Predicting distributed working memory activity in a large-scale mouse brain: the importance of the cell type-specific connectome

Xingyu Ding[1,*], Sean Froudist-Walsh[1,2,*], Jorge Jaramillo[1,3,*], Junjie Jiang[1,4], and Xiao-Jing Wang[1,+]

[1]Center for Neural Science, New York University, New York, NY 10003, USA
[2]Computational Neuroscience Unit, Faculty of Engineering, University of Bristol, Bristol BS8 1UB, UK
[3]Campus Institute for Dynamics of Biological Networks, Goettingen, Germany
[4]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.
[*]co-first authors
[+]lead contact: xjwang@nyu.edu

December 8, 2022

Abstract

Recent advances in connectomic and neurophysiological tools make it possible to probe whole-brain mechanisms in the mouse that underlie cognition and behavior. Based on experimental data, we developed a large-scale model of the mouse brain for a cardinal cognitive function called working memory, the brain's ability to internally hold and process information without sensory input. In the model, interregional connectivity is constrained by mesoscopic connectome data. The density of parvalbumin-expressing interneurons in the model varies systematically across the cortex. We found that the long-range cell type-specific targeting and density of cell classes define working memory representations. A core cortical subnetwork and the thalamus produce distributed persistent activity, and the network exhibits numerous attractor states. Novel cell type-specific graph theory measures predicted the activity patterns and core subnetwork. This work highlights the need for cell type-specific connectomics, and provides a theory and tools to interpret large-scale recordings of brain activity during cognition.

Introduction

In contrast to our substantial knowledge of local neural computation, such as orientation selectivity in the primary visual cortex or spatial map of grid cells in the medial entorhinal cortex, much less is understood about distributed processes in multiple interacting brain regions underlying cognition and behavior. This has recently begun to change, as advances in new technologies enable neuroscientists to probe neural activity at single-cell resolution and on a large-scale by electrical recording or calcium imaging of behaving animals (Jun et al. 2017; Steinmetz et al. 2019; Stringer et al. 2019; Musall et al. 2019; Steinmetz et al. 2021), ushering in a new era of neuroscience investigating distributed neural dynamics and brain functions (Wang 2022).

To be specific, consider a core cognitive function called working memory, the ability to temporally maintain information in mind without external stimulation (Baddeley 2012). Working memory has long been studied in neurophysiology using delay-dependent tasks, where stimulus-specific information must be stored in working memory across a short time period between a sensory input and a memory-guided behavioral response (Fuster and Alexander 1971; Funahashi et al. 1989; Goldman-Rakic 1995; Wang 2001). Delay-period mnemonic persistent neural activity has been observed in multiple brain regions, suggesting distributed working memory representation (Suzuki and Gottlieb 2013; Leavitt et al. 2017; Christophel et al. 2017; Xu 2017; Dotson et al. 2018). Connectome-based computational models of the macaque cortex found that working memory activity depends on interareal connectivity (Murray et al. 2017; Jaramillo et al. 2019), macroscopic gradients of synaptic excitation (Wang 2020; Mejias and Wang 2022) and dopamine modulation (Froudist-Walsh et al. 2021a).

Mnemonic neural activity during a delay period is also distributed in the mouse brain (Liu et al. 2014; Schmitt et al. 2017; Guo et al. 2017; Bolkan et al. 2017; Gilad et al. 2018). The new recording and imaging techniques as well as optogenetic methods for causal analysis (Yizhar et al. 2011), that are widely applicable to behaving mice, hold promise for elucidating the circuit mechanism of distributed brain functions in rodents. Recurrent synaptic excitation represents a neural basis for the maintenance of persistent neural firing (Goldman-Rakic 1995; D. J. Amit 1995; Wang 2021). In the monkey cortex, the number of spines (sites of excitatory synapses) per pyramidal cell increases along the cortical hierarchy, consistent with the idea that mnemonic persistent activity in association cortical areas including the prefrontal cortex is sustained by recurrent excitation stronger than in early sensory areas. Such a macroscopic gradient is lacking in the mouse cortex (Gilman et al. 2017), raising the possibility that the brain mechanism for distributed working memory representations may be fundamentally different in mice and monkeys.

In this paper we report a cortical mechanism of distributed working memory that does not depend on a gradient of synaptic excitation. We developed an anatomically-based model of the mouse brain for working memory, built on the recently available mesoscopic connectivity data of the mouse thalamocortical system (Oh et al. 2014; Gămănuţ et al. 2018; Harris

et al. 2019; Kim et al. 2017). Our model is validated by capturing large-scale neural activity observed in recent mouse experiments (Guo et al. 2017; Gilad et al. 2018). Using this model, we found that a decreasing gradient of synaptic inhibition mediated by parvalbumin (PV) positive GABAergic cells (Kim et al. 2017; Fulcher et al. 2019; Wang 2020) shapes the distributed pattern of working memory representation.

A focus of this work is to examine whether anatomical connectivity can predict the emergent large-scale neural activity pattern underlying working memory. Interestingly, traditional graph-theory measures of inter-areal connections, which ignore cell types of projection targets, are uncorrelated with activity patterns. We propose new cell type-specific graph theory measures to overcome this problem, and differentiate contributions of cortical areas in terms of their distinct role in loading, maintaining, and reading out the content of working memory. Through computer-simulated perturbations akin to optogenetic inactivations, a core subnetwork was uncovered for the generation of persistent activity. This core subnetwork can be predicted based on the cell type-specific interareal connectivity, highlighting the necessity of knowing the cell type targets of interareal connections in order to relate anatomy with physiology and behavior. This work provides a computational and theoretical platform for cross-scale understanding of cognitive processes across the mouse cortex.

Results

A decreasing gradient of PV interneuron density from sensory to association cortex

Our large-scale circuit model of the mouse cortex uses inter-areal connectivity provided by anatomical data within the 43-area parcellation in the common coordinate framework v3 atlas (Oh et al. 2014) (Fig. 1A, Fig 1 - supplement 1A). The model is endowed with area-to-area variation of parvalbumin-expressing interneurons (PV) in the form of a gradient measured from the qBrain mapping platform (Fig 1 - supplement 1B) (Kim et al. 2017). The PV cell density (the number of PV cells per unit volume) is divided by the total neuron density, to give the PV cell fraction, which better reflects the expected amount of synaptic inhibition mediated by PV neurons (Fig. 1B-C, neuron density is shown in Fig 1 - supplement 1C). Cortical areas display a hierarchy defined by mesoscopic connectome data acquired using anterograde fluorescent tracers (Oh et al. 2014) (Fig. 1D-E). In Fig. 1F, the PV density is plotted as a function of the cortical hierarchy, which shows a moderate negative correlation between the two. Therefore, primary sensory areas have a higher density of PV interneurons than association areas, although the gradient of PV densities does not align perfectly with the cortical hierarchy.

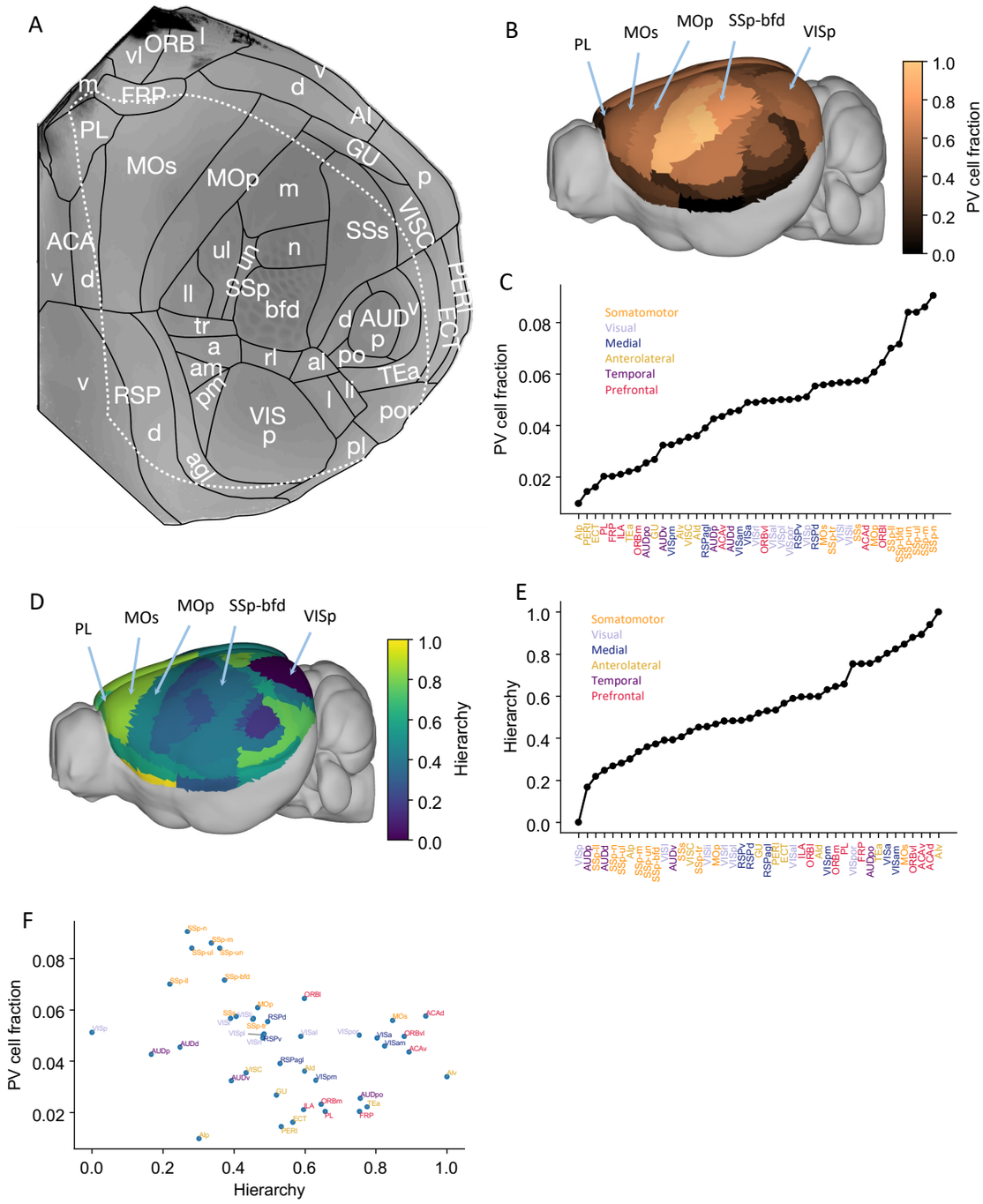

Figure 1: Anatomical basis of the multi-regional mouse cortical model. (A). Flattened view of mouse cortical areas. Figure adapted from (Harris et al. 2019). (B). Normalized PV cell fraction for each brain area, visualized on a 3d surface of the mouse brain. Five areas are highlighted : VISp, Primary somatosensory area, barrel field (SSp-bfd), primary motor (MOp), MOs and PL. (C). The PV cell fraction for each cortical area, ordered. Each area belongs to one of five modules, shown in color. (Harris et al. 2019). (D). Hierarchical position for each area on a 3d brain surface. Five areas are highlighted as in (B), and color represents the hierarchy position. (E). Hierarchical positions for each cortical area. The hierarchical position is normalized and the hierarchical position of VISp is set to be 0. Same as (C), colors represent which module an area belongs to. (F). Correlation between PV cell fraction and hierarchy (Pearson correlation coefficient $r = -0.35$, $p < 0.05$).

A whole-mouse cortex model with a gradient of interneurons

In our model, each cortical area is described by a local circuit (Fig. 2A), using a mean-field reduction (Wong and Wang 2006) of a spiking neural network (Wang 2002). A minimal version of it has two neural pools selective for different stimuli and a shared inhibitory neural pool. The model makes the following assumptions. First, local inhibitory strength is proportional to PV interneuron density across the cortex. Second, the inter-areal long-range connection matrix is given by the anterograde tracing data (Oh et al. 2014; Knox et al. 2018; Wang et al. 2020). Third, targeting is biased onto inhibitory cells for top-down compared with bottom-up projections, therefore feedforward connections have a greater net excitatory effect than feedback connections (counterstream inhibitory bias, CIB) (Mejias and Wang 2022; Javadzadeh and Hofer 2022; Wang 2022).

Distributed working memory activity depends on the gradient of inhibitory neurons and the cortical hierarchy

We simulated the large-scale network to perform a simple visual delayed response task that requires one of two stimuli to be held in working memory. We shall first consider the case in which the strength of local recurrent excitation is insufficient to generate persistent activity when parcellated areas are disconnected from each other. Consequently, the observed distributed mnemonic representation must depend on long-range interareal excitatory connection loops. Later in the paper we will discuss the network model behavior when some local areas are capable of sustained persistent firing in isolation.

The main question is: when distributed persistent activity emerges after a transient visual input (a 500 ms current pulse to a selective excitatory population) is presented to the primary visual cortex (VISp), what determines the spatial pattern of working memory representation? After we remove the external stimulus, the firing rate in area VISp decreases rapidly to baseline. Neural activity propagates throughout the cortex after stimulus offset (Fig. 2B). Neural activities in the higher visual cortical areas (e.g. VISrl and VISpl) show similar dynamics to VISp. In stark contrast, many frontal and lateral areas (including prelimbic (PL), infralimbic (ILA), secondary motor (MOs) and ventral agranular insula (AIv) areas) sustained a high firing rate during the delay period (Fig. 2B). Areas that are higher in the cortical hierarchy show elevated activity during the delay period (Fig. 2C). This persistent firing rate could last for more than 10 seconds and is a stable attractor state of the network (Inagaki et al. 2019).

The cortical hierarchy and PV fraction predict the delay period firing rate of each cortical area (Fig. 2C-E). Thus the activity pattern of distributed working memory depends on both local and large-scale anatomy. The delay activity pattern has a stronger correlation with hierarchy ($r = 0.91$) than with the PV fraction ($r = -0.43$). The long-range connections thus play a predominant important role in defining the persistent activity pattern.

Activity in early sensory areas such as VISp displays a rigorous response to the transient

