_{IE} S_E^i + \mu_{IE} \sum_{j=1}^{N} FLN_{ij} S_E^j) - W_{II} S_I^i + I_{ext,I}^i), \\
\tau_r \frac{dI_{noi}^i}{dt} &= -I_{noi}^i + \sqrt{\tau_r \sigma_{noi}^2} \xi^i,
\end{aligned} \tag{5}
$$

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, which is generated as described in the previous section. The parameter J^i corresponds to the gradient of excitation. This 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 $I_{ext,E}^i = I_{ext,E}$ and $I_{ext,I}^i = I_{ext,I}$. The noise term I_{noi} is an Ornstein-Uhlenbeck process, representing the AMPA synaptic noise with a short time constant $\tau_r = 2ms$ [59]. 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 $\frac{dr_I}{dt} = 0$. Thereafter, we have the steady state equations as follows:

$$\begin{aligned}
\frac{-S_E}{\tau_E} + \gamma_E(1 - S_E)r_E &= 0, \\
\frac{-S_I}{\tau_I} + \gamma_I r_I &= 0, \\
-r_E + \phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\
-r_I + \phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0.
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
\frac{-S_E}{\tau_E} + \gamma_E(1 - S_E)\phi_{exc}(JW_{EE}S_E - W_{EI}S_I + I_{ext,E}) &= 0, \\
\frac{-S_I}{\tau_I} + \gamma_I\phi_{inh}(JW_{IE}S_E - W_{II}S_I + I_{ext,I}) &= 0.
\end{aligned} \tag{7}$$

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

$$\begin{aligned}
S_I = \gamma_I\tau_I(c_1 I_{inh,t} - c_0) &= \alpha c_1 JW_{IE}S_E + \alpha(c_1 I_{ext,I} - c_0), \\
\alpha = \frac{\gamma_I\tau_I}{1 + \gamma_I\tau_I c_1 W_{II}} &= \frac{1}{\frac{1}{\gamma_I\tau_I} + c_1 W_{II}},
\end{aligned} \tag{8}$$

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 (8) into the steady state S_E equation (7). After this manipulation, the steady state of the single cortical area is determined by the NMDA gating variable S_E as follows:

$$-S_E + \gamma_E \tau_E (1 - S_E)\phi_{exc}(\alpha_1 S_E + \alpha_2) = 0,$$

$$\alpha_1 = J(W_{EE} - \alpha c_1 W_{EI} W_{IE}),$$

$$\alpha_2 = I_{ext,E} - \alpha W_{EI}(c_1 I_{ext,I} - c_0). \tag{9}$$

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 (3), 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 hierarchy 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

$$S_E = \frac{-(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E}) \pm \sqrt{(a(\alpha_1 - \alpha_2) + b - \frac{1}{\gamma_E \tau_E})^2 - 4(-a\alpha_1)(a\alpha_2 - b)}}{2(-a\alpha_1)}, \tag{10}$$

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 critical hierarchy 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 hierarchy value 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 (2). By combining equation (2) and equation (9) the steady state is given by

$$-S_E(1-e^{-d(a\alpha_1 S_E+a\alpha_2-b)})+\gamma_E\tau_E(1-S_E)(a\alpha_1 S_E+a\alpha_2-b)=0. \tag{11}$$

The steady states equation (11) is highly nonlinear, and we can not provide an analytic solution. Instead, we solve equation (11) using the Matlab numerical solver *fsolve*. The steady state in equation (11) 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:

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

We assume stable steady states of 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 large-scale network model as

$$\frac{-S_E^i}{\tau_E} + \gamma_E(1 - S_E^i)\phi_{exc}(J^i(W_{EE}S_E^i + \mu_{EE}L_E^i) - W_{EI}S_I^i + I_{ext,E}^i) = 0,$$

$$\frac{-S_I^i}{\tau_I} + \gamma_I\phi_{inh}(J^i(W_{IE}S_E^i + \mu_{IE}L_E^i) - W_{II}S_I^i + I_{ext,I}^i) = 0. \quad (13)$$

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

$$S_I^i = \gamma_I\tau_I(c_1 I_{inh,t} - c_0) = \alpha c_1 W_{IE} J^i S_E^i + \alpha c_1 \mu_{IE} J^i L_E^i + \alpha(c_1 I_{ext,I}^i - c_0), \quad (14)$$

where the definition of α is same as in equation (8). Therefore, for the ith cortical area the excitatory gating variable S_E^i obeys the following steady state equation

$$-S_E^i + \gamma_E\tau_E(1 - S_E^i)\phi_{exc}((W_{EE} - W_{EI}\alpha c_1 W_{IE})J^i S_E^i$$
$$+(\mu_{EE} - W_{EI}\alpha c_1\mu_{IE})J^i L_E^i + (I_{ext,E}^i - W_{EI}\alpha(c_1 I_{ext,I}^i - c_0))) = 0. \quad (15)$$

