nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-020-00753-w

Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks

Robert Kim [1,2,3] ✉ and Terrence J. Sejnowski [1,4,5] ✉

Cortical neurons process information on multiple timescales, and areas important for working memory (WM) contain neurons capable of integrating information over a long timescale. However, the underlying mechanisms for the emergence of neuronal timescales stable enough to support WM are unclear. By analyzing a spiking recurrent neural network model trained on a WM task and activity of single neurons in the primate prefrontal cortex, we show that the temporal properties of our model and the neural data are remarkably similar. Dissecting our recurrent neural network model revealed strong inhibitory-to-inhibitory connections underlying a disinhibitory microcircuit as a critical component for long neuronal timescales and WM maintenance. We also found that enhancing inhibitory-to-inhibitory connections led to more stable temporal dynamics and improved task performance. Finally, we show that a network with such microcircuitry can perform other tasks without disrupting its pre-existing timescale architecture, suggesting that strong inhibitory signaling underlies a flexible WM network.

Temporal receptive fields are organized hierarchically across the cortex[1,2]. Areas important for higher cognitive functions are capable of integrating and processing information in a robust manner and reside at the top of the hierarchy[1–3]. The prefrontal cortex (PFC) is a higher-order cortical region that supports a wide range of complex cognitive processes including WM—an ability to encode and maintain information over a short period of time[4,5]. However, the underlying circuit mechanisms that give rise to the stable temporal receptive fields associated strongly with WM are not known and challenging to probe experimentally. A better understanding of possible mechanisms could elucidate not only how areal specialization in the cortex emerges, but also how local cortical microcircuits carry out WM computations.

Previous experimental studies reported that baseline activities of single neurons in the primate PFC contain unique temporal receptive field structures. Using decay time constants of spike-count autocorrelation functions obtained from neurons at rest, these studies demonstrated that the primate PFC is composed mainly of neurons with large time constants or timescales[1,6–8]. In addition, neurons with longer timescales carried more information during the delay period of a WM task compared with short-timescale neurons[8]. Chaudhuri et al.[2] proposed a large-scale computational model where heterogeneous timescales were organized naturally in a hierarchical manner that closely matched the hierarchy observed in the primate neocortex. The framework utilized a gradient of recurrent excitation to establish varying degrees of temporal dynamics[2]. Although their findings suggest that recurrent excitation is correlated with area-specific timescales, it is still unclear if recurrent excitation indeed directly regulates neuronal timescales and WM computations.

Recent experimental studies paint a different picture, in which diverse inhibitory interneurons form intricate microcircuits in the PFC to execute memory formation and retrieval[9–13]. Both somatostatin (SST) and vasoactive intestinal peptide (VIP) interneurons have been shown to form a microcircuit that can disinhibit excitatory cells via inhibition of parvalbumin (PV) interneurons[14,15]. Furthermore, SST and VIP neurons at the center of such disinhibitory microcircuitry were causally implicated with impaired associative and working memory via optogenetic manipulations[9,10,12,13]. Consistent with these observations, the primate anterior cingulate cortex, which is at the top of the timescale hierarchy,[1] was found to contain more diverse and stronger inhibitory inputs compared with the lateral PFC[16]. A recent theoretical study also showed that inhibitory-to-inhibitory synapses, although far fewer in number compared to excitatory connections, are critical components for implementing robust maintenance of memory patterns[17].

To characterize how strong inhibitory signaling enables WM maintenance and leads to slow temporal dynamics, we constructed a spiking