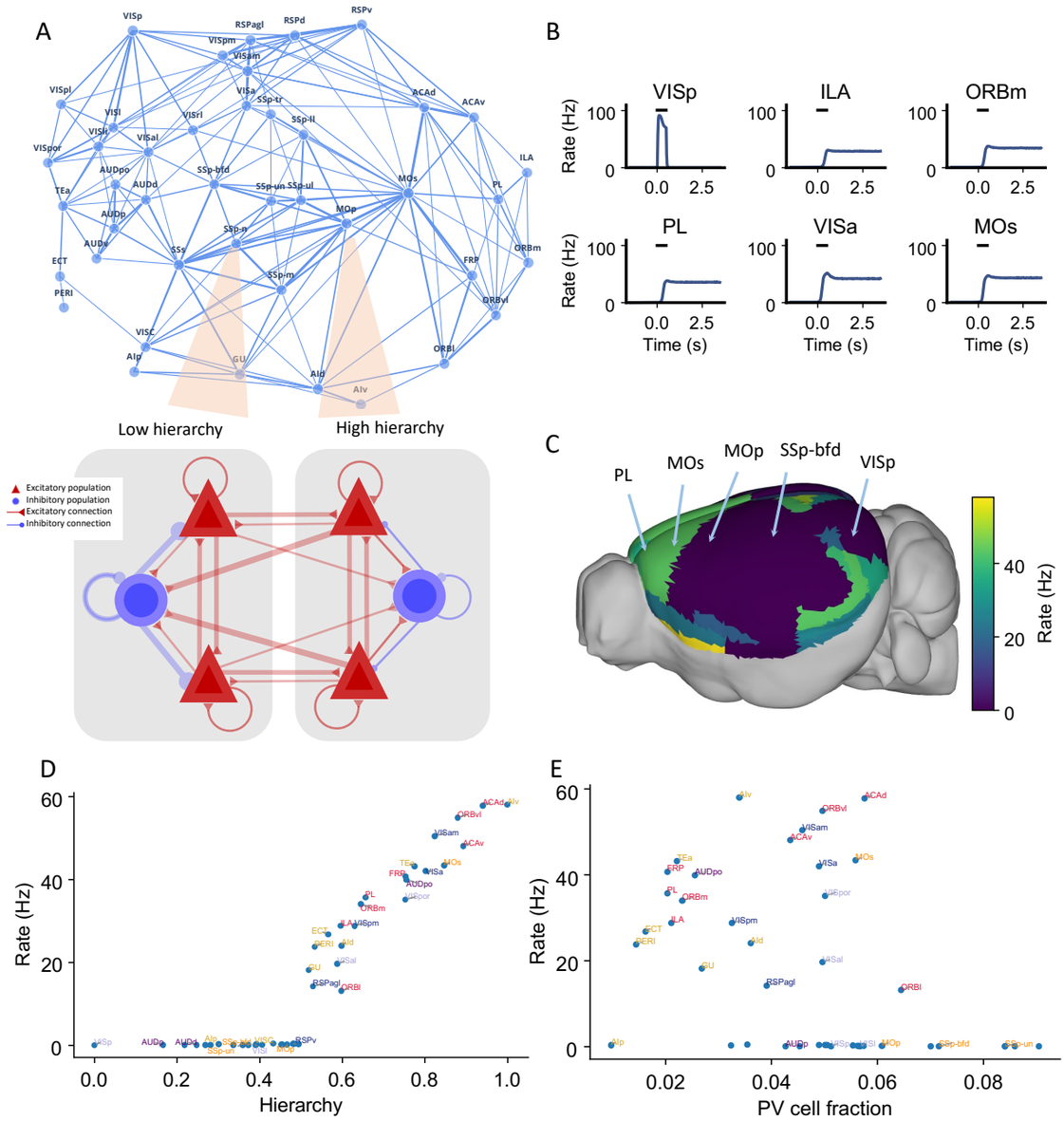

Figure 2: Distributed working memory activity depends on the gradient of PV density and the cortical hierarchy. (A). Model design of the large-scale model for distributed working memory. Top, connectivity map of the cortical network. Each node corresponds to a cortical area and an edge is a connection, where the thickness of the edge represents the strength of the connection. Only strong connections are shown (without directionality for the sake of clarity). Bottom, local and long-range circuit design. Each local circuit contains two excitatory populations (red), each selective to a particular stimulus and one inhibitory population (blue). Long-range connections are scaled by mesoscopic connectivity strength (Oh et al. 2014) and follows counterstream inhibitory bias (CIB) (Mejias and Wang 2022). (B). The activity of 6 selected areas during a working memory task is shown. A visual input of 500ms is applied to area VISp, which propagates to the rest of the large-scale network. (C). Delay period firing rate for each area on a 3d brain surface. Similar to Fig. 1B, the positions of 5 areas are labeled. (D). Delay-period firing rate is positively correlated with cortical hierarchy ($r = 0.91$, $p < 0.05$). (E). Delay-period firing rate is negatively correlated with PV cell fraction ($r = -0.43$, $p < 0.05$).

input but returns to a low firing state after stimulus withdrawal. In contrast, many frontal areas show strong persistent activity. When the delay period firing rates are plotted versus hierarchy, we observe a gap in the distribution of persistent activity (Fig. 2D) that marks an abrupt transition in the cortical space. This leads to the emergence of a subnetwork of areas capable of working memory representations.

We also used our circuit model to simulate delayed response tasks with different sensory modalities (Fig 2 - supplement 1), by stimulating primary somatosensory area SSp-bfd and primary auditory area AUDp. The pattern of delay period firing rates for these sensory modalities is similar to the results obtained for the visual task: sensory areas show transient activity, while frontal and lateral areas show persistent activity after stimulus withdrawal. Moreover, the cortical hierarchy could predict the delay period firing rate of each cortical area well ($r = 0.89$, $p < 0.05$). Our model thus predicts that working memory may share common activation patterns across sensory modalities, which is partially supported by cortical recordings during a memory-guided response task (Inagaki et al. 2018).

Thalamocortical interactions maintain distributed persistent activity

To investigate how thalamocortical interactions affect the large-scale network dynamics, we designed a thalamocortical network similar to the cortical network (Fig. 3A). Several studies have shown that thalamic areas are also involved in the maintenance of working memory (Bolkan et al. 2017; Guo et al. 2017; Schmitt et al. 2017). However, the large-scale thalamocortical mechanisms underlying memory maintenance are unknown. We set the strength of connections between the thalamus and cortex using data from the Allen Institute (Oh et al. 2014) (Fig 3 - supplement 1). All thalamocortical connections in the model are mediated by AMPA synapses. There are no recurrent connections in the thalamus within or across thalamic nuclei (Jones 2007). The effect of thalamic reticular nucleus neurons was included indirectly as a constant inhibitory current to all thalamic areas. Similarly to cortical areas, the thalamus is organized along a measured hierarchy (Harris et al. 2019). For example, the dorsal part of the lateral geniculate nucleus (LGd) is lower than the cortical area VISp in the hierarchy, consistent with the fact that LGd sends feedforward inputs to VISp. Thalamocortical projections in the model are slightly more biased toward excitatory neurons in the target area if they are feedforward projections and towards inhibitory neurons if they are feedback.

Here, we weakened the strength of cortical interareal connections as compared to the cortex model of Fig. 2. Now, persistent activity can still be generated (Fig. 3B, blue) but is maintained with the help of the thalamocortical loop, as observed experimentally (Guo et al. 2017). Indeed, in simulations where the thalamus was inactivated, the cortical network no longer showed sustained activity (Fig. 3B, red).

In the thalamocortical model, the delay activity pattern of the cortical areas correlates with the hierarchy, again with a gap in the firing rate separating the areas engaged in

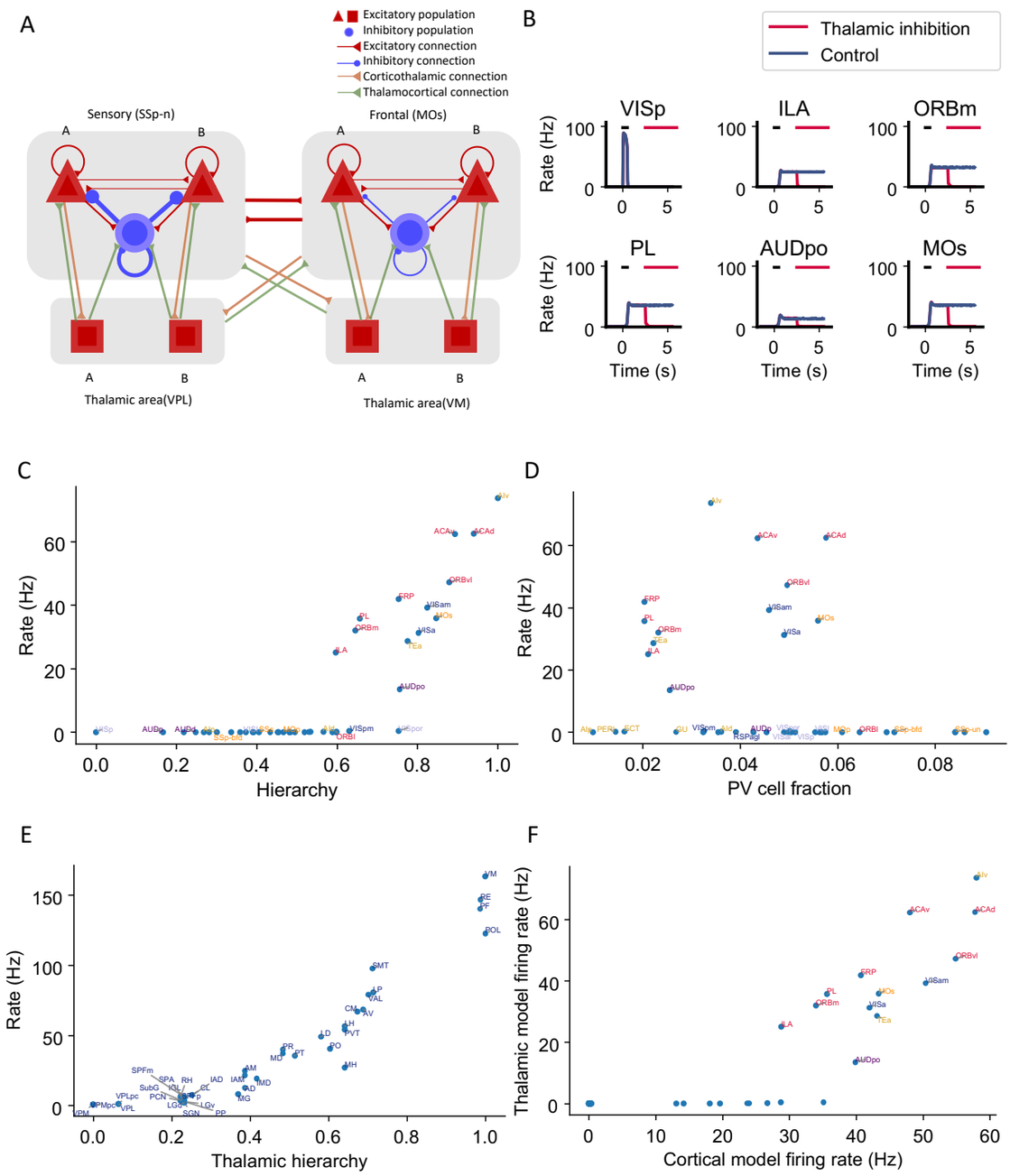

Figure 3: Thalamocortical interactions help maintain distributed persistent activity. (A). Model schematic of the thalamocortical network. The structure of the cortical component is the same as our default model in Fig. 2A, but with modified parameters. Each thalamic area includes two excitatory populations (red square) selective to different stimuli. Long range projections between thalamus and cortex also follow the counterstream inhibitory bias rule as in the cortex. Feedforward projections target excitatory neurons with stronger connections and inhibitory neurons with weaker connections; the opposite holds for feedback projections. (B). The activity of 6 sample cortical areas in a working memory task is shown during control (blue) and when thalamic areas are inhibited in the delay period (red). Black dashes represent the external stimulus applied to VISp. Red dashes represent external inhibitory input given to all thalamic areas. (C). Delay period firing rate of cortical areas in the thalamocortical network. The activity pattern has a positive correlation with cortical hierarchy ($r = 0.78$, $p < 0.05$). (D). Same as (C) but plotted against PV cell fraction. The activity pattern has a negative correlation with PV cell fraction, but it is not significant ($r = -0.26$, $p = 0.09$). (E). Delay firing rate of thalamic areas in thalamocortical network. The firing rate has a positive correlation with thalamic hierarchy ($r = 0.94$, $p < 0.05$). (F). Delay period firing rate of cortical areas in thalamocortical network has a positive correlation with delay firing rate of the same areas in a cortex-only model ($r = 0.77$, $p < 0.05$). Note that only the areas showing persistent activity in both models are considered for correlation analyses.

persistent activity from those that do not (Fig. 3B, Fig. 3C). Sensory areas show a low delay firing rate, and frontal areas show strong persistent firing. Unlike the cortex, the firing rate of thalamic areas continuously increases along the hierarchy (Fig. 3E). On the other hand, cortical dynamics in the thalamocortical and cortical models show many similarities. Early sensory areas do not show persistent activity in either model. Many frontal and lateral areas show persistent activity and there is an abrupt transition in cortical space in the thalamocortical model, like in the cortex only model. Quantitatively, the delay firing pattern of the cortical areas is correlated with the hierarchy and the PV fraction (Fig. 3C, Fig. 3D). Furthermore, the delay period firing rate of cortical areas in the thalamocortical model correlates well with the firing rate of the same areas in the cortical model (Fig. 3F). This comparison suggests that the cortical model captures most of the dynamical properties in the thalamocortical model; therefore in the following analyses, we will mainly focus on the cortex-only model for simplicity.

Cell type-specific connectivity measures predict distributed persistent firing patterns

Structural connectivity constrains large-scale dynamics (Mejias and Wang 2022; Froudist-Walsh et al. 2021a; Cabral et al. 2011). However, we found that standard graph theory measures could not predict the pattern of delay period firing across areas. There is no significant correlation between input strength and delay period firing rate ($r = 0.25$, $p = 0.25$, Fig. 4A(i), A(ii)) and input strength cannot predict which areas show persistent activity (prediction accuracy = 0.51, Fig. 4A(iii)). We hypothesized that this is because currently available connectomic data used in this model do not specify the type of neurons targeted by the long-range connections. For instance, when two areas are strongly connected with each other, such a loop would contribute to the maintenance of persistent activity if projections are mutually excitatory, but not if one of the two projections predominantly targets inhibitory PV cells. Therefore, cell type-specificity of interareal connections must be taken into account in order to relate the connectome with the whole-brain dynamics and function. To examine this possibility, we introduced a *cell type projection coefficient* (see *Calculation of network structure measures* in the Methods), which is smaller with a higher PV cell fraction in the target area (Fig 4 - supplement 1). The cell type projection coefficient also takes cell type targets of long range connections into account, which, in our model, is quantified by counterstream inhibitory bias (CIB). As a result, the modified cell type-specific connectivity measures increase if the target area has a low density of PV interneurons and/or if long-range connections predominantly target excitatory neurons in the target area.

We found that cell type-specific graph measures accurately predict delay-period firing rates. The cell type-specific input strength of the early sensory areas is weaker than the raw input strength (Fig. 4B(i)). The firing rate across areas is positively correlated with cell type-specific input strength (Fig. 4B(ii)). Cell type-specific input strength also accurately

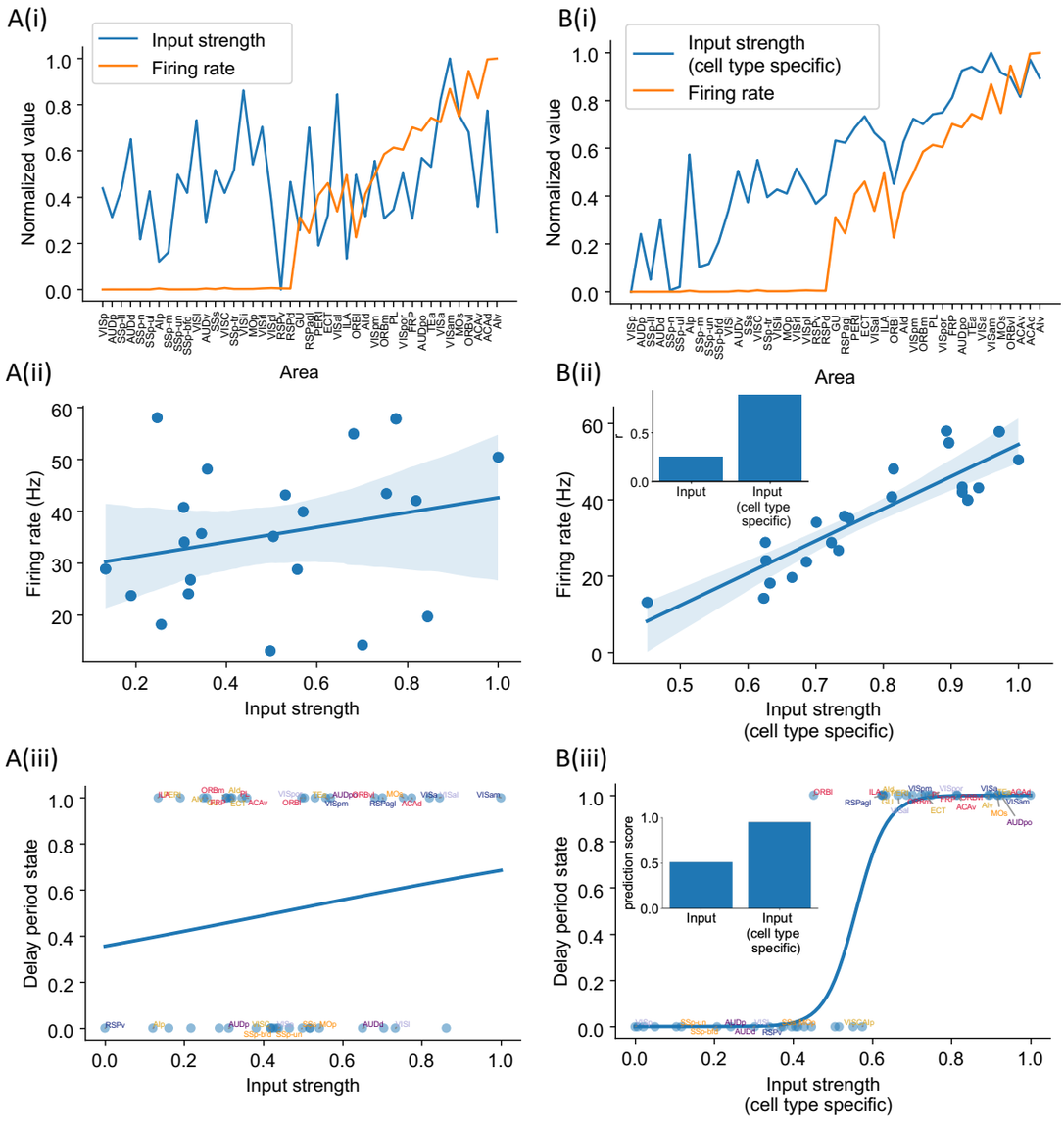

Figure 4: Cell type-specific connectivity measures are better at predicting firing rate pattern than nonspecific ones. (A(i)). Delay period firing rate (orange) and input strength for each cortical area. Input strength of each area is the sum of connectivity weights of incoming projections. Areas are plotted as a function of their hierarchical positions. Delay period firing rate and input strength are normalized for better comparison. (A(ii)). Input strength does not show significant correlation with delay period firing rate for areas showing persistent activity in the model ($r = 0.25$, $p = 0.25$). (A(iii)). Input strength cannot be used to predict whether an area shows persistent activity or not (prediction accuracy = 0.51). (B(i)). Delay period firing rate (orange) and cell type-specific input strength for each cortical area. Cell type-specific input strength considers how the long-rang connections target different cell types and is the sum of modulated connectivity weights of incoming projections. Same as (A(i)), areas are sorted according to their hierarchy and delay period firing rate and input strength are normalized for better comparison. (B(ii)). Cell type-specific input strength has a strong correlation with delay period firing rate of cortical areas showing persistent activity ($r = 0.89$, $p < 0.05$). Inset: Comparison of the correlation coefficient for raw input strength and cell type-specific input strength. (B(iii)). Cell type-specific input strength predicts whether an area shows persistent activity or not (prediction accuracy = 0.95). Inset: comparison of the prediction accuracy for raw input strength and cell type-specific input strength.

predicts which areas show persistent activity (Fig. 4B(iii)). Similarly, we found that the cell type-specific eigenvector centrality, but not standard eigenvector centrality (Newman 2018), was a good predictor of delay period firing rates (Fig. 4 - supplement 2).