First, we will analyze the steady state equation (15) for the case when the transfer function is threshold-linear. The above equation (15) can be written as the steady state of the following set of dynamical equations

$$\frac{dS_E^i}{dt} = f(S_E^i, L_E^i, J^i) \quad (16)$$
$$= -\gamma_E\tau_E\chi_1 J^i (S_E^i)^2 + (\gamma_E\tau_E\chi_1 J^i - \gamma_E\tau_E(\chi_2 J^i L_N^i + \chi_3) - 1)S_E^i$$
$$+ \gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3),$$

with

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

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 (16) 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 (10)). However, in our large-scale network model, the quadratic formula of the network model is also dependent on the hierarchy value through J^i and the long-range excitatory inputs L_E^i. Therefore, the bifurcation happening in the hierarchical space is determined by the following expression:

$$(\gamma_E\tau_E\chi_1 J^i - \gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3) - 1)^2 + 4(\gamma_E\tau_E\chi_1 J^i)(\gamma_E\tau_E(\chi_2 J^i L_E^i + \chi_3)) \tag{17}$$

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

$$+ (1 + 2\gamma_E\tau_E(\chi_3 + \tau_E\chi_3^2) + 2\gamma_E\tau_E\chi_1(\tau_E\chi_3 - 1)J^i + \gamma_E(\tau_E\chi_1 J^i)^2) \tag{18}$$

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

$$+ 2\gamma_E\tau_E(-\chi_1 + \tau_E\chi_1\chi_3 + (\chi_2 + \tau_E\chi_1\chi_3)L_E^i)J^i + (1 + \gamma_E\tau_E^2\chi_3^2 + 2\gamma_E\tau_E\chi_3) \tag{19}$$

$$= 0,$$

where $J^i = 1 + \eta h^i$ and L_E^i are the scaled hierarchy value and long-range excitatory inputs of ith brain area, respectively. The equation (17) is a constrain equation in the two-dimensional space of hierarchy h and long-range excitatory inputs L_E. Therefore, equation (17) 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 (18)), 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 in an active state only when it has long-range excitatory inputs such that $L_E^i > L_E^*$. From another viewpoint, the bifurcation equation could be a quadratic equation of the scaled hierarchical value J^i (eq. 19). For a ith brain area with long-range excitatory inputs L_E^i, only when it has a hierarchical position $J^i > J^*$ it displays non-zero firing rates. The critical line given by equation (17) 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:

$$\begin{aligned}
&\frac{dS_E^i}{dt} = f(S_E^i, L_E^i, J^i) \\
&= -S_E^i(1 - e^{-d(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3)}) \\
&+\gamma_E \tau_E (1 - S_E^i)(\chi_1 J^i S_E^i + \chi_2 J^i L_E^i + \chi_3) = 0.
\end{aligned} \tag{20}$$

We use numerical methods to solve equation (20). Numerically solving the equation (20) will give a steady state surface shown in Fig. 4D, Fig.S4, Fig.S5A-B and Fig. 6. 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 very similar geometry as the cusp bifurcation normal form solution surface [43]. 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 [43, 26]. The cusp normal form solution surface is given by the set of solutions to the steady state equation $\beta_1 + \beta_2 x - x^3 = 0$ 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 [43, 26], and with this, its bifurcation diagram. The cusp bifurcation point is given by $\beta_1 = \beta_2 = 0$. In Fig. S5A-B, for illustration proposes, we overlay a β_1 and β_2

axes to highlight the resemblance of our network's solution surface with the cusp surface [43].

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 largest. 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 [43], 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 (17), we have the bifurcation point (S_E^*, L_E^*, J^*) fulfill the below equation.

$$\begin{aligned} &f(S_E^*, L_E^*, J^*) = \\ &(\gamma_E \tau_E \chi_1 J^* - \gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3) - 1)^2 \\ &+4(\gamma_E \tau_E \chi_1 J^*)(\gamma_E \tau_E (\chi_2 J^* L_E^* + \chi_3)) = 0. \end{aligned} \tag{21}$$

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 (17) change from complex conjugate to real. This point in parameter space corresponds to the point of 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:

$$
\begin{aligned}
& f(S_E^i, L_E^i, J^i) = f(S_E^*, L_E^*, J^*) \\
& + \begin{pmatrix} \frac{\partial f}{\partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial f}{\partial J^i}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
& + \frac{1}{2} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix}^T \begin{pmatrix} \frac{\partial^2 f}{\partial (S_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial S_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial L_E^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (L_E^i)^2}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial L_E^i \partial J^i}\Big|_{S_E^*, L_E^*, J^*} \\ \frac{\partial^2 f}{\partial J^i \partial S_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial J^i \partial L_E^i}\Big|_{S_E^*, L_E^*, J^*} & \frac{\partial^2 f}{\partial (J^i)^2}\Big|_{S_E^*, L_E^*, J^*} \end{pmatrix} \begin{pmatrix} S_E^i - S_E^* \\ L_E^i - L_E^* \\ J^i - J^* \end{pmatrix} \\
& + O(3),
\end{aligned}
\tag{22}
$$