A core subnetwork for persistent activity across the cortex

Many areas show persistent activity in our model. However, are all active areas equally important in maintaining persistent activity? When interpreting large-scale brain activity, we must distinguish different types of contribution to working memory. For instance, inactivation of an area like VISp impairs performance of a delay-dependent task because it is essential for a (visual) "input" to access working memory; on the other hand a "readout" area may display persistent activity only as a result of sustained inputs from other areas that form a "core", which are causally important for maintaining a memory representation.

We propose four types of areas related to distributed working memory: input, core, readout, and nonessential (Fig. 5A). External stimuli first reach input areas, which then propagate activity to the core and non-essential areas. Core areas form recurrent loops and support distributed persistent activity across the network. By definition, disrupting any of the core areas would affect persistent activity globally. The readout areas also show persistent activity. Yet, inhibiting readout areas has little effect on persistent activity elsewhere in the network. We can assign the areas to the four classes based on three properties: a) the effect of inhibiting the area during stimulus presentation on delay activity in the rest of the network; b) the effect of inhibiting the area during the delay period on delay activity in the rest of the network; c) the delay activity of the area itself on trials without inhibition.

In search of a core working memory subnetwork in the mouse cortex, in model simulations we inactivated each area either during stimulus presentation or during the delay period, akin to optogenetic inactivation in mice experiments. The effect of inactivation was quantified by calculating the decrement in the firing rate compared to control trials for the areas that were not inhibited (Fig. 5B). The VISp showed a strong inhibition effect during the stimulus period, as expected for an Input area. We identified eight areas with a substantial inhibition effect during the delay period (Fig. 5C), which we identify as a core for working memory. Core areas are distributed across the cortex. They include frontal areas PL, ILA, medial part of the orbital area (ORBm), which are known to contribute to working memory (Liu et al. 2014; Bolkan et al. 2017). Other associative and sensory areas (AId, VISpm, ectorhinal area (ECT), perihinal area (PERI), gustatory area (GU)) are also in the core. Similarly, we used the above criteria to classify areas as Readout or Non-essential (Fig. 5D).

The core subnetwork can be identified by the presence of strong excitatory loops

Inhibition protocols across many areas are computationally costly. We sought a structural indicator that is easy to compute and is predictive of whether an area is engaged in working

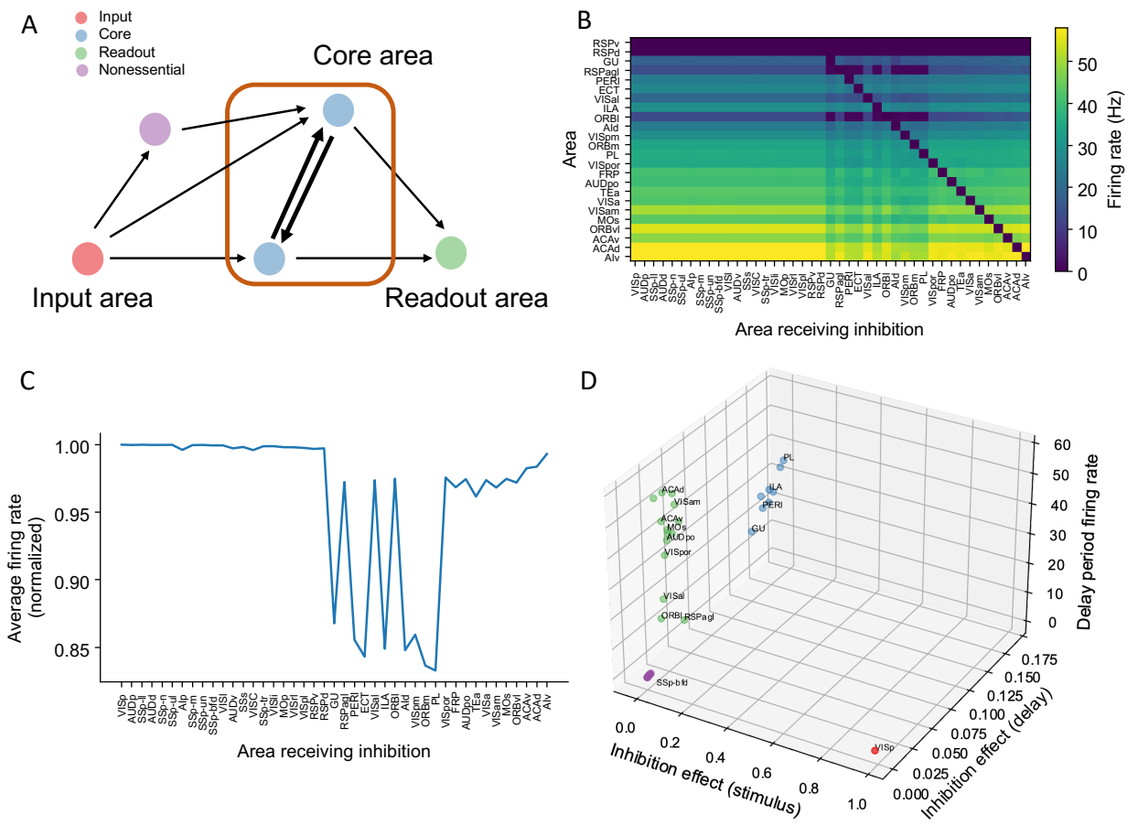

Figure 5: A core subnetwork generates persistent activity across the cortex. (A). We propose four different types of areas. Input areas (red) are responsible for coding and propagating external signals, which are then propagated through synaptic connections. Core areas (blue) form strong recurrent loops and generate persistent activity. Readout areas (green) inherit persistent activity from core areas. Nonessential areas (purple) may receive inputs and send outputs but they do not affect the generation of persistent activity. (B). Delay period firing rate for cortical areas engaged in working memory (Y axis) after inhibiting different cortical areas (X axis). Areas in the X axis and Y axis are both sorted according to hierarchy. Firing rates of areas with small firing rate (<1Hz) are partially shown (only RSPv and RSPd are shown because their hierarchical positions are close to areas showing persistent activity). (C). The average firing rate for areas engaged in persistent activity under each inhibition simulation. The X axis shows which area is inhibited, and the Y axis shows the average delay period activity for all areas showing persistent activity. Note that when calculating the average firing rate, the inactivated area was excluded in order to focus on the inhibition effect of one area on other areas. Average firing rates on the Y axis are normalized using the average firing in control (no inhibition) simulation. (D). Classification of 4 types of areas based on their delay period activity after stimulus- and delay-period inhibition. The inhibition effect, due to either stimulus or delay period inhibition, is the change of average firing rate normalized by the average firing rate in the control condition. Areas with strong inhibition effect during stimulus period are classified as Input areas; areas with strong inhibition effect during delay period and strong delay period firing rate are classified as Core areas; areas with weak inhibition effect during delay period but strong delay period firing rate during control are classified as Readout areas; areas with weak inhibition effect during delay period and weak delay period firing rate during control are classified as Nonessential areas.

memory function. Such an indicator could also guide the interpretation of large-scale neural recordings in experimental studies. In the dynamical regime where individual cortical areas do not show persistent activity independently, distributed working memory patterns must be a result of long-range recurrent loops across areas. We thus introduced a quantitative measurement of the degree to which each area is involved in long-range recurrent loops (Fig. 6A).

The core subnetwork can be identified by the presence of strong loops between excitatory cells. Here we focus on length-2 loops (Fig. 6A); the strength of a loop is the product of two connection weights for a reciprocally connected pair of areas; and the loop strength measure of an area is the sum of the loop strengths of all length-2 loops that the area is part of. Results were similar for longer loops (Fig. 6B, also see Fig. 6 - supplement 1 for results of longer loops). The raw loop strength had no positive statistical relationship to the core working memory subnetwork (Fig. 6C(i), Fig. 6C(ii)). We then defined cell type-specific loop strength (see Methods). The cell type-specific loop strength is the loop strength calculated using connectivity multiplied by the cell type projection coefficient. The cell type-specific loop strength, but not the raw loop strength, predicts which area is a core area with high accuracy (Fig. 6D(i), Fig. 6D(ii), prediction accuracy = 0.93). This demonstrates that traditional connectivity measures are informative but not sufficient to explain dynamics during cognition in the mouse brain. Cell type-specific connectivity, and new metrics that account for such connectivity, are necessary to infer the role of brain areas in supporting large-scale brain dynamics during cognition.

Multiple attractor states emerge from the mouse mesoscopic connectome and local recurrent interactions

Different tasks lead to dissociable patterns of internally sustained activity across the brain, described as separate attractor states. We developed a protocol to identify other attractor states, then analyzed the relationship between network properties and the attractor states (Fig. 7A-C). For different parameters, the number of attractors and the attractor patterns change. Two parameters are especially relevant here. These are the long-range connection strength (μ_{EE}) and local excitatory connection strength ($g_{E,self}$). These parameters affect the number of attractors in a model of the macaque cortex (Mejias and Wang 2022). Increasing the long range connection strength decreases the number of attractors (Fig. 7D). Stronger long-range connections implies that the coupling between areas is stronger. If areas are coupled with each other, the activity state of an area will be highly correlated to that of its neighbors. This leads to less variability and fewer attractors.

To quantify how the patterns of attractors change for different parameters, two quantities are introduced. The *attractor fraction* is the fraction of all detected attractor states to which an area belongs. An area "belongs" to an attractor state if it is in a high activity state in that attractor. The *attractor size* is defined by the number of areas belonging to that attractor.

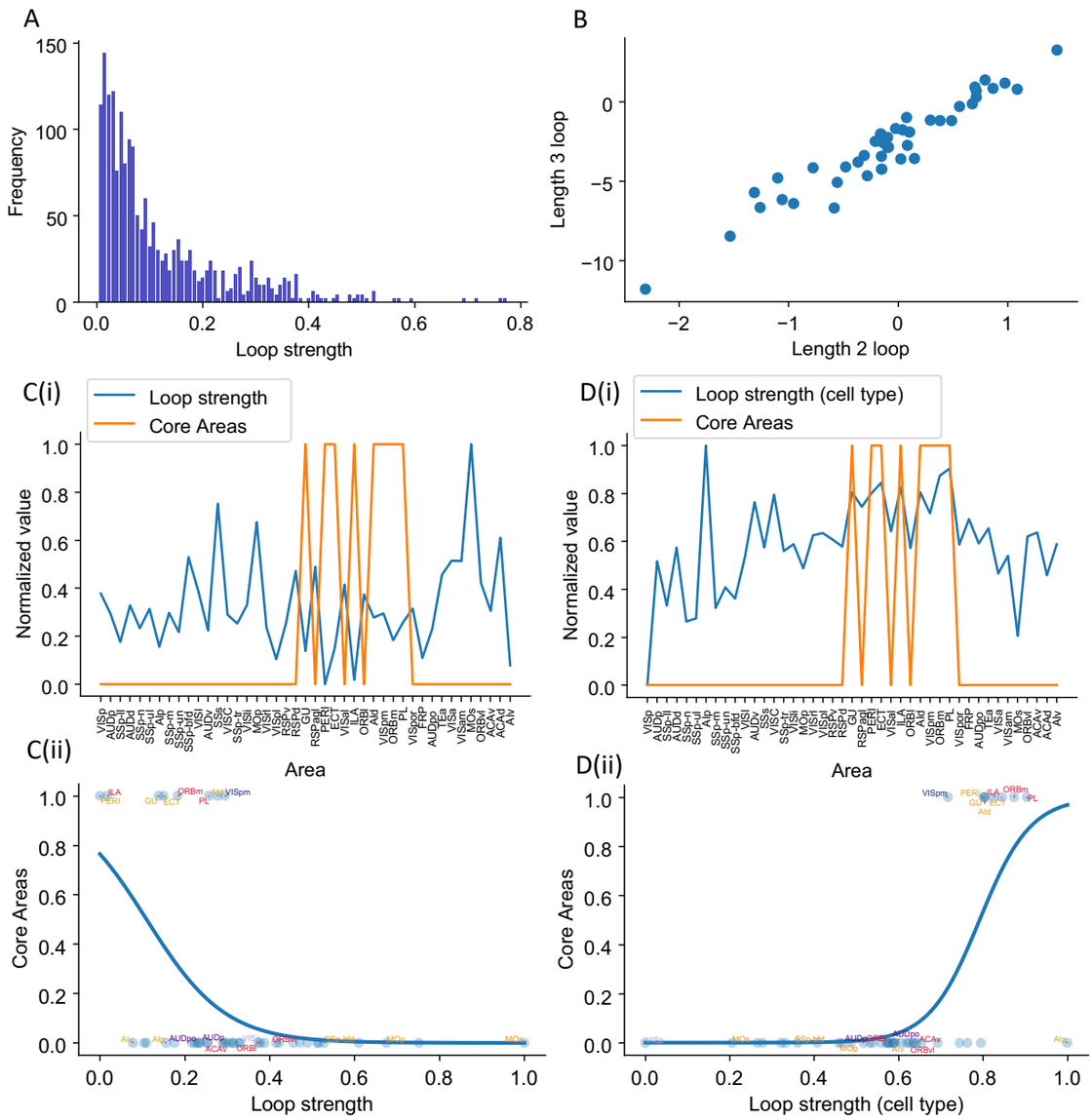

Figure 6: The core subnetwork can be identified structurally by the presence of strong excitatory loops. (A). Distribution of length-2 loops. X axis is the single loop strength of each loop (product of connectivity strengths within loop) and Y axis is their relative frequency. (B). Loop strengths of each area calculated using different length of loops (e.g., length 3 vs length 2) are highly correlated (r = 0.96, p < 0.05). (C(i)). Loop strength (blue) is plotted alongside Core Areas (orange), a binary variable that takes the value 1 if the area is a Core Area, 0 otherwise. Areas are sorted according to their hierarchy. The loop strength is normalized to a range of (0, 1) for better comparison. (C(ii)). A high loop strength value does not imply that an area is a core area. Blue curve shows the logistic regression curve fits to differentiate the core areas versus non core areas. (D(i)). Same as (C), but for cell type-specific loop strength. (D(ii)). A high cell type-specific loop strength predicts that an area is a core area (prediction accuracy = 0.93). Same as (C), but for cell type-specific loop strength.

As we increased the long-range connection strength, the attractor size distribution became bimodal. The first mode corresponded to large attractors, with many areas. The second mode corresponded to small attractors, with few areas (Fig. 7D).

When the local excitatory strength is increased, the number of attractors increased as well (Fig. 7E). In this regime some areas are endowed with sufficient local reverberation to sustain persistent activity even when decoupled from the rest of the system, therefore the importance of long-range coupling is diminished and a greater variety of attractor states is enabled. This can be understood by a simple example of two areas 1 and 2, each capable of two stimulus-selective persistent activity states; even without coupling there are $2 \times 2 = 4$ attractor states with elevated firing. Thus, local and long-range connection strength have opposite effects on the number of attractors.

The cell type-specific input strength predicted firing rates across many attractors. In an example parameter regime ($\mu_{EE} = 0.04$ nA and $g_{E,self} = 0.44$ nA), we identified 143 attractors. We correlated the input strength and cell type-specific input strength with the many attractor firing rates (Fig. 7F). The raw input strength is weakly correlated with activity patterns. The cell type-specific input strength is strongly correlated with activity across attractors. This shows that the cell type-specific connectivity measures are better at predicting the firing rates in many scenarios. These results further prove the importance of having cell type-specific connectivity for modeling brain dynamics.

Different attractor states rely on distinct subsets of core areas. In one example attractor, we found 5 areas that show persistent activity: VISa, VISam, FRP, MOs and ACAd (Fig. 7G) (parameter regime, $\mu_{EE} = 0.03$ nA and $g_{E,self} = 0.44$ nA). We repeated the previous inhibition analysis to identify core areas for this attractor state. Inhibiting one area, MOs, during the delay had the strongest effect on delay activity in the other parts of the attractor (Fig. 7H). MOs also showed strong persistent activity during delay period. This is consistent with its role in short-term memory and planning (Li et al. 2015; Inagaki et al. 2019). According to our definition, MOs is a core area for this attractor. To calculate a loop strength that was specific to this attractor, we only examined connections between these five areas. The cell type-specific loop strength was strongest in area MOs (Fig. 7I). Thus, we can identify likely core areas for individual attractor states from cell type-specific structural measures. This also demonstrates that different attractor states can be supported by distinct core areas.

Discussion

We developed a connectome-based dynamical model of the mouse brain. The model was capable of internally maintaining sensory information across many brain areas in distributed activity in the absence of any input. To our knowledge this is the first biologically-based model of the entire mouse cortex and the thalamocortical system for a cognitive function. Together with our recent work (Mejias and Wang 2022; Froudist-Walsh et al. 2021a; Froudist-Walsh et al. 2021b), it provides a study of contrast between mice and monkeys.

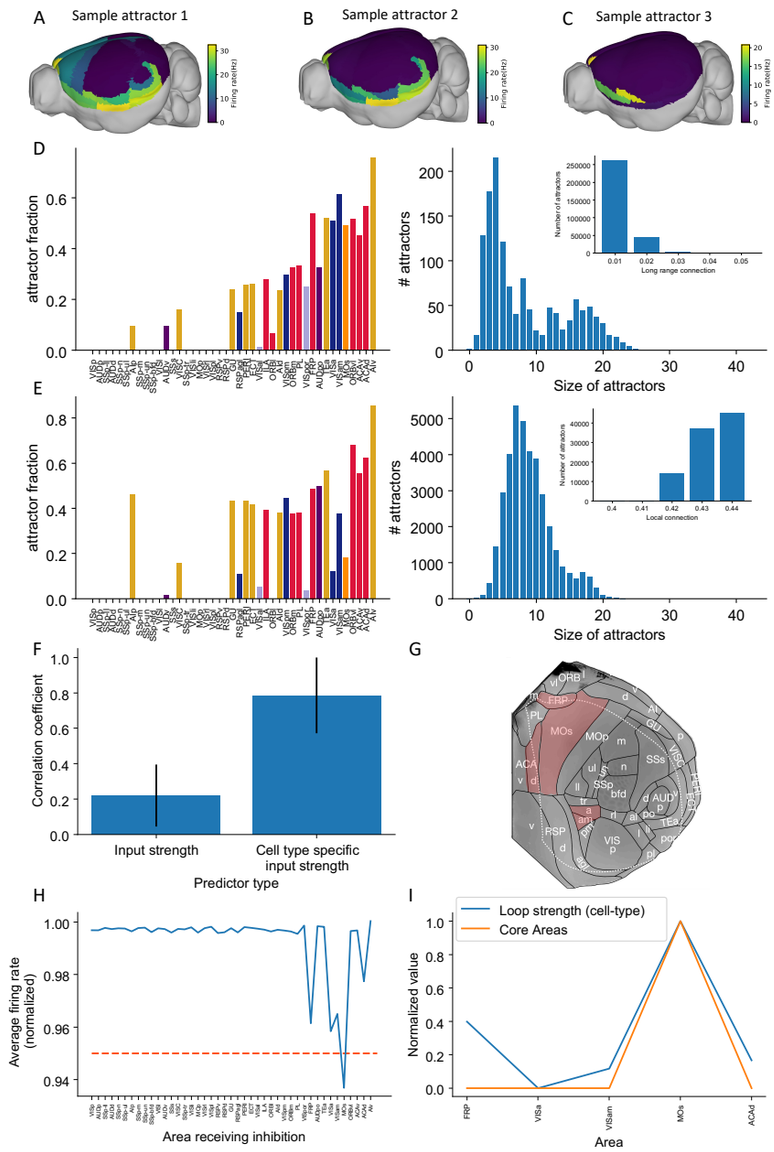

Figure 7: Multiple attractors coexist in the mouse working memory network. (A-C) Example attractor patterns with a fixed parameter set. Each attractor pattern can be reached via different external input patterns applied to the brain network. Delay activity is shown on a 3D brain surface. Color represents the firing rate of each area. (D-E) The distribution of attractor fractions (left) and number of attractors as a function of size (right) for different parameter combinations are shown. Attractor fraction of an area is the ratio between the number of attractors that include the area and the total number of identified attractors. In (D), local excitatory strengths are fixed ($g_{E,self}$ = 0.44 nA) while long-range connection strengths vary in the range μ_{EE} = 0.01-0.05 nA. Left and right panels of (D) show one specific parameter μ_{EE} = 0.03 nA. Inset panel of (D) shows the number of attractors under different long-range connection strengths while $g_{E,self}$ is fixed at 0.44 nA. In (E), long range connection strengths are fixed (μ_{EE} = 0.02 nA) while local excitatory strengths varies in the range $g_{E,self}$ = 0.4-0.44 nA. Left and right panels of (E) show one specific parameter $g_{E,self}$ = 0.43 nA. Inset panel of (E) shows the number of attractors under different local excitatory strengths, while μ_{EE} is fixed at 0.02 nA. (F). Prediction of the delay period firing rate using input strength and cell type-specific input strength for each attractor state identified under μ_{EE} = 0.04 nA and $g_{E,self}$ = 0.44 nA. 143 distinct attractors were identified and the average correlation coefficient using cell type-specific input strength is better than that using input strength. (G). A example attractor state identified under the parameter regime μ_{EE} = 0.03 nA and $g_{E,self}$ = 0.44 nA. The 5 areas with persistent activity are shown in red. (H). Effect of single area inhibition analysis for the attractor state in (G). For a regime where 5 areas exhibit persistent activity during the delay period, inactivation of the premotor area MOs yields a strong inhibition effect (<0.95 orange dashed line) and is therefore a Core area for the attractor state in (G). (I). Cell type-specific loop strength (blue) is plotted alongside Core areas (orange) for the attractor state in (G). Only 5 areas with persistent activity are used to calculate the loop strength. Loop strength is normalized to be within the range of 0 and 1. High cell type-specific loop measures predict that an area is a Core area (prediction accuracy is 100% correct). The number of areas is limited, so prediction accuracy is very high.

Our main findings are threefold. First, mnemonic activity pattern is shaped by the differing densities of PV interneurons across cortical areas. Areas with high PV density encoded information only transiently. Those with low PV density sustained activity for longer periods. Thus, the gradient of PV cells (Kim et al. 2017) has a definitive role in separating rapid information processing in sensory areas from sustained mnemonic information representation in associative areas of the mouse cortex. This is consistent with the view that each local area operates in the “inhibition-stabilizing regime" where recurrent excitation alone would lead to instability but the local network is stabilized by feedback inhibition even in the primary visual cortex (R. J. Douglas et al. 1995; Murphy and Miller 2009). Second, we deliberately considered two different dynamical regimes: when local recurrent excitation is not sufficient to sustain persistent activity and when it does. In the former case, distributed working memory must emerge from long-range interactions between parcellated areas, thereby the concept of synaptic reverberation (Lorente de Nó 1933; P. S. Goldman-Rakic 1995; Wang 2001; Wang 2021) is extended to the large-scale global brain. Note that currently it is unclear whether persistent neural firing observed in a delay dependent task is generated locally or depends on long-distance reverberation among multiple brain regions. Our work made the distinction explicit and offers specific predictions to be tested experimentally. Third, presently available connectomic data are not sufficient to account for neural dynamics and distributed cognition, and we propose cell type-specific connectomic measures that are shown to predict the observed distributed working memory representations.

We found that the cortical structures form recurrent loops with the thalamus, and the thalamocortical loops aided in sustaining activity throughout the delay period. The specific pattern of cortico-cortical connections was also critical to working memory. However, standard graph theory measures based on the connectome were unable to predict the pattern of working memory activity. By focusing on cell type-specific interactions between areas, we were able to reveal a core of cortical areas. The core is connected by excitatory loops, and is responsible for generating a widely distributed pattern of sustained activity. Outside the core, we identified "readout" areas that inherited activity from the core. Readout areas could use this information for further computations. This clarifies the synergistic roles of the connectome and gradients of local circuit properties in producing a distributed cognitive function. This additionally highlights the need for a cell type-specific connectome.

Previous large-scale models of the human and macaque cortex have replicated the functional connectivity (Deco et al. 2014; Demirtaş et al. 2019; Honey et al. 2007; Schmidt et al. 2018; Shine et al. 2018; Cabral et al. 2011; Wang et al. 2019) and propagation of information along the cortical hierarchy (Chaudhuri et al. 2015; Joglekar et al. 2018; Diesmann et al. 1999). Recently, large-scale models of brain activity during cognitive tasks have been developed (Mejias and Wang 2022; Froudist-Walsh et al. 2021a; Klatzmann et al. 2022). In these models, the number of dendritic spines per pyramidal cell increases along the hierarchy. This enables associative regions of the cortex to maintain information in working memory. The increase of dendritic spines along the hierarchy is a robust feature of primate cortical organisation

(Elston 2007), which does not exist in the mouse cortex (Gilman et al. 2017). Yet, in the mouse cortex, other properties do vary along the cortical hierarchy (Kim et al. 2017; Fulcher et al. 2019). We took advantage of the recent discovery of a gradient of PV interneurons in the mouse cortex (Kim et al. 2017), and implemented it directly in our large-scale model. We demonstrated how the increasing gradient of excitation along the primate cortical hierarchy and the decreasing gradient of PV inhibition in the mouse cortex could serve a similar role. Both gradients enable sustained activity to emerge in associative areas. Thus, while the neural activity underlying working memory may be widely distributed in both rodents and primates, the circuit-level mechanisms may differ. This should be considered when interpreting studies of working memory in rodent models of cognition and disease.

In the macaque, long-range connectivity is a strong predictor of the working memory activity (Mejias and Wang 2022; Froudist-Walsh et al. 2021a). Thus, at least some of the functional specialization of brain areas is due to differences in interareal connections. In contrast, we found that traditional graph theory metrics of connectivity were unable to predict the working memory activity in the mouse brain. This may be due to the almost fully connected pattern of interareal connectivity in the mouse cortex (Gămănuţ et al. 2018). This implies that, qualitatively, all areas have a similar set of cortical connections. In our model, we allowed the cell type target of interareal connections to change according to the relative position of the areas along the cortical hierarchy. Feedforward connections had a greater net excitatory effect than feedback connections. This hypothesis (Mejias and Wang 2022) has received some recent experimental support (Yoo et al. 2021; Huang et al. 2019; Javadzadeh and Hofer 2022).

By introducing cell type-specific graph theory metrics, we were able to predict the pattern and strength of delay period activity with high accuracy. Connectome databases are an invaluable resource for basic neuroscience. However, they may be insufficient for constraining models of brain activity. In the future, connectome databases should be supplemented by cell type-specific information.

We demonstrated how cell type-specific graph-theory measures can accurately identify the core subnetwork, which can also be identified independently using a simulated large-scale optogenetic experiment. We found a core subnetwork of areas that, when inhibited, caused a substantial drop in activity in the remaining cortical areas. This core working memory subnetwork included frontal cortical areas with well documented patterns of sustained activity during working memory tasks, such as prelimbic (PL), infralimbic (ILA) and medial orbitofrontal cortex (ORBm) (Schmitt et al. 2017; Liu et al. 2014; Wu et al. 2020). However, the core subnetwork for the visual working memory task we assessed was distributed across the cortex. It also included temporal and higher visual areas, suggesting that long-range recurrent connections between the frontal cortex and temporal and visual areas are responsible for generating persistent activity and maintaining visual information in working memory in the mouse.

The core visual working memory subnetwork generates activity that is then inherited

by many readout areas. Readout areas also exhibit persistent activity. However, inhibiting readout areas does not significantly affect the activity of other areas (Figure 5). Readout areas can use the stored information for further computations or to affect behavior. The readout areas in our model were a mixture of higher visual areas, associative areas and premotor areas of cortex. Notably, we classified the secondary motor cortex (MOs), which contains the anterior lateral motor (ALM) area, as a readout area despite its high level of persistent activity. ALM has received a lot of attention in mouse studies of working memory and motor preparation (Guo et al. 2017; Guo et al. 2014; Inagaki et al. 2019; Li et al. 2015; Wu et al. 2020; Voitov and Mrsic-Flogel 2022). If ALM is indeed a readout area for sensory working memory tasks, (e.g., Schmitt et al. 2017), then the following prediction arises. Inhibiting ALM should have a relatively small effect on sustained activity in core areas (such as PL) during the delay period. In contrast, inhibiting PL and other core areas may disrupt sustained activity in ALM. Even if ALM is not part of the core for sensory working memory, it could form part of the core for motor preparation tasks (Figure 7G). We found a high cell type-specific loop strength for area ALM, like that in core areas, which supports this possibility (Figure 7I). Furthermore, we found some attractor states for which the MOs was classified as a core area, while PL was not even active during the delay period. Our result is supported by a recent study that found no behavioral effect after PL inhibition in a motor planning task (Wang et al. 2021). Therefore, the core subnetwork required for generating persistent activity is likely task-dependent. Outside of this core subnetwork, there is a large array of readout brain areas that can use the stored information to serve behavior. Future modeling work may help elucidate the biological mechanisms responsible for switching between attractor landscapes for different tasks.

Neuroscientists are now observing task-related neural activity at single-cell resolution across much of the brain (Stringer et al. 2019; Steinmetz et al. 2019). This makes it important to identify ways to distinguish the core areas for a function from those that display activity that serves other purposes. We show that a large-scale inhibition protocol can identify the core subnetwork for a particular task. We further show how this core can be predicted based on the interareal loops that target excitatory neurons. Were such a cell type-specific interareal connectivity dataset available, it may help interpretation of large-scale recording experiments. This could also focus circuit manipulation on regions most likely to cause an effect on the larger network activity and behavior. Our approach identifies the brain areas that are working together to support working memory. It also identifies those that are benefiting from such activity to serve other purposes. Our simulation and theoretical approach is therefore ideally suited to understand the large-scale anatomy, recording and manipulation experiments which are at the forefront of modern systems neuroscience.

Neuroscience has rapidly moved into a new era of investigating large-scale brain circuits. Technological advances have enabled the measurement of connections, cell types and neural activity across the mouse brain. We developed a model of the mouse brain and theory of working memory that is suitable for the large-scale era. Previous reports have emphasized the

importance of gradients of dendritic spine expression and interareal connections in sculpting task activity in the primate brain (Mejias and Wang 2022; Froudist-Walsh et al. 2021a). Although these anatomical properties from the primate cortex are missing in the mouse brain (Gămănuţ et al. 2018; Gilman et al. 2017), other properties such as interneuron density (Kim et al. 2017) may contribute to areal specialization. Indeed, our model clarifies how gradients of interneurons and cell type-specific interactions define large-scale activity patterns in the mouse brain during working memory, which enables sensory and associative areas to have complementary contributions. Future large-scale modeling studies can leverage cell type-specific connectivity to study other important cognitive computations beyond working memory, including learning and decision making (Abbott et al. 2017; Abbott et al. 2020).