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 (22) obtaining

$$
\begin{aligned}
\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) = \zeta_1 (S_E^i)^2 + \zeta_2 S_E^i + \zeta_3, \\
\zeta_1^i &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^i &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} L_E^i + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^i &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} L_E^i + J^i L_E^{i,*} - J^i L_E^i) \\
&\quad -\gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} + \gamma_E \tau_E \chi_2 (J^i L_E^i - J^{i,*} L_E^{i,*})),
\end{aligned} \tag{23}
$$

The above equation (23) corresponds to the bifurcation in hierarchy space normal form. We solve the steady state of the above equation (23) self-consistently and predict the firing rate of the delay period working memory states. The self-consistent equations read as

$$
\begin{aligned}
S_E^i &= \frac{-\zeta_2^{i,sce} - \sqrt{(\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce}}}{2\zeta_1^{i,sce}} \\
\zeta_1^{i,sce} &= -\gamma_E \tau_E \chi_1 J^{i,*}, \\
\zeta_2^{i,sce} &= (-2\gamma_E \tau_E \chi_1 S_E^{i,*}(J^i - J^{i,*}) - \gamma_E \tau_E \chi_2 L_E^{i,*}(J - J^{i,*}) \\
&\quad -\gamma_E \tau_E \chi_2 J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + \gamma_E \tau_E \chi_1 J^i - (1 + \gamma_E \tau_E \chi_3)), \\
\zeta_3^{i,sce} &= (-\gamma_E \tau_E \chi_1 J^{i,*}(S_E^{i,*})^2 + \gamma_E \tau_E \chi_2 S_E^{i,*}(J^{i,*} \sum_{j=1}^{N} FLN_{ij} S_E^j + J^i L_E^{i,*} \\
&\quad -J^i \sum_{j=1}^{N} FLN_{ij} S_E^j) - \gamma_E \tau_E \chi_1 J^{i,,*} S_E^{i,*} + (1 + \gamma_E \tau_E \chi_3) S_E^{i,*} \\
&\quad +\gamma_E \tau_E \chi_2 (J^i \sum_{j=1}^{N} FLN_{ij} S_E^j - J^{i,*} L_E^{i,*})), \\
\Delta^i &= (\zeta_2^{i,sce})^2 - 4\zeta_1^{i,sce}\zeta_3^{i,sce} \geq 0, \\
S_E^i &\geq 0.
\end{aligned} \tag{24}
$$

To solve the self-consistent equations, first, we solve equation (19) 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 (16) and solving numerically the equation (16). Third, we insert (S_E^*, L_E^*, J^*) into the self-consistent equations in equation (24). Lastly, we solve the self-consistent equations in equation (24) and then predict the firing rate pattern of activity during an active state as shown in Fig. 4B. The normal form in equation (23) can be further reduced to a more general expression as reads below

$$\begin{aligned}\frac{dS_E^i}{dt} &= f(S_E^i, L_E^i, J^i) \\ &= a_1(S_E^i)^2 + (a_2 J^i + a_3 L_E^i + a_4)S_E^i + (a_5 J^i + a_6 L_E^i + a_7 J^i L_E^i + a_8), \\ & a_1, \ldots, a_8 \in R, \end{aligned} \tag{25}$$

where $a_1, \ldots, a_8$ are parameters calculated re-arrenging terms in equation (23). 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 [43]. 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 finding active states.

In a network between 1000-10000 brain areas, it is infeasible for our computational resources to find all the possible active states. Therefore, we try to get 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 all the 1000 cortical areas into 20 groups along the hierarchical position. Therefore, each group has 50 cortical regions that are 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 (13) 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

$$\begin{aligned} S_E^i = & \frac{\tau_E \gamma_E \phi_{exc}(J^i(W_{EE} S_E^i + \mu_{EE} L_E^i) - W_{EI} S_I^i + I_{ext,E}^i)}{1 + \tau_E \gamma_E \phi_{exc}(J^i(W_{EE} S_E^i + \mu_{EE} L_E^i) - W_{EI} S_I^i + I_{ext,E}^i)}, \\ S_I^i = & \tau_I \gamma_I \phi_{inh}(J^i(W_{IE} S_E^i + \mu_{IE} L_E^i) - W_{II} S_I^i + I_{ext,I}^i), \end{aligned} \tag{26}$$

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 (26) until the mean absolute difference between two consecutive iterations is smaller than 10^{-10}, or the iteration number is large 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 had before. Once the sum of the absolute difference of all the brain areas is large than 0.05, we would define it as a new unique state and keep it. Through this process, we obtained 4333 (one of the states in Fig. 5A is the resting state) distinct active states after trying $1,048,576$ initial states. For all the distinct active states, we checked the local stability of the state by calculating all the eigenvalues of the Jacobian matrix. We found that the real part of all eigenvalues in all the active states is negative. Therefore, all the states are locally stable.