Acknowledgements

We thank Daniel P. Bliss and Ulises Pereira for support with analysis tools at the beginning of the project, and members of the Wang Lab at New York University for discussions related to the project. This work was funded by US National Institutes of Health (NIH) grant R01MH062349, Office of Naval Research (ONR) grant N00014, National Science Foundation (NSF) NeuroNex grant 2015276, Simons Foundation grant 543057SPI, and NIH U19NS123714.

Author contributions

XJW: designed the research, worked with the other authors throughout the project and co-wrote the paper; XYD: carried out all the computer simulations and analysis of simulation data, and co-wrote the paper; SFW, JJ and JJJ: contributed to all aspects of this project in interactions with XYD and co-wrote the paper.

Declaration of Interests

No competing interests declared.

Methods

Anterograde tracing, connectivity data

We used the mouse connectivity map from Allen institute (Oh et al. 2014) to constrain our large-scale circuit model of the mouse brain. The Allen Institute measured the connectivity among cortical and subcortical areas using an anterograde tracing method. In short, they injected virus and expressed fluorescent protein in source areas and performed fluorescent imaging in target areas to measure the strength of projections from source areas. Unlike retrograde tracing methods used in other studies (Markov et al. 2014b), the connectivity

strength measured using this method does not need to be normalized by the total input or output strength. This means that connectivity strength between any two areas is comparable.

The entries of the connectivity matrix from the Allen Institute can be interpreted as proportional to the total number of axonal fibers projecting from unit volume in one area to unit volume in another area. Before incorporating the connectivity into our model, we normalized the data as follows. In each area, we model the dynamics of an "average" neuron, assuming that the neuron receive inputs from all connected areas. Thus, we multiplied the connectivity matrix by the volume Vol_j of source area j and divided by the average neuron density d_i in target area i:

$$W_{norm,ij} = W_{raw,ij}\frac{Vol_j}{d_i}, \tag{1}$$

where $W_{raw,ij}$ is the raw, i.e., original, connection strength from unit volume in source area j to unit volume in target area i, Vol_j is the volume of source area j (Wang et al. 2020), and d_i is the neuron density in source area i (Erö et al. 2018). $W_{norm,ij}$ is the matrix that we use to set the long rang connectivity in our circuit model. We can define the cortico-thalamic connectivity $W_{ct,norm,ij}$ and thalamo-cortical connectivity $W_{tc,norm,ij}$ in a similar manner.

Interneuron density along the cortex

Kim and colleagues measured the density of typical interneuron types in the brain (Kim et al. 2017). They expressed fluorescent proteins in genetically labeled interneurons and counted the number of interneurons using fluorescent imaging. We took advantage of these interneuron density data and specifically used the PV density to set local and long-range inhibitory weights. We first normalized the PV density in each area:

$$PV_i = \frac{PV_{raw,i} - min(PV_{raw,i})}{max(PV_{raw,i}) - min(PV_{raw,i})} \tag{2}$$

$PV_{raw,i}$ is the PV interneuron density of all layers in area i, and PV_i is a normalized value of PV_{raw}, which will be used in subsequent modeling.

Hierarchy in the cortex

The concept of hierarchy is important for understanding the cortex. Hierarchy can be defined based on mapping corticocortical long range connections onto feedforward or feedback connections (Felleman and Essen 1991; Markov et al. 2014a; Harris et al. 2019). Harris and colleagues measured the corticocortical projections and target areas in a series of systematic experiments in mice (Harris et al. 2019). Projection patterns were clustered into multiple groups and the label "feedforward" or "feedback" was assigned to each group. Feedforward and feedback projections were then used to determine relative hierarchy between areas. For example, if the projections from area A to area B are mostly feedforward, then area B has a higher hierarchy than area A. This optimization process leads to a quantification of the

relative hierarchy of cortical areas $h_{raw,i}$. We defined the normalized hierarchy value h_i as

$$h_i = \frac{h_{raw,i} - min(h_{raw,i})}{max(h_{raw,i}) - min(h_{raw,i})} \tag{3}$$

where $h_{raw,i}$ is the raw, i.e., original hierarchical ordering from (Harris et al. 2019). Due to data acquisition issues, 6 areas did not have a hierarchy value assigned to them (SSp-un, AUDv, GU, VISC, ECT, PERI) (Harris et al. 2019). We estimated hierarchy through a weighted sum of the hierarchy value of 37 known areas, while the weight is determined through the connectivity strength. The parameters α_h and β_h are selected so that $h_{i,estimate}$ are close to h_i for areas with known hierarchy.

$$h_{i,estimate} = \alpha_h \frac{\sum_{j=1}^{37} w_{raw,ij} h_j}{\sum_{j=1}^{37} w_{raw,ij}} + \beta_h. \tag{4}$$

For the thalamocortical model, we also used the hierarchy value for thalamic areas (Harris et al. 2019). The hierarchy of thalamic areas are comparable to cortical areas, so in order to use it in the model, we also normalized them.

$$h_{th,i} = \frac{h_{th,raw,i} - min(h_{raw,i})}{max(h_{raw,i}) - min(h_{raw,i})} \tag{5}$$

To estimate the hierarchy value of thalamic areas with missing values, we used the known hierarchy value of the thalamic area next to the missing one as a replacement.

Description of the local circuit

Our large-scale circuit model includes 43 cortical areas. Each area includes two excitatory populations, labeled A and B, and one inhibitory population, C. The two excitatory populations are selective to different stimuli. The synaptic dynamics between populations are based on previous firing rate models of working memory (Wang 1999; Wong and Wang 2006). The equations that define the dynamics of the synaptic variables are

$$\frac{dS_A}{dt} = -\frac{S_A}{\tau_N} + \gamma(1 - S_A) r_A \tag{6}$$

$$\frac{dS_B}{dt} = -\frac{S_B}{\tau_N} + \gamma(1 - S_B) r_B \tag{7}$$

$$\frac{dS_C}{dt} = -\frac{S_C}{\tau_G} + \gamma_I r_C \tag{8}$$

where S_A and S_B are the NMDA synaptic variables of excitatory populations A and B, while S_C is the GABA synaptic variable of the inhibitory population C. r_A, r_B and r_C are the firing rates of populations A, B and C, respectively. τ_N and τ_G are the time constants of NMDA and GABA synaptic conductances. γ and γ_I are the parameters used to scale the contribution of presynaptic firing rates. The total currents received I_i $(i = A, B, C)$ are given by

$$I_A = g_{E,self}S_A + g_{E,cross}S_B + g_{EI}S_C + I_{0A} + I_{LR,A} + x_A(t) \tag{9}$$

$$I_B = g_{E,self}S_B + g_{E,cross}S_A + g_{EI}S_C + I_{0B} + I_{LR,B} + x_B(t) \tag{10}$$

$$I_C = g_{IE}S_A + g_{IE}S_B + g_{II}S_C + I_{0C} + I_{LR,C} + x_C(t). \tag{11}$$

In these equations, $g_{E,self}$, $g_{E,cross}$ denote the connection strength between excitatory neurons with same or different selectivity, respectively. These connection strengths are the same for different areas, since there is no significant gradient for excitatory strength in mice. g_{IE} are the connection strengths from excitatory to inhibitory neurons, while g_{EI}, and g_{II} are connection strengths from inhibitory to excitatory neurons and from inhibitory to inhibitory neurons, respectively. These connections will be scaled by PV density in the corresponding area. We will discuss the details in the next section. I_{0i} $(i = A, B, C)$ are constant background currents to each population. $I_{LR,i}$ $(i = A, B, C)$ are the long range (LR) currents received by each population. The term $x_i(t)$ where $i = A, B, C$ represents noisy contributions from neurons external to the network. It is modeled as an Ornstein-Uhlenbeck process:

$$\tau_{noise}\frac{dx_i}{dt} = -x_i + \sqrt{\tau_{noise}}\sigma_i\zeta_i(t), \tag{12}$$

where $\zeta_i(t)$ is Gaussian white noise, τ_{noise} describes the time constant of external AMPA synapses and σ_i sets the strength of the noise for each population. $\sigma_A = \sigma_B = 5pA$ while $\sigma_C = 0pA$.

The steady state firing rate of each population is calculated based on a transfer function $\phi_i(I)$ of input current received by each population I_i $(i = A, B, C)$ given by

$$\phi_{A,B}(I_{A,B}) = \frac{aI_{A,B} - b}{1 - exp[-d(aI_{A,B} - b)]} \tag{13}$$

$$\phi_C(I_C) = [\frac{1}{g_I}(c_1 I - c_0) + r_0]^+ \tag{14}$$

Note that the transfer functions $\phi_i(t)$ are the same for two excitatory populations. x^+ denotes the positive part of the function x. The firing rate of each population follows equations:

$$\tau_r\frac{dr_{A,B}}{dt} = -r_{A,B} + \phi_{A,B}(I_{A,B}) \tag{15}$$

$$\tau_r\frac{dr_C}{dt} = -r_C + \phi_C(I_C). \tag{16}$$

Interneuron gradient and local connections

We scaled local interneuron connectivity with the interneuron density that was obtained using fluorescent labeling (Kim et al. 2017). Specifically, local I-I connections and local I-E connections are scaled by the interneuron density by setting the connection strength $g_{k,i}(k = EI, II)$ as a linear function of PV density PV_i in area i.

$$g_{EI,i} = J_{EI,min} + J_{EI,scaling}PV_i \tag{17}$$

$$g_{II,i} = J_{II,min} + J_{II,scaling}PV_i \tag{18}$$

where $J_{k,min}$ $(k = EI, II)$ is the intercept and $J_{k,scaling}$ $(k = EI, II)$ is the slope.

Hierarchy and long range connections

Long range (LR) connections between areas are scaled by connectivity data from the Allen Institute (Oh et al. 2014). We consider long-range connections that arise from excitatory neurons because most long-range connections in the cortex correspond to excitatory connections (Petreanu et al. 2009). Long-range connections will target excitatory populations in other brain areas with the same selectivity (Zandvakili and Kohn 2015) and will also target inhibitory neurons. These long-range connections are given by the following equations:

$$I_{A,B,LR,i} = \mu_{EE} W_{E,ij} S_{A,B,j} \tag{19}$$

$$I_{C,LR,i} = \mu_{IE} W_{I,ij} (S_{A,j} + S_{B,j}), \tag{20}$$

where W_E is the normalized long-range connectivity to excitatory neurons, and W_I is the normalized long-range connectivity to inhibitory neurons. μ_{EE} and μ_{IE} are coefficients scaling the long-range E to E and E to I connection strengths, respectively.

Here, we assume that the long-range connections will be scaled by a coefficient that is based on the hierarchy of source and target area. To quantify the difference between long-range feedforward and feedback projections, we introduce m_{ij} to measure the "feedforwardness" of projections between two areas. According to our assumption of counterstream inhibitory bias (CIB), long-range connections to inhibitory neurons are stronger for feedback connections and weaker for feedforward connections, while the opposite holds for long range connections to excitatory neurons. Following this hypothesis, we define m_{ij} as a sigmoid function of the difference between the hierarchy value of source and target areas. For feedforward projections, $m_{ij} > 0.5$; for feedback projections, $m_{ij} < 0.5$. Excitatory and inhibitory long-range connection strengths are implemented by multiplying the long-range connectivity strength W_{ij} by m_{ij} and $(1 - m_{ij})$, respectively:

$$m_{ij} = \frac{1}{1 + \beta e^{-(h_i - h_j)}} \tag{21}$$

$$W_{E,ij} = m_{ij} W_{ij} \tag{22}$$

$$W_{I,ij} = (1 - m_{ij}) W_{ij} \tag{23}$$

with

$$W_{ij} = (W_{norm,ij})^{k_{scale}} \tag{24}$$

The normalized connectivity $W_{norm,ij}$ is then rescaled to translate the broad range of connectivity values (over five orders of magnitude) to a range more suitable for our firing rate models. k_{scale} is the coefficient used for this scaling. $k_{scale} < 1$ effectively makes the range much smaller than the original normalized connectivity $W_{norm,ij}$.

Thalamocortical network model

Corticothalamic connectivity. We introduced thalamic areas in the network to examine their effect on cortical dynamics. Each thalamic area includes 2 excitatory populations, A and B, with no inhibitory population. These two populations share the same selectivity with the corresponding cortical areas. Unlike cortical areas, there are no recurrent connections between thalamic neurons (Sherman 2007). Thalamic currents have the following contributions (tc stands for thalamocortical connections and ct for corticothalamic connections):

$$I_{th,A,B} = I_{ct,A,B} + I_{th,0,A,B} + I_{th,noise,A,B} \tag{25}$$

where $I_{th,i}$ $(i = A, B)$ is the total current received by each thalamic population, $I_{ct,i}$ $(i = A, B)$ is the long range current from cortical areas to target thalamic area, $I_{th,0,i}$ $(i = A, B)$ is the background current for each population, and $I_{th,noise,i}$ $(i = A, B)$ is the noise input to thalamic population A and B, which we set to 0 in our simulations. $I_{ct,i}$ $(i = A, B)$ has the following form:

$$I_{ct,A,B,i} = g_{ct} W_{ct,E,ij} S_{k,j} \tag{26}$$

where $W_{ct,E,ij}$ is the LR connectivity to thalamic neurons, and $S_{k,j}$ is the synaptic variable of population k $(k = A, B)$ in cortical area j. Since all thalamic neurons are excitatory, we model corticothalamic projections as in the previous section:

$$m_{ct,ij} = \frac{1}{1 + \beta e^{-(h_{th,i} - h_j)}} \tag{27}$$

$$W_{ct,E,ij} = m_{ct,ij} W_{ct,ij} \tag{28}$$

$$\tag{29}$$

where

$$W_{ct,ij} = (W_{ct,norm,ij})^{k_{scale}} \tag{30}$$

$W_{ct,norm,ij}$ is the normalized connection strength from cortical area j to thalamic area i. $m_{ct,ij}$ is the coefficient quantifying how the long range connections target excitatory neurons based on cortical hierarchy h_j and thalamic hierarchy $h_{th,i}$.

The thalamic firing rates are described by:

$$\tau_r \frac{dr_{th,A,B}}{dt} = -r_{th,A,B} + \phi_{th,A,B}(I_{th,A,B}) \tag{31}$$

with the activation function for thalamic neurons given by:

$$\phi_{th,A,B}(I_{th,A,B}) = \frac{aI_{th,A,B} - b}{1 - exp[-d(aI_{th,A,B} - b)]} \tag{32}$$

Thalamic neurons are described by AMPA synaptic variables (Jaramillo et al. 2019):

$$\frac{dS_{th,A,B}}{dt} = -\frac{S_{th,A,B}}{\tau_A} + \gamma_A r_{th,A,B} \tag{33}$$

Thalamocortical connectivity. The connections from thalamic neurons to cortical neurons follow these equations

$$I_{tc,A,B,i} = g_{E,tc} W_{E,tc,ij} S_{th,A,B,j} \tag{34}$$

$$I_{tc,C,i} = g_{I,tc} W_{I,tc,ij} (S_{th,A,j} + S_{th,B,j}) \tag{35}$$

and connectivity

$$m_{tc,ij} = \frac{1}{1 + \beta e^{-(h_i - h_{th,j})}} \tag{36}$$

$$W_{E,tc,ij} = m_{tc,ij} W_{tc,ij} \tag{37}$$

$$W_{I,tc,ij} = (1 - m_{tc,ij}) W_{tc,ij} \tag{38}$$

and connectivity matrix

$$W_{tc,ij} = (W_{tc,norm,ij})^{k_{scale}} \tag{39}$$

The thalamocortical input is added to the total input current of each cortical population.

$$I_A = g_{E,self} S_A + g_{E,cross} S_B + g_{EI} S_C + I_{0A} + I_{LR,A} + I_{tc,A} + x_A(t) \tag{40}$$

$$I_B = g_{E,self} S_B + g_{E,cross} S_A + g_{EI} S_C + I_{0B} + I_{LR,B} + I_{tc,B} + x_B(t) \tag{41}$$

$$I_C = g_{IE} S_A + g_{IE} S_B + g_{II} S_C + I_{0C} + I_{LR,C} + I_{tc,C} + x_C(t) \tag{42}$$

Calculation of network structural measures

We considered three types of structural measures. The first one is input strength. Input strength of area i is the summation of the connection strengths onto node i. It quantifies the total external input onto area i.

$$W_{input,i} = \sum_{j=1}^{n} W_{ij} \tag{43}$$

The second one is eigenvector centrality (Newman 2018). Eigenvector centrality of area i is the ith element of the leading eigenvector of the connectivity matrix. It quantifies how many areas are connected with the target area i and how important these neighbors are.

$$W = Q \Lambda Q^{-1} \tag{44}$$

$$C_{eig,i} = q_{i1} \tag{45}$$

The third structural measure is loop strength, which quantifies how each area is involved in strong recurrent loops. We first define the strength of a single loop k

$$L_k = \prod_{A_i, A_j \in loop_k} w_{ij}, \tag{46}$$

and then the loop strength S_{A_i} of a single area A_i

$$S_{A_i} = \sum_{A_i \in loop_k} L_k \tag{47}$$

We now focus on cell type-specific structural measures. Cell type specificity is introduced via a coefficient k_{cell} that scales all long range connection strengths (cell type projection coefficient):

$$k_{cell} = m_{ij} - PV_i(1 - m_{ij}) \tag{48}$$

Thus, we can define cell type-specific input strength as:

$$W_{input,i,cellspec} = \sum_{j=1}^{n} (m_{ij} - PV_i(1 - m_{ij}))W_{ij} \tag{49}$$

Similarly, cell type-specific eigenvector centrality is defined as

$$\tilde{W}_{ij} = (m_{ij} - PV_i(1 - m_{ij}))w_{ij} \tag{50}$$

$$\tilde{W} = \tilde{Q}\tilde{\Lambda}\tilde{Q}^{-1} \tag{51}$$

$$C_{eig,i,cellspec} = \tilde{q}_{i1} \tag{52}$$

and cell type-specific loop strength:

$$L_{k,cellspec} = \prod_{A_i,A_j \in loop_k} (m_{ij} - PV_i(1 - m_{ij}))w_{ij} \tag{53}$$

$$S_{A_i,cellspec} = \sum_{A_i \in loop_k} L_{k,cellspec} \tag{54}$$

Stimulation protocol and inhibition analysis

We simulate a working memory task by applying an external current I_{stim} to one of the excitatory populations. The external current is a pulsed input with start time T_{on} and offset time T_{off}. Without losing generality, we assume that the external input is provided to population A. In most of the simulations in this study, we simulate a visual working memory task, with the external applied to VISp. The simulation duration is T_{trial} and we used a time step of dt.

We apply inhibition analysis to understand the robustness of attractors and, more importantly, to investigate which areas play an important role in maintaining the attractor state. Excitatory input was applied to the inhibitory population I to simulate opto-genetic inhibition. The external input I_{inh} is strong as compared to I_{stim} and results in an elevated firing rate of the inhibitory population, which in turn decreases the firing rate of the excitatory populations. Usually the inhibition is applied to a single area. When inhibition is applied during the stimulus period, its start and end times are equal to T_{on} and T_{off}, respectively. When inhibition is applied during delay period, its start time is later than T_{off} to allow the system settle to a stable state. Thus, the onset of inhibition starts 2 seconds after T_{off} and lasts until the end of trial. In the case of thalamocortical network simulations, we inhibit thalamic areas by introducing a hyperpolarizng current to both excitatory populations, since we do not have inhibitory populations in thalamic areas in the model.

To quantify the effect of single area or multiple areas inhibition, we calculate the average firing rate of areas that satisfy two conditions: i) the area shows persistent activity before

inhibition and ii) the area does not receive inhibitory input. The ratio between such average firing rate after inhibition and before inhibition is used to quantify the overall effect of inhibition. If the ratio is lower than 100%, this suggests that inhibiting certain area(s) disrupts the maintenance of the attractor state. Note that the inhibition effect is typically not very strong, and only in rare cases, inhibition of a single area leads to loss of activity of other areas (Fig. 5B, Fig. 5C). To quantify such differences, we use a threshold of 5% to differentiate them. We will use (relatively) "weak inhibition effect" and "strong inhibition effect" to refer to them afterwards.

We used the three measures to classify areas into 4 types (Fig. 5D): i) inhibition effect during delay period, ii) inhibition effect during stimulus period, and iii) delay period firing rate. Areas with strong inhibition effect during stimulus period are classified as input areas; areas with strong inhibition effect during delay period and strong delay period firing rate are classified as core areas; areas with weak inhibition effect during delay period but strong firing rate are classified as readout areas; areas with weak inhibition effect during delay period and weak firing rate during delay period are classified as nonessential areas.

Simulation of multiple attractors

Multiple attractors coexist in the network and its properties and number depends on the connectivity and dynamics of each node. In this study we did not try to capture all the possible attractors in the network, but rather compare the number of attractors for different networks. Here we briefly describe the protocol used to identify multiple attractors in the network. We first choose k areas and then generate a subset of areas as the stimulation areas. We cover all possible subsets, which means we run 2^k simulations in total. The external stimulus is given to all areas in the subset simultaneously with same strength and duration. The delay period activity is then quantified using a similar protocol as the standard simulation protocol. The selection of k areas corresponds to a qualitative criterion. First we choose the areas with small PV fraction or high hierarchy, since these areas are more likely to show persistent activity. Second, the number of possible combination grows exponentially as we increase k, and if we use k = 43, the number of combinations is around 8.8e+12, which is beyond our simulation power. As a trade-off between the simulation power and coverage of areas, we choose k = 18, which correspond to 2.6e+5 different combinations of stimulation. For each parameter setting, we run 2.6e+5 simulations to capture possible attractor patterns. For each attractor pattern, a binary vector is generated by thresholding delay firing rate using a firing rate threshold of 5Hz. An attractor pattern is considered distinct if and only if the binary vector is different from all identified attractors. In these way we can identify different attractors in the simulation. We also apply same simulation pipeline to identify attractors for different parameters. Specifically we change the long range connectivity strength μ_{EE} and local excitatory connections $g_{E,self}$.

References

Abbott, L. F. et al. (2020). "The mind of a mouse". *Cell* 182, pp. 1372–1376.

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

Abbott, Larry F. et al. (2017). "An International Laboratory for Systems and Computational Neuroscience". *Neuron* 96.6, pp. 1213–1218.

Baddeley, Alan (2012). "Working Memory: Theories, Models, and Controversies." *Annual review of psychology* 63.1, pp. 1–29. pmid: 21961947.

Bolkan, Scott S., Joseph M. Stujenske, Sebastien Parnaudeau, Timothy J. Spellman, Caroline Rauffenbart, Atheir I. Abbas, Alexander Z. Harris, Joshua A. Gordon, and Christoph Kellendonk (2017). "Thalamic Projections Sustain Prefrontal Activity during Working Memory Maintenance". *Nature Neuroscience* 20.7 (7), pp. 987–996.

Cabral, Joana, Etienne Hugues, Olaf Sporns, and Gustavo Deco (2011). "Role of Local Network Oscillations in Resting-State Functional Connectivity". *NeuroImage* 57.1, pp. 130–139.

Chaudhuri, Rishidev, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang (2015). "A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex". *Neuron* 88.2, pp. 419–431.

Christophel, Thomas B., P. Christiaan Klink, Bernhard Spitzer, Pieter R. Roelfsema, and John-Dylan Haynes (2017). "The Distributed Nature of Working Memory". *Trends in Cognitive Sciences* 21.2, pp. 111–124.

D. J. Amit (1995). "The Hebbian paradigm reintegrated: local reverberations as internal representations". *Behav. Brain Sci.* 18, pp. 617–626.

Deco, Gustavo, Adrián Ponce-Alvarez, Patric Hagmann, Gian Luca Romani, Dante Mantini, and Maurizio Corbetta (2014). "How Local Excitation–Inhibition Ratio Impacts the Whole Brain Dynamics". *Journal of Neuroscience* 34.23, pp. 7886–7898. pmid: 24899711.

Demirtaş, Murat, Joshua B. Burt, Markus Helmer, Jie Lisa Ji, Brendan D. Adkinson, Matthew F. Glasser, David C. Van Essen, Stamatios N. Sotiropoulos, Alan Anticevic, and John D. Murray (2019). "Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics". *Neuron* 101.6, 1181–1194.e13.

Diesmann, Markus, Marc-Oliver Gewaltig, and Ad Aertsen (1999). "Stable Propagation of Synchronous Spiking in Cortical Neural Networks". *Nature* 402.6761 (6761), pp. 529–533.

Dotson, Nicholas M., Steven J. Hoffman, Baldwin Goodell, and Charles M. Gray (2018). "Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized". *Neuron* 99.1, 215–226.e4.

Egorov, Alexei V., Bassam N. Hamam, Erik Fransén, Michael E. Hasselmo, and Angel A. Alonso (2002). "Graded Persistent Activity in Entorhinal Cortex Neurons". *Nature* 420.6912, pp. 173–178.

Elston, Guy N. (2007). "Specialization of the Neocortical Pyramidal Cell during Primate Evolution". In: *Evolution of Nervous Systems*. Elsevier, pp. 191–242.

Erlich, Jeffrey C., Max Bialek, and Carlos D. Brody (2011). "A Cortical Substrate for Memory-Guided Orienting in the Rat". *Neuron* 72.2, pp. 330–343.

Erö, Csaba, Marc-Oliver Gewaltig, Daniel Keller, and Henry Markram (2018). "A Cell Atlas for the Mouse Brain". *Frontiers in Neuroinformatics* 12, p. 84.

Felleman, D J and D C Van Essen (1991). "Distributed Hierarchical Processing in the Primate Cerebral Cortex." *Cerebral Cortex* 1.1, pp. 1–47. pmid: 1822724.

Froudist-Walsh, Sean, Daniel P. Bliss, Xingyu Ding, Lucija Rapan, Meiqi Niu, Kenneth Knoblauch, Karl Zilles, Henry Kennedy, Nicola Palomero-Gallagher, and Xiao-Jing Wang (2021a). "A Dopamine Gradient Controls Access to Distributed Working Memory in the Large-Scale Monkey Cortex". *Neuron* 109.21, 3500–3520.e13.

Froudist-Walsh, Sean, Ting Xu, Meiqi Niu, Lucija Rapan, Karl Zilles, Daniel S. Margulies, Xiao-Jing Wang, and Nicola Palomero-Gallagher (2021b). "Gradients of Receptor Expression in the Macaque Cortex". *bioRxiv*, p. 2021.02.22.432173.

Fulcher, Ben D., John D. Murray, Valerio Zerbi, and Xiao-Jing Wang (2019). "Multimodal Gradients across Mouse Cortex". *Proceedings of the National Academy of Sciences* 116.10, pp. 4689–4695. pmid: 30782826.

Funahashi, Shintaro, Charles J. Bruce, and Patricia S. Goldman-Rakic (1989). "Mnemonic Coding of Visual Space in the Monkey's Dorsolateral Prefrontal Cortex". *Journal of Neurophysiology* 61.2, pp. 331–349.

Fuster, Joaquin M. and Garrett E. Alexander (1971). "Neuron Activity Related to Short-Term Memory". *Science* 173.3997, pp. 652–654. pmid: 4998337.

Gămănuţ, Răzvan, Henry Kennedy, Zoltán Toroczkai, Mária Ercsey-Ravasz, David C. Van Essen, Kenneth Knoblauch, and Andreas Burkhalter (2018). "The Mouse Cortical Connectome, Characterized by an Ultra-Dense Cortical Graph, Maintains Specificity by Distinct Connectivity Profiles". *Neuron* 97.3, 698–715.e10.

Gao, Zhenyu, Courtney Davis, Alyse M Thomas, Michael N Economo, Amada M Abrego, Karel Svoboda, Chris I De Zeeuw, and Nuo Li (2018). "A Cortico-Cerebellar Loop for Motor Planning." *Nature* 39, p. 1062. pmid: 30333626.

Gilad, Ariel, Yasir Gallero-Salas, Dominik Groos, and Fritjof Helmchen (2018). "Behavioral Strategy Determines Frontal or Posterior Location of Short-Term Memory in Neocortex". *Neuron* 99.4, 814–828.e7.

Gilman, Joshua P., Maria Medalla, and Jennifer I. Luebke (2017). "Area-Specific Features of Pyramidal Neurons—a Comparative Study in Mouse and Rhesus Monkey". *Cerebral Cortex* 27.3, pp. 2078–2094.

Goldman-Rakic, P. S (1995). "Cellular Basis of Working Memory". *Neuron* 14.3, pp. 477–485.

Guo, Zengcai V., Hidehiko K. Inagaki, Kayvon Daie, Shaul Druckmann, Charles R. Gerfen, and Karel Svoboda (2017). "Maintenance of Persistent Activity in a Frontal Thalamocortical Loop". *Nature* 545.7653 (7653), pp. 181–186.

Guo, Zengcai V., Nuo Li, Daniel Huber, Eran Ophir, Diego Gutnisky, Jonathan T. Ting, Guoping Feng, and Karel Svoboda (2014). "Flow of Cortical Activity Underlying a Tactile Decision in Mice". *Neuron* 81.1, pp. 179–194.

Harris, Julie A. et al. (2019). "Hierarchical Organization of Cortical and Thalamic Connectivity". *Nature* 575.7781 (7781), pp. 195–202.

Harvey, Christopher D., Philip Coen, and David W. Tank (2012). "Choice-Specific Sequences in Parietal Cortex during a Virtual-Navigation Decision Task". *Nature* 484.7392, pp. 62–68. pmid: 22419153.

Honey, Christopher J., Rolf Kötter, Michael Breakspear, and Olaf Sporns (2007). "Network Structure of Cerebral Cortex Shapes Functional Connectivity on Multiple Time Scales". *Proceedings of the National Academy of Sciences* 104.24, pp. 10240–10245. pmid: 17548818.

Huang, Chengcheng, Douglas A. Ruff, Ryan Pyle, Robert Rosenbaum, Marlene R. Cohen, and Brent Doiron (2019). "Circuit Models of Low-Dimensional Shared Variability in Cortical Networks". *Neuron* 101.2, 337–348.e4.

Inagaki, Hidehiko K., Lorenzo Fontolan, Sandro Romani, and Karel Svoboda (2019). "Discrete Attractor Dynamics Underlies Persistent Activity in the Frontal Cortex". *Nature* 566.7743 (7743), pp. 212–217.

Inagaki, Hidehiko K., Miho Inagaki, Sandro Romani, and Karel Svoboda (2018). "Low-Dimensional and Monotonic Preparatory Activity in Mouse Anterior Lateral Motor Cortex". *The Journal of Neuroscience* 38.17, pp. 4163–4185.

Jaramillo, Jorge, Jorge F. Mejias, and Xiao-Jing Wang (2019). "Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations". *Neuron* 101.2, 321–336.e9.

Javadzadeh, Mitra and Sonja B. Hofer (2022). "Dynamic Causal Communication Channels between Neocortical Areas". *Neuron* 0.0.

Joglekar, Madhura R., Jorge F. Mejias, Guangyu Robert Yang, and Xiao-Jing Wang (2018). "Inter-Areal Balanced Amplification Enhances Signal Propagation in a Large-Scale Circuit Model of the Primate Cortex". *Neuron* 98.1, 222–234.e8. pmid: 29576389.

Jones, Edward G. (2007). "Neuroanatomy: Cajal and after Cajal". *Brain Research Reviews* 55.2, pp. 248–255.

Jun, James J. et al. (2017). "Fully Integrated Silicon Probes for High-Density Recording of Neural Activity". *Nature* 551.7679 (7679), pp. 232–236.

Kim, Yongsoo, Guangyu Robert Yang, Kith Pradhan, Kannan Umadevi Venkataraju, Mihail Bota, Luis Carlos García del Molino, Greg Fitzgerald, Keerthi Ram, Miao He, Jesse Maurica Levine, Partha Mitra, Z. Josh Huang, Xiao-Jing Wang, and Pavel Osten (2017). "Brain-Wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism". *Cell* 171.2, 456–469.e22.

Klatzmann, Ulysse, Sean Froudist-Walsh, Daniel P. Bliss, Panagiota Theodoni, Jorge F. Mejias, Meiqi Niu, Lucija Rapan, Nicola Palomero-Gallagher, Claire Sergent, Stanislas

Dehaene, and Xiao-Jing Wang (2022). "A Connectome-Based Model of Conscious Access in Monkey Cortex". *bioRxiv*, p. 2022.02.20.481230.

Knox, Joseph E., Kameron Decker Harris, Nile Graddis, Jennifer D. Whitesell, Hongkui Zeng, Julie A. Harris, Eric Shea-Brown, and Stefan Mihalas (2018). "High-Resolution Data-Driven Model of the Mouse Connectome". *Network Neuroscience* 3.1, pp. 217–236. pmid: 30793081.