Table 1. Parameters for Numerical Simulations		
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. S4 Fig. S5A Fig. S6B)
$c1$, $c0$	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), 29pA (Fig. S2H left), 8pA (Fig. 5B), 6pA (Fig. S7A), 10pA (Fig. S7B), 16pA (Fig. S7C)

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-expoential function:

$$ae^{\frac{-\Delta T}{\tau}} + c, \tag{27}$$

Double-expoential function:

$$ae^{\frac{-\Delta T}{\tau_1}} + (1-a)e^{\frac{-\Delta T}{\tau_2}} + c, \tag{28}$$

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 function 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.

We determine the final time constant of each brain area based on the root-mean-square error (RMSE) of the fitting. 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.

Supplementary figures

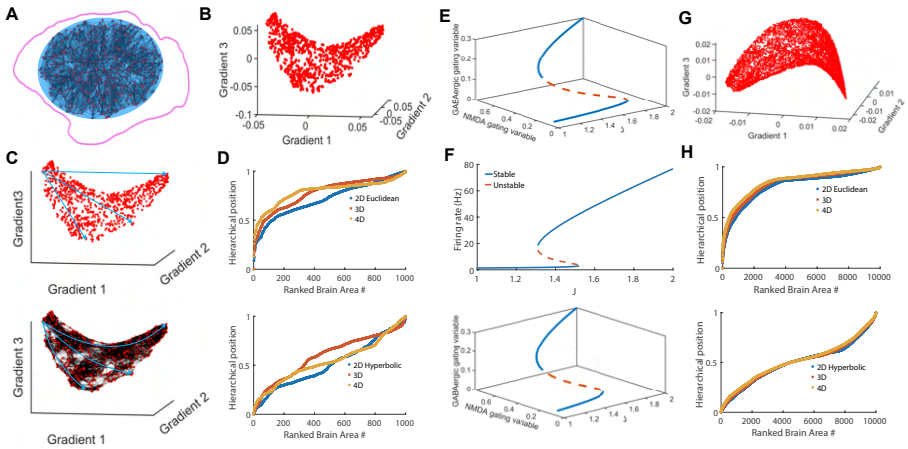

Figure S1: **The supplementary figure of generated connectivity, hierarchy, and local circuit model.** (A) The generated super neocortex network with 1000 brain areas is consistent with the macaque neocortex network statistic. (B) The three-dimensional embedding hyperbolic shapes of generated super neocortex 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 two-dimensional, three-dimensional, and four-dimensional embedding of generated super neocortex model. (E) The NMDA and GABAergic gating variables' steady states vary with the hierarchical position of an isolated brain area with simplified dynamics. Additionally, this isolated brain area has the same parameter settings as Panel F of Figure 1 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 generated super neocortex model with 10000 brain areas.

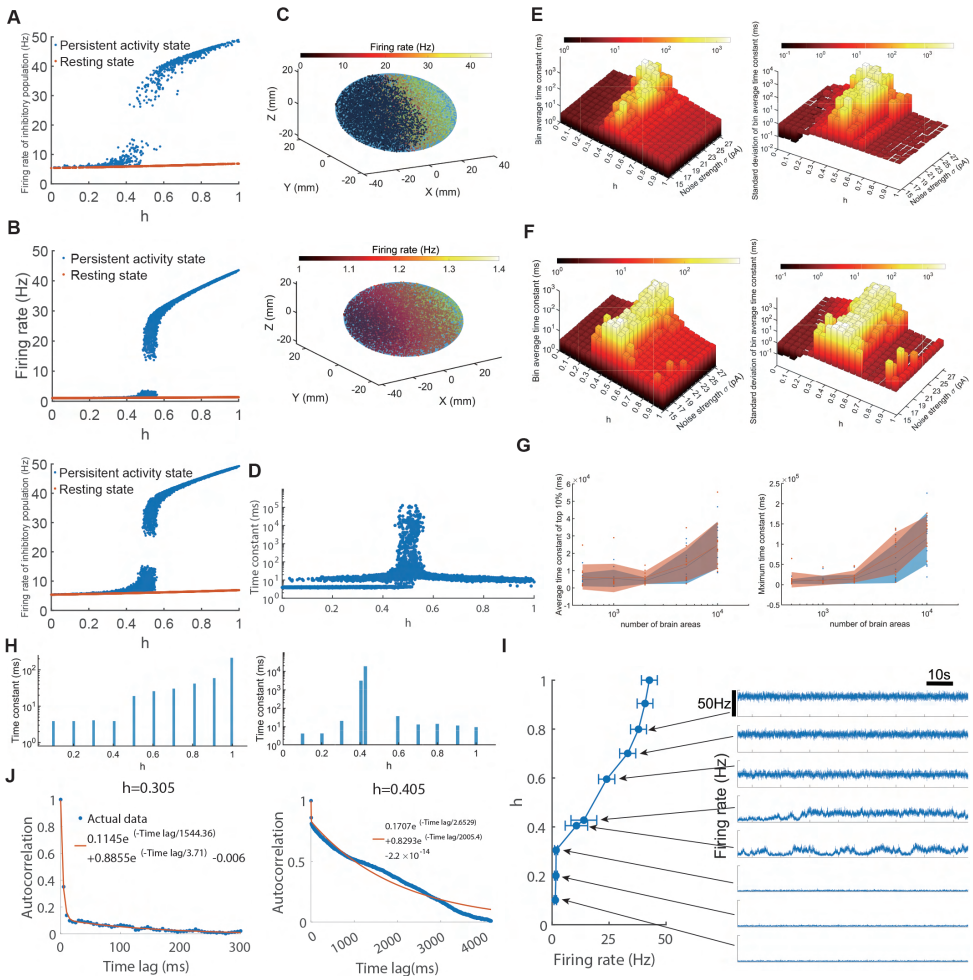

Figure S2: **The supplementary figure of bifurcation in hierarchical space.** (A) Firing rate of all the inhibitory populations for the same parameter set of panel A of Fig. 2 of the main text. (B) The firing rate of the excitatory (upper) and inhibitory (bottom) population of both active (blue) and 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 at 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.

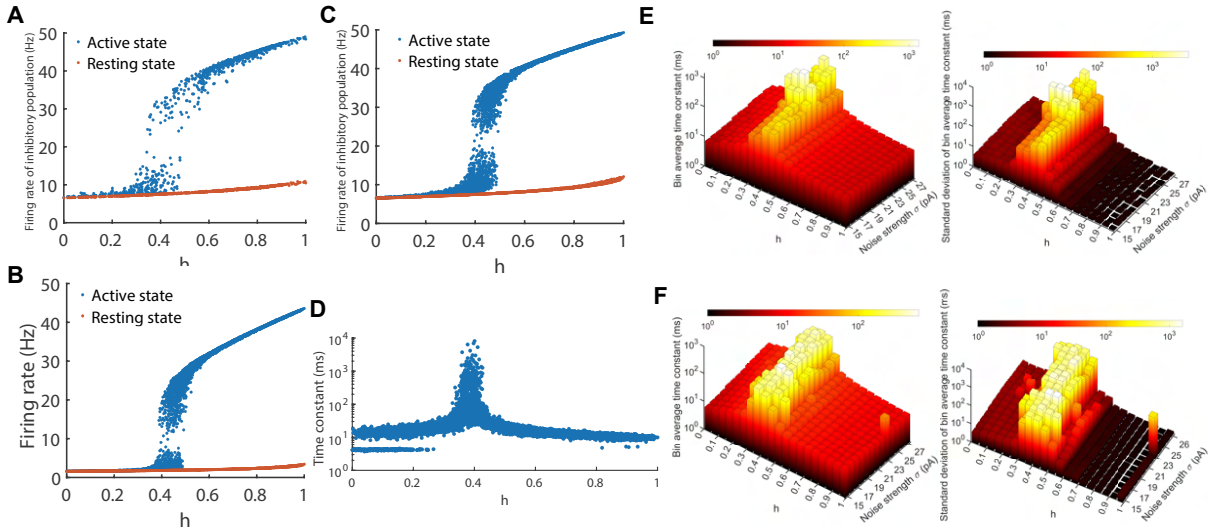

Figure S3: **The delay activity state and time constant for brain networks with** 10000 **brain areas and** $d = 0.157$**.** (A) Firing rate of all the inhibitory populations for the same set of panel A of Fig.3 of the main text with $d = 0.157$. (B) The firing rate of the excitatory population of both 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.