Kopec, Charles D., Jeffrey C. Erlich, Bingni W. Brunton, Karl Deisseroth, and Carlos D. Brody (2015). "Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements". *Neuron* 88.2, pp. 367–377. pmid: 26439529.

Leavitt, Matthew L., Diego Mendoza-Halliday, and Julio C. Martinez-Trujillo (2017). "Sustained Activity Encoding Working Memories: Not Fully Distributed". *Trends in Neurosciences* 40.6, pp. 328–346.

Li, Nuo, Tsai-Wen Chen, Zengcai V. Guo, Charles R. Gerfen, and Karel Svoboda (2015). "A Motor Cortex Circuit for Motor Planning and Movement". *Nature* 519.7541 (7541), pp. 51–56.

Li, Nuo, Kayvon Daie, Karel Svoboda, and Shaul Druckmann (2016). "Robust Neuronal Dynamics in Premotor Cortex during Motor Planning". *Nature* 532.7600, pp. 459–464.

Liu, Ding, Xiaowei Gu, Jia Zhu, Xiaoxing Zhang, Zhe Han, Wenjun Yan, Qi Cheng, Jiang Hao, Hongmei Fan, Ruiqing Hou, Zhaoqin Chen, Yulei Chen, and Chengyu T. Li (2014). "Medial Prefrontal Activity during Delay Period Contributes to Learning of a Working Memory Task". *Science* 346.6208, pp. 458–463.

Markov, Nikola T, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, Pascal Barone, Colette Dehay, Kenneth Knoblauch, and Henry Kennedy (2014a). "Anatomy of Hierarchy: Feedforward and Feedback Pathways in Macaque Visual Cortex." *Journal of Comparative Neurology* 522.1, pp. 225–259. pmid: 23983048.

Markov, Nikola T. et al. (2014b). "A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex". *Cerebral Cortex* 24.1, pp. 17–36.

Mejias, Jorge F. and Xiao-Jing Wang (2022). "Mechanisms of Distributed Working Memory in a Large-Scale Network of Macaque Neocortex". *eLife* 11, e72136.

Murphy, B. K. and K. D. Miller (2009). "Balanced amplification: a new mechanism of selective amplification of neural activity patterns". *Neuron* 61, pp. 635–648.

Murray, John D., Jorge Jaramillo, and Xiao-Jing Wang (2017). "Working Memory and Decision-Making in a Frontoparietal Circuit Model". *Journal of Neuroscience* 37.50, pp. 12167–12186. pmid: 29114071.

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

Newman, M. E. J. (2018). *Networks*. Second edition. Oxford: Oxford University Press.

Oh, Seung Wook et al. (2014). "A Mesoscale Connectome of the Mouse Brain". *Nature* 508.7495 (7495), pp. 207–214.

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

Petreanu, Leopoldo, Tianyi Mao, Scott M Sternson, and Karel Svoboda (2009). "The Subcellular Organization of Neocortical Excitatory Connections." *Nature* 457.7233, pp. 1142–1145. pmid: 19151697.

Pinto, Lucas, Kanaka Rajan, Brian DePasquale, Stephan Y. Thiberge, David W. Tank, and Carlos D. Brody (2019). "Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions". *Neuron* 104.4, 810–824.e9.

R. J. Douglas, C. Koch, M. Mahowald, K. M. Martin, and H. H. Suarez (1995). "Recurrent excitation in neocortical circuits". *Science* 269, pp. 981–985.

Schmidt, Maximilian, Rembrandt Bakker, Kelly Shen, Gleb Bezgin, Markus Diesmann, and Sacha Jennifer van Albada (2018). "A Multi-Scale Layer-Resolved Spiking Network Model of Resting-State Dynamics in Macaque Visual Cortical Areas". *PLOS Computational Biology* 14.10, e1006359.

Schmitt, L. Ian, Ralf D. Wimmer, Miho Nakajima, Michael Happ, Sima Mofakham, and Michael M. Halassa (2017). "Thalamic Amplification of Cortical Connectivity Sustains Attentional Control". *Nature* 545.7653 (7653), pp. 219–223.

Sherman, S Murray (2007). "The Thalamus Is More than Just a Relay". *Current opinion in neurobiology* 17.4, pp. 417–422.

Shine, James M, Matthew J Aburn, Michael Breakspear, and Russell A Poldrack (2018). "The Modulation of Neural Gain Facilitates a Transition between Functional Segregation and Integration in the Brain". *eLife* 7. Ed. by Gustavo Deco, e31130.

Steinmetz, Nicholas A., Peter Zatka-Haas, Matteo Carandini, and Kenneth D. Harris (2019). "Distributed Coding of Choice, Action and Engagement across the Mouse Brain". *Nature* 576.7786 (7786), pp. 266–273.

Steinmetz, Nicholas A. et al. (2021). "Neuropixels 2.0: A Miniaturized High-Density Probe for Stable, Long-Term Brain Recordings". *Science* 372.6539, eabf4588.

Stringer, Carsen, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris (2019). "Spontaneous Behaviors Drive Multidimensional, Brainwide Activity". *Science* 364.6437, eaav7893.

Suzuki, Mototaka and Jacqueline Gottlieb (2013). "Distinct Neural Mechanisms of Distractor Suppression in the Frontal and Parietal Lobe". *Nature Neuroscience* 16.1 (1), pp. 98–104.

Voitov, Ivan and Thomas D. Mrsic-Flogel (2022). "Cortical Feedback Loops Bind Distributed Representations of Working Memory". *Nature*, pp. 1–9.

Wang, Peng, Ru Kong, Xiaolu Kong, Raphaël Liégeois, Csaba Orban, Gustavo Deco, Martijn P. van den Heuvel, and B. T. Thomas Yeo (2019). "Inversion of a Large-Scale Circuit Model Reveals a Cortical Hierarchy in the Dynamic Resting Human Brain". *Science Advances* 5.1, eaat7854.

Wang, Quanxin et al. (2020). "The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas". *Cell* 181.4, 936–953.e20. pmid: 32386544.

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

Wang, Xiao-Jing (1999). "Synaptic Basis of Cortical Persistent Activity: The Importance of NMDA Receptors to Working Memory". *Journal of Neuroscience* 19.21, pp. 9587–9603.

Wang, Xiao-Jing (2001). "Synaptic Reverberation Underlying Mnemonic Persistent Activity". *Trends in Neurosciences* 24.8, pp. 455–463.

Wang, Xiao-Jing (2020). "Macroscopic Gradients of Synaptic Excitation and Inhibition in the Neocortex". *Nature Reviews Neuroscience* 21.3, pp. 169–178.

Wang, Xiao-Jing (2021). "50 years of mnemonic persistent activity: Quo vadis?" *Trends in Neurosci.* 44, pp. 888–902.

Wang, Xiao-Jing (2022). "Theory of the Multiregional Neocortex: Large-Scale Neural Dynamics and Distributed Cognition". *Annual Review of Neuroscience* 45.1, pp. 533–560.

Wang, Yu, Xinxin Yin, Zhouzhou Zhang, Jiejue Li, Wenyu Zhao, and Zengcai V. Guo (2021). "A Cortico-Basal Ganglia-Thalamo-Cortical Channel Underlying Short-Term Memory". *Neuron* 109.21, 3486–3499.e7.

Wong, Kong-Fatt and Xiao-Jing Wang (2006). "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions". *Journal of Neuroscience* 26.4, pp. 1314–1328. pmid: 16436619.

Wu, Zheng, Ashok Litwin-Kumar, Philip Shamash, Alexei Taylor, Richard Axel, and Michael N. Shadlen (2020). "Context-Dependent Decision Making in a Premotor Circuit". *Neuron* 106.2, 316–328.e6.

Xu, Yaoda (2017). "Reevaluating the Sensory Account of Visual Working Memory Storage". *Trends in Cognitive Sciences* 21.10, pp. 794–815.

Yizhar, Ofer, Lief E. Fenno, Thomas J. Davidson, Murtaza Mogri, and Karl Deisseroth (2011). "Optogenetics in Neural Systems". *Neuron* 71.1, pp. 9–34.

Yoo, Sang-Ah, Julio C. Martinez-Trujillo, Stefan Treue, John K. Tsotsos, and Mazyar Fallah (2021). "Feature-Based Attention Induces Surround Suppression during the Perception of Visual Motion". *bioRxiv*, p. 2021.02.17.431646.

Zandvakili, Amin and Adam Kohn (2015). "Coordinated Neuronal Activity Enhances Corticocortical Communication." *Neuron* 87.4, pp. 827–839. pmid: 26291164.

Zhang, Xiaoxing, Wenjun Yan, Wenliang Wang, Hongmei Fan, Ruiqing Hou, Yulei Chen, Zhaoqin Chen, Chaofan Ge, Shumin Duan, Albert Compte, and Chengyu T Li (2019). "Active Information Maintenance in Working Memory by a Sensory Cortex". *eLife* 8, e43191.

Zhu, Jia, Qi Cheng, Yulei Chen, Hongmei Fan, Zhe Han, Ruiqing Hou, Zhaoqin Chen, and Chengyu T. Li (2020). "Transient Delay-Period Activity of Agranular Insular Cortex Controls Working Memory Maintenance in Learning Novel Tasks". *Neuron* 105.5, 934–946.e5.

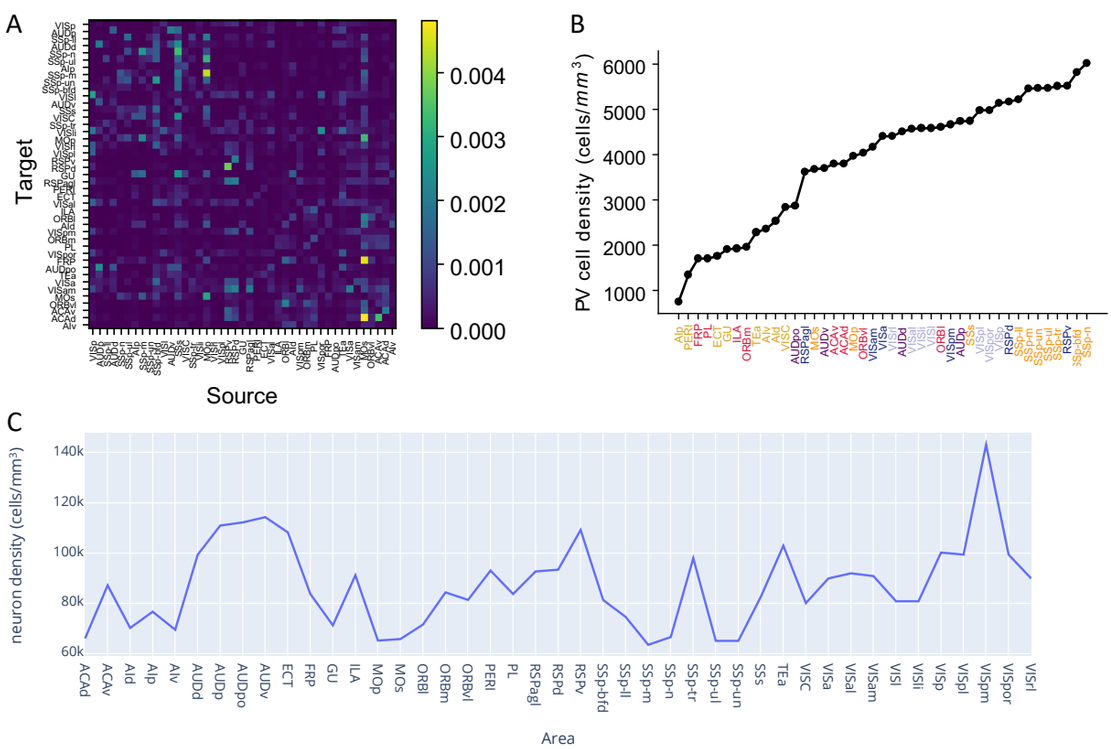

Figure1 - Supplement 1. Anatomical details of the mouse cortex. (A). Connectivity matrix depicting cortico-cortical connections between 43 cortical areas. Areas are sorted according to their hierarchy. (B). The raw PV cell density for each cortical area (Y axis), with areas sorted (X axis). Each area belongs to one of five modules, shown in color (see also Figure 1). (Harris et al. 2019). (C). Neuron density for each cortical area. The data is from Erö et al. 2018.

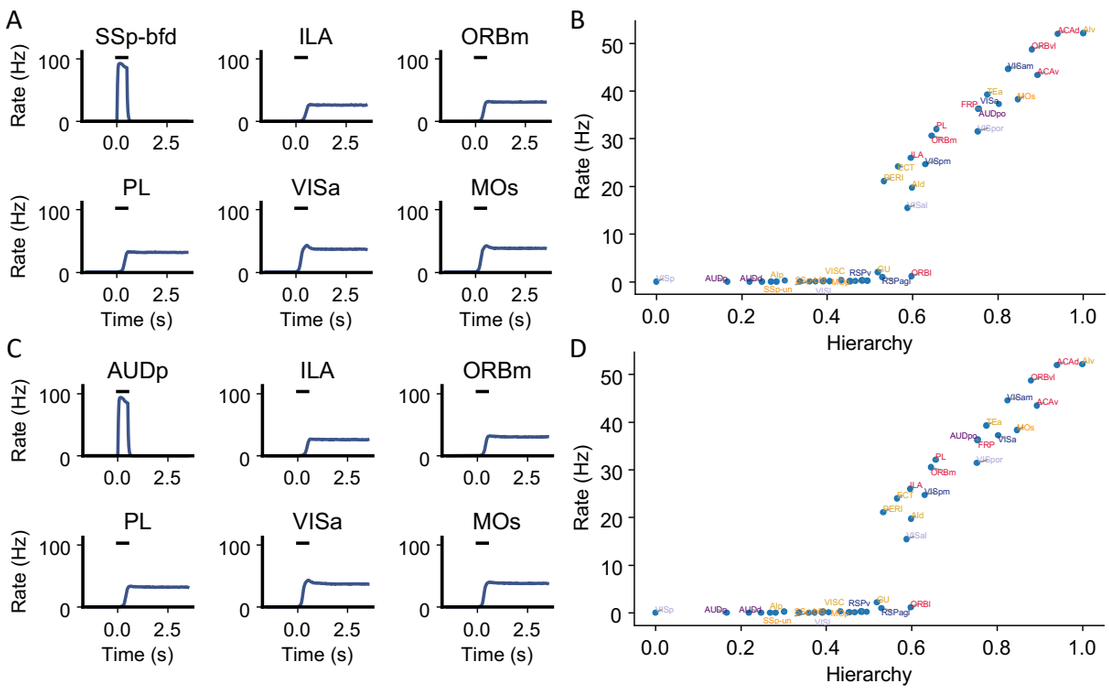

Figure2 - Supplement 1. Example simulation for different sensory modalities. The simulation protocol is the same as the default one in Figure 2, except that the external input is applied to primary sensory areas related to two other sensory modalities: somatosensory and auditory. (A). The activity of 6 selected areas during the working memory task is shown. A somatosensory input of 500ms is applied to primary somatosensory area SSp-bfd, which propagates to the rest of the large-scale network. (B). Similar to the simulation where a primary visual area is stimulated (Fig 2D), delay period firing is positively correlated with cortical hierarchy. (r = 0.89, p<0.05) (C) and (D) are similar to (A) and (B) except that the input is given to primary auditory area AUDp. (D). Delay period firing is also positively correlated with cortical hierarchy. (r = 0.89, p<0.05)

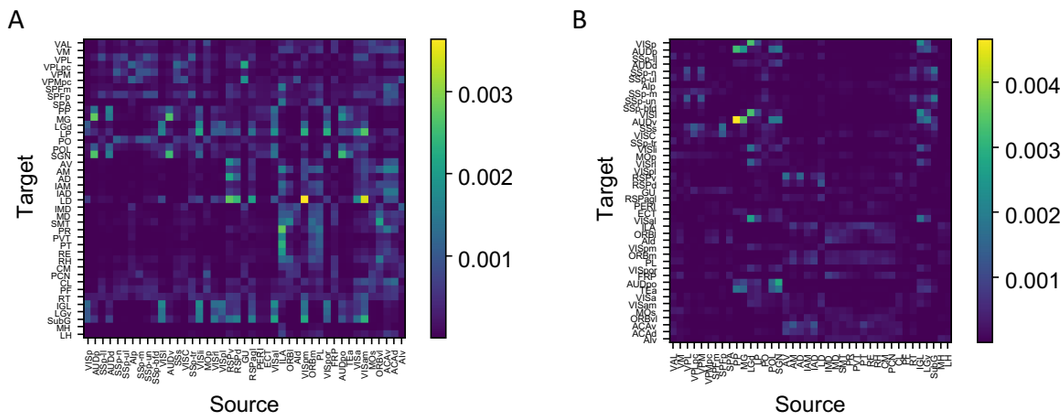

Figure 3 - Supplement 1. Anatomical data of thalamus and cortical connectivity. (A). Connectivity matrix of corticothalamic connections: 43 cortical areas to 40 thalamic areas. (B). Connectivity matrix of thalamocortical connections: 40 thalamic areas to 43 cortical areas.