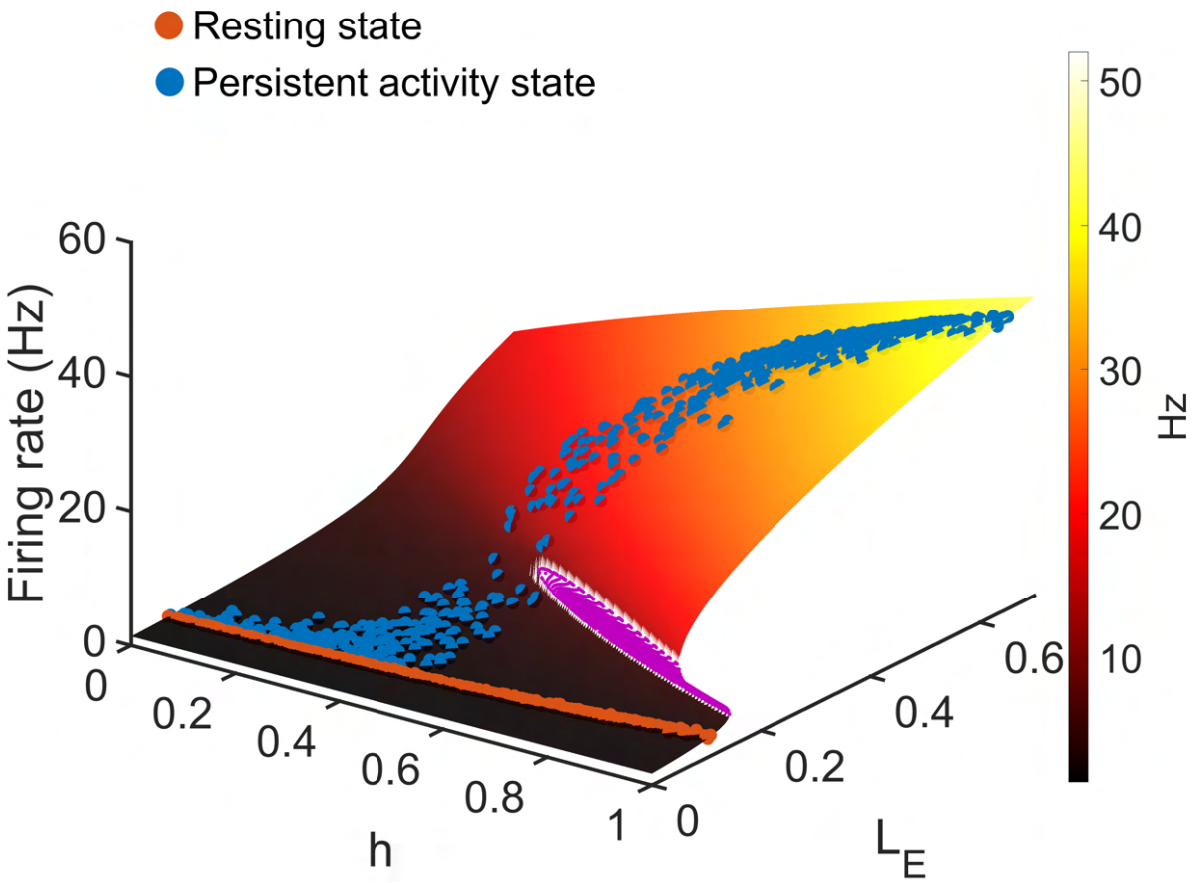

Figure S4: **The solution surface of $d = 0.157$.** The neocortex model's resting (brown) and persistent activity state (blue) lie on top of the solution surface ($d = 0.157$). In this case, the active state has a continuous transition without a firing rate gap.

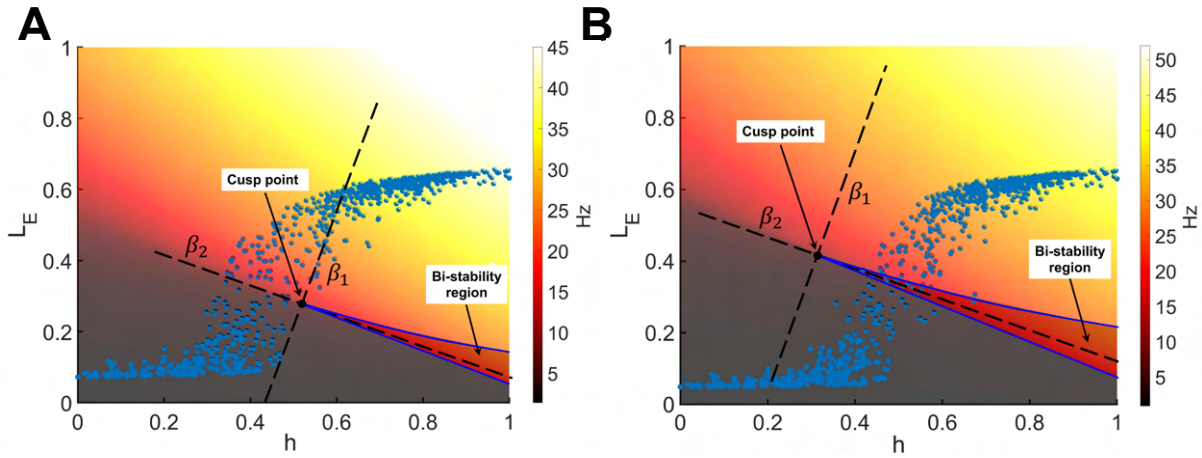

Figure S5: **Cusp geometry determines the bifurcation in hierarchical space.** (A and B) The geometry of the solution surface and the cusp point determine the bifurcation in hierarchical space. The blue dots correspond to the persistent activity state of the neocortex model with $d = 0.157$ (panel A) 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, for hierarchy values h and long-range excitatory input current L_E in the ranges $0.6 - 1$ and $0.1 - 0.2$, respectively, the solution surface is folded.