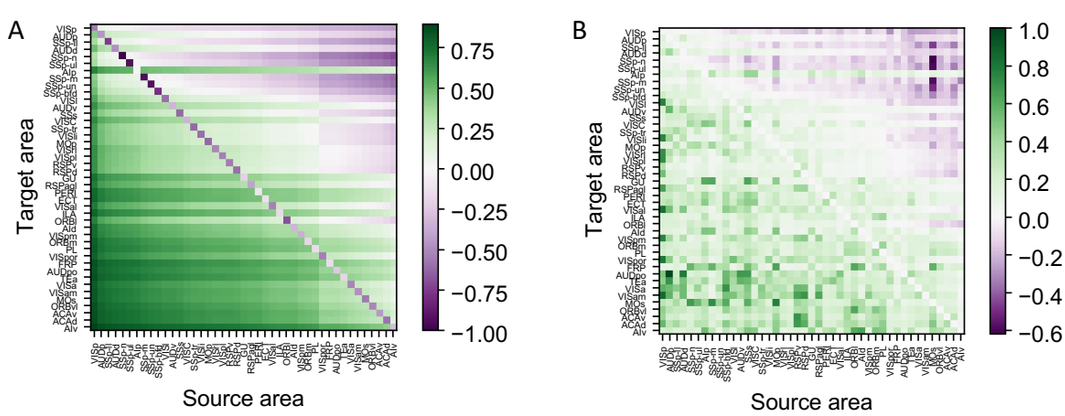

Figure 4 - Supplement 1. Details of cell type-specific connectivity measures. (A). The matrix of cell type projection coefficients between cortical areas. The cell type projection coefficient is given by the formula $k_{cell} = m_{ij} - PV_i(1 - m_{ij})$. (B). The matrix of connectivity strengths, modified by cell type projection coefficient between cortical areas. The modified connectivity strength is given by $\tilde{W}_{ij} = (m_{ij} - PV_i(1 - m_{ij}))w_{ij}$.

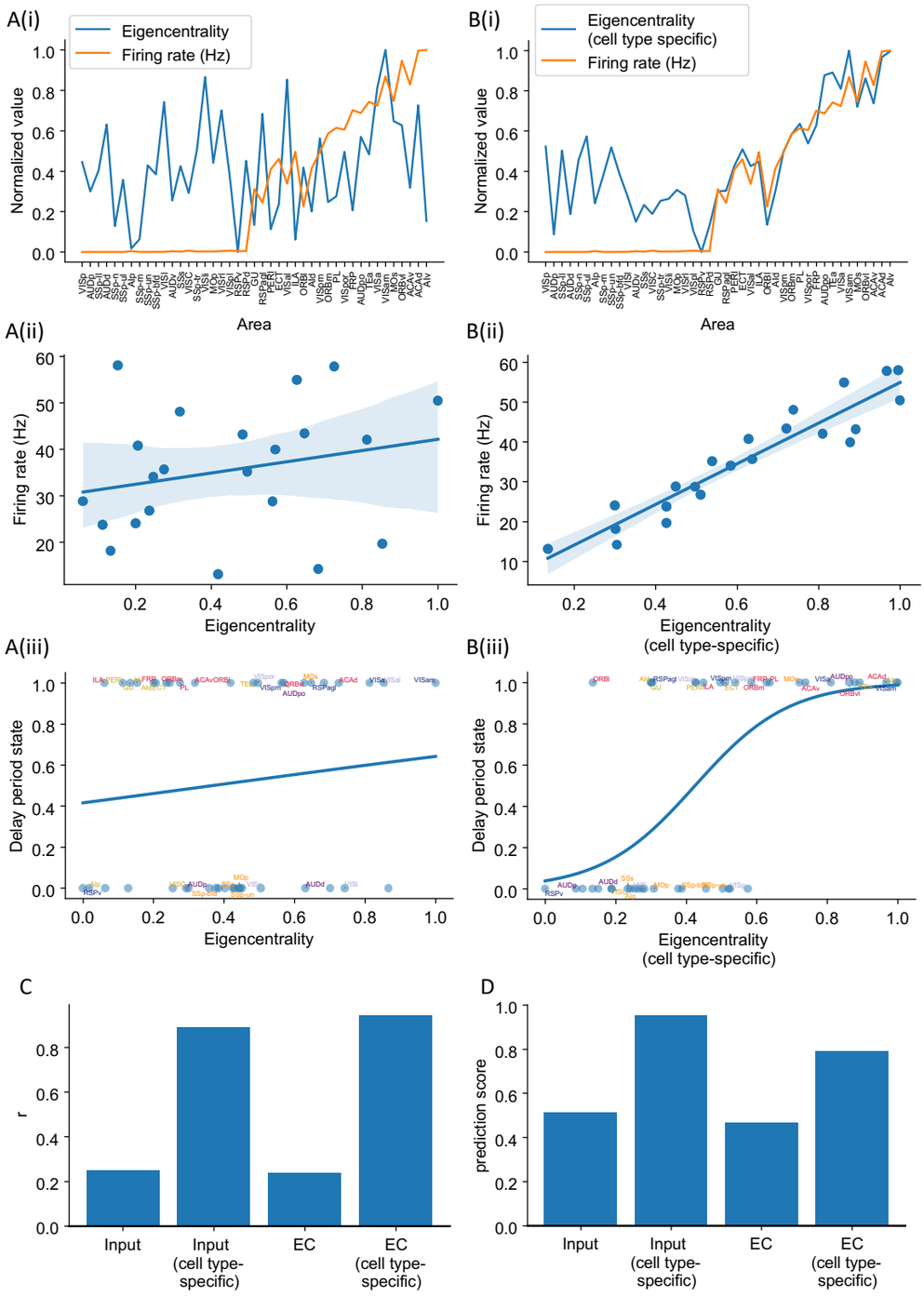

Figure 4 - Supplement 2. Cell type-specific eigenvector centrality measures are better at predicting firing rate patterns than raw eigenvector centrality measures. The analysis is the same as in Figure 4, where we compared cell type-specific input strength and raw input strength. Eigenvector centrality (EC, eigencentrality) of area i is the ith element of the leading eigenvector of the connectivity matrix. It quantifies how many areas are connected with the target area i and how important are these neighbors. Details are in the Methods section. (A(i)). Delay period firing rate (orange) and eigenvector centrality for each cortical area (blue). (A(ii)). Eigenvector centrality does not show a significant correlation with delay period firing rate for areas showing persistent activity in the model ($r = 0.24$, $p = 0.29$). (A(iii)). Eigenvector centrality cannot be used to predict whether an area shows persistent activity or not (prediction accuracy = 0.46). (B(i)). Delay period firing rate (orange) and cell type-specific eigenvector centrality for each cortical area (blue). (B(ii)). Cell type-specific eigenvector centrality has a strong correlation with the firing rate of cortical areas showing persistent activity ($r = 0.94$, $p < 0.05$). (B(iii)). Cell type-specific eigenvector centrality predicts whether an area shows persistent activity or not (prediction accuracy = 0.79). (C). Comparison of the correlation coefficient r for raw eigenvector centrality and cell type-specific eigenvector centrality in predicting delay firing rate. Raw input strength and cell type-specific input strength are also included for comparison. (D). Comparison of the prediction accuracy for raw eigenvector centrality and cell type-specific eigenvector centrality. Raw input strength and cell type-specific input strength are also included for comparison.

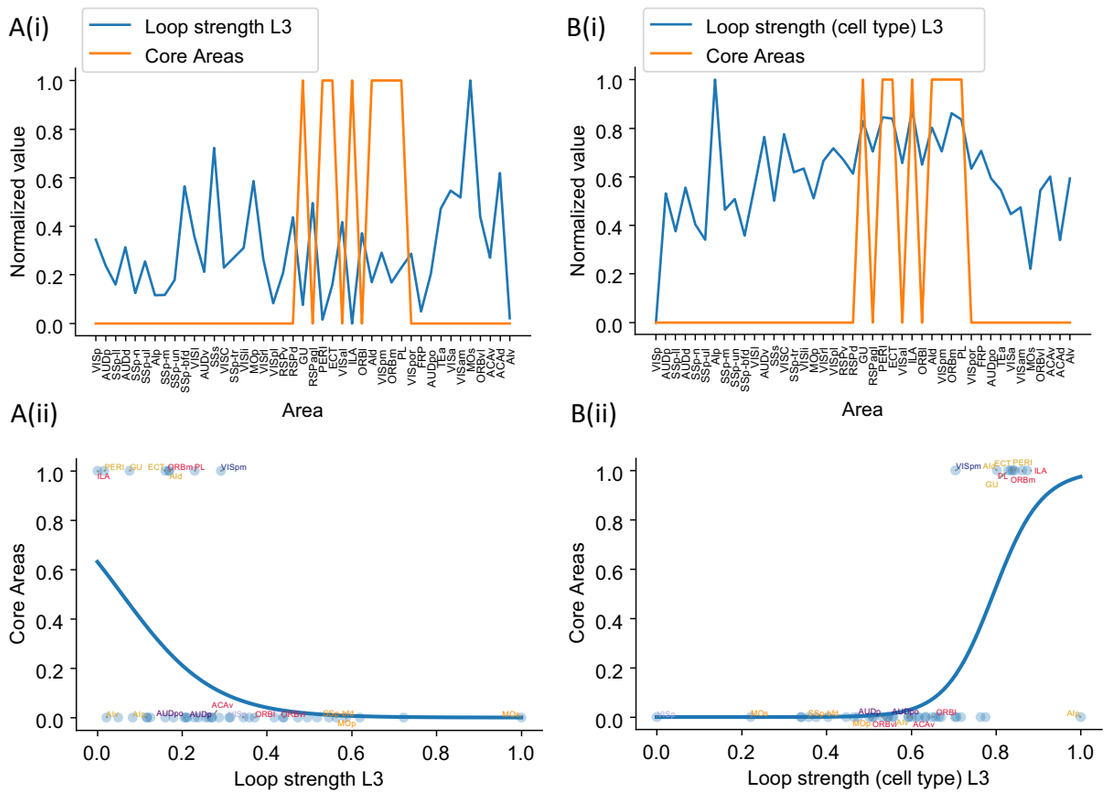

Figure 6 - Supplement 1. Cell type-specific loop strengths (Length 3 loops) are also better at predicting firing rate patterns than raw loop measures. Loop strengths (length 3 loops or L3) is calculated using similar method as loop strengths (length 2 loops). The only difference is we considered loops with length 3 (eg. A1->A2->A3->A1). The analysis is the same as in Figure 6, where we compared cell type-specific loop strengths (length 2 loops) and raw loop strengths. (A(i)). Loop strength (blue) is plotted alongside Core Areas (orange), a binary variable that takes the value 1 if the area is indeed a Core Area, 0 otherwise. Loop strength is normalized to a range of (0, 1) for better comparison. (A(ii)). A high loop strength value does not imply that an area is a Core Area. (B(i)). Same as (A), but for cell type-specific loop strength. (B(ii)). High cell type-specific loop measures predicts that an area is a Core Area (prediction accuracy = 0.95). Same as (A), but for cell type-specific loop strength.

Area	Supporting literature
ALM (MOs)	(Kopec et al. 2015; Guo et al. 2014; Li et al. 2016), (Inagaki et al. 2019; Erlich et al. 2011; Guo et al. 2017), (Gilad et al. 2018; Gao et al. 2018; Wu et al. 2020)
mPFC (PL/ILA)	(Liu et al. 2014; Schmitt et al. 2017), (Bolkan et al. 2017)
OFC	(Wu et al. 2020)
PPC (VISa)	(Harvey et al. 2012)
AIa (AId,AIv)	(Zhu et al. 2020)
Area p (VISpl)	(Gilad et al. 2018)
dorsal cortex	(Pinto et al. 2019)
entorhinal (in vitro persistent activity)	(Egorov et al. 2002)
piriform	(Zhang et al. 2019; Wu et al. 2020)
VM/VAL	(Guo et al. 2017)
MD	(Schmitt et al. 2017; Bolkan et al. 2017)
superior colliculus	(Kopec et al. 2015)
cerebellar nucleus	(Gao et al. 2018)

Table 1: Supplementary experimental evidence. The listed literature include experiments that provide supporting evidence for working memory activity in cortical and subcortical brain areas in the mouse or rat. These studies show either that a given area is involved in working memory tasks and/or exhibit delay period activity. Area name corresponds to what has been reported in the literature. Some areas do not correspond exactly to the names from the Allen common coordinate framework.

Parameter	Description	Task/Figure	Value
	Cortical circuit parameters		
τ_{NMDA}	NMDA synapse time constant	All figures	60 ms
τ_{GABA}	GABA synapse time constant	All figures	5 ms
τ_{AMPA}	AMPA synapse time constant	All figures	2 ms
τ_{rates}	neuron time constant	All figures	20 ms
τ_{noise}	noise time constant	All figures	2 ms
a, b, d	parameters in excitatory F-I curve.	All figures	140 Hz/nA, 54 Hz, 308 ms
g_I, c_1, c_0, r_{0I}	parameters in inhibitory F-I curve.	All figures	4, 615 Hz/nA, 177 Hz, 5.5 Hz
γ	parameters in NMDA excitatory synaptic equations.	All figures	1.282
γ_I	parameters in GABA synaptic equations.	All figures	2
γ_A	parameters in AMPA excitatory synaptic equations.	All figures	2
$g_{E,self}$	local self excitatory connections	Figures 1-6	0.4 nA
$g_{E,cross}$	local cross population excitatory connections	All figures	10.7 pA
g_{IE}	local E to I connections	All figures	0.4087 nA
$J_{EI,min}, J_{EI,scaling}$	local I to E connection strength	All figures	0.192 nA, 0.16 nA
$J_{II,min}, J_{II,scaling}$	local I to I connection strength	All figures	0.105 nA, 0.075 nA
I_{0A}, I_{0B}	background current for excitatory neurons	All figures	0.305 nA
I_{0C}	background current for inhibitory neurons	All figures	0.26 nA
σ_A, σ_B	standard deviation of excitatory noise current	All figures	5 pA
σ_C	standard deviation of inhibitory noise current	All figures	0 pA
r_{0E}	background current for excitatory neurons	All figures	5 Hz
r_{0I}	background current for excitatory neurons	All figures	5.5 Hz
μ_{EE}	long range E to E connection strength	Figures 1, 2, 4, 5, 6	0.1 nA
μ_{IE}	long range E to I connection strength	Figures 1, 2, 4, 5, 6	0.167 nA
β	parameters in m_{ij}	All figures	2.42
k_{scale}	parameters for scaling the connectivity matrix	All figures	0.3
α_h, β_h	parameters for estimation of hierarchy	All figures	1.33, -0.22
I_{stim}	external stimulus strength	All figures	0.5 nA
I_{inh}	external input to inhibitory neurons	All figures	5 nA
T_{on}	stimulus start time	All figures	2 s
T_{off}	stimulus end time	All figures	2.5 s
T_{trial}	simulation time for each trial	All figures	10 s
dt	simulation time step	All figures	0.5 ms
	Thalamocortical network		
μ_{EE}	long range E to E connection strength	Figure 3	0.01 nA
μ_{IE}	long range E to I connection strength	Figure 3	0.0167 nA
g_{ct}	cortico thalamic connections strength	Figure 3	0.32 nA
$g_{E,tc}$	thalamo-cortical connections to excitatory neurons	Figure 3	0.6 nA
$g_{I,tc}$	thalamo-cortical connections to inhibitory neurons	Figure 3	1.38 nA
	Simulation of multiple attractors		
μ_{EE}	long range E to E connection strength	Figure 7	0.01, 0.02, 0.03, 0.04, 0.05 nA
μ_{IE}	long range E to I connection strength	Figure 7	0.0167, 0.033, 0.05, 0.066, 0.083 nA
$g_{E,self}$	local self excitatory connections	Figure 7	0.4, 0.41, 0.2, 0.43, 0.44 nA

Table 2: Parameters for numerical simulations