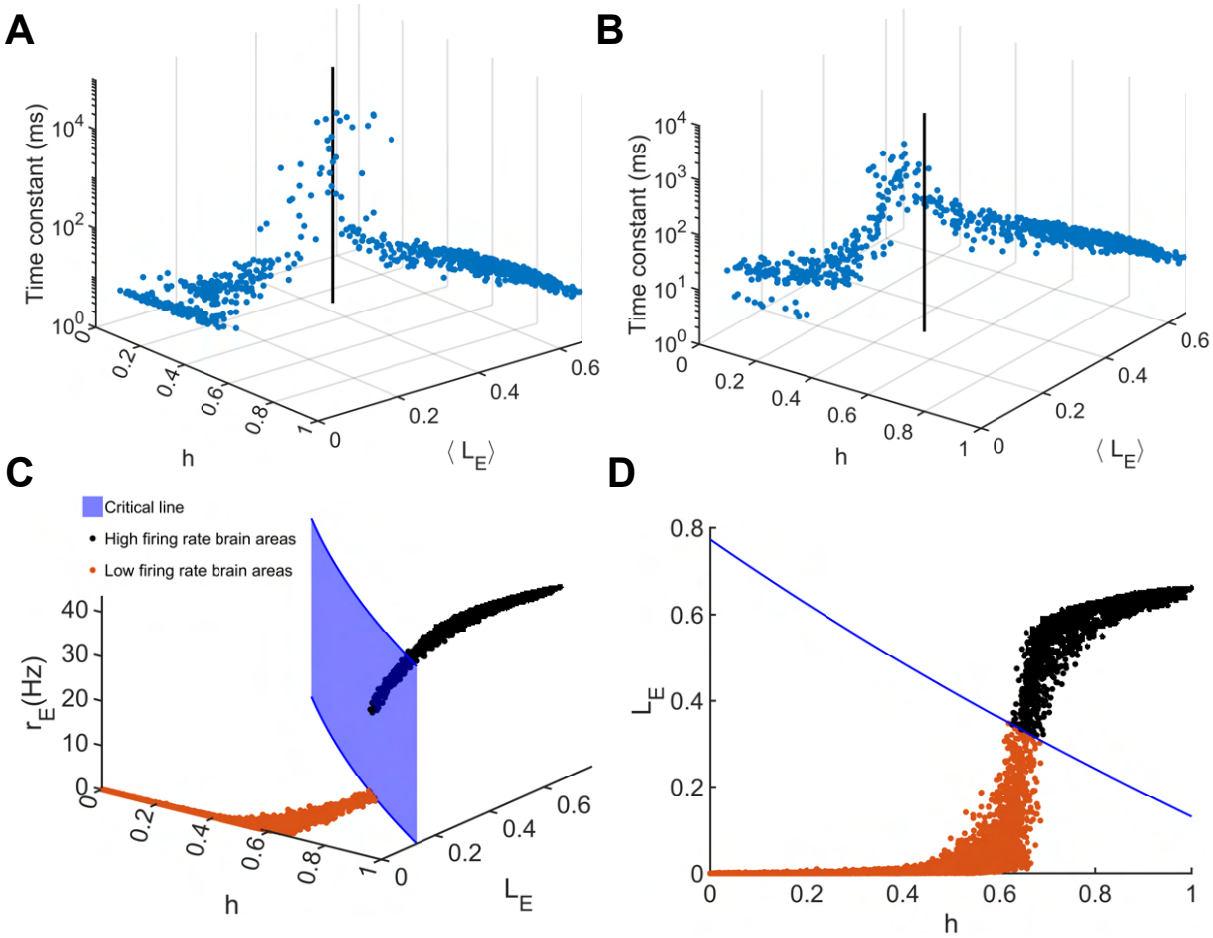

Figure S6: **Time constant around cusp point and geometry of solution surface of network with 10000 brain areas with threshold-linear transfer function.** (A-B) The time constant of all the brain areas arrange in the h and $\langle L_E \rangle$ space for $d = 0.17$ and $d = 0.157$ with 1000 brain areas, 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 D of Fig.2 and S3, 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.

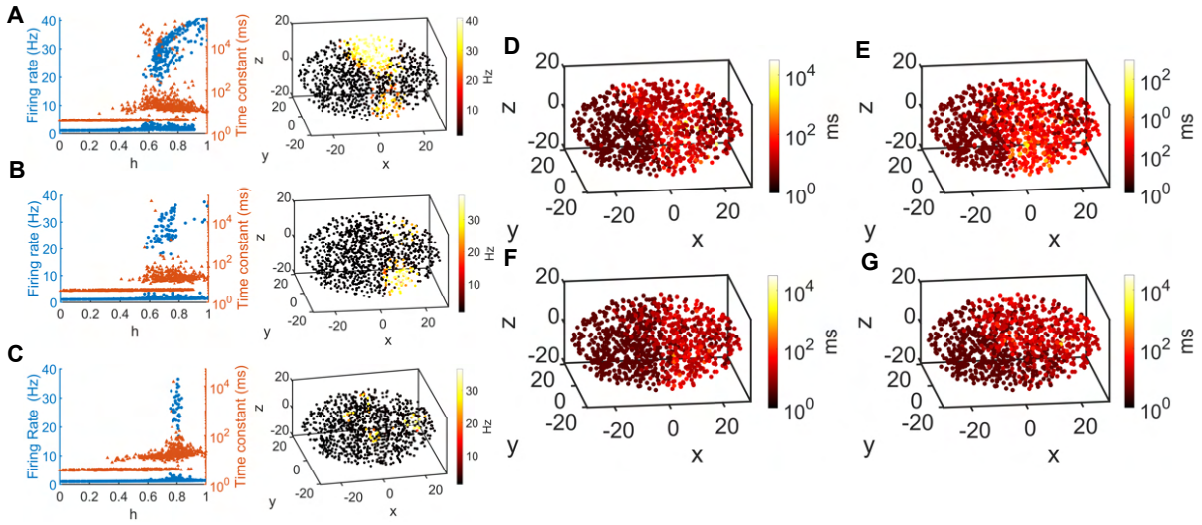

Figure S7: **The supplementary figure of the diversity of distributed working memory states.** (A-C) The firing rate (blue) and time constant (brown) of active state $S1$, $S3$, and $S4$ (left) for each cortical area, respectively. The spatial distribution of active state $S1$, $S3$, and $S4$ (right) firing rate in the generative model ellipsoid, respectively. (D-E) The spatial distribution of active state $S1$, $S2$, $S3$, and $S4$ time constants in the generative model ellipsoid, respectively.

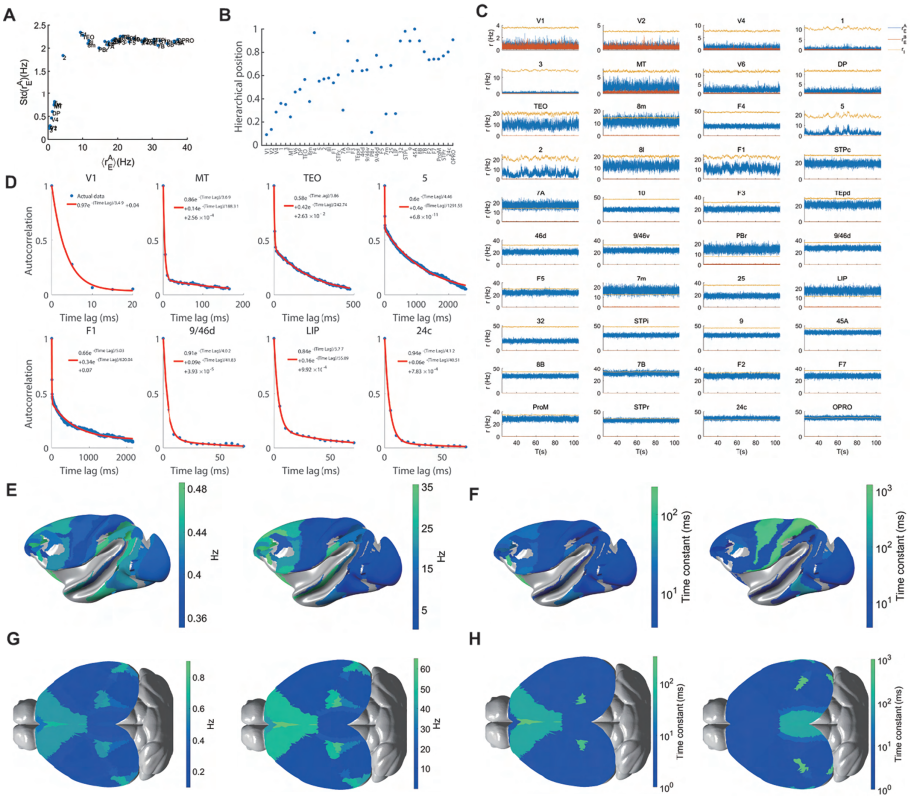

Figure S8: **The supplementary figure of bifurcation in the hierarchical space of connectome-based cortical models of macaque monkey and mouse.** (A) The average firing rate versus the standard deviation of each brain area of the macaque neocortex with 40 brain areas. (B) The normalized hierarchical position or spine count data of the 40 brain areas. (C) The firing rate of excitatory population A (blue), B (brown), and inhibitory population (yellow) of the 40 brain areas. (D) The autocorrelation function and the related single or double exponential fitting function of the eight chosen brain areas in panel D of Fig.1 in the main text. (E) Spatial activity map of resting state (left) and the monotonic delay period working memory state (right) of the macaque neocortex model with 40 brain areas with the model in [24]. (F) The spatial time constant map of 40 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 [25]. (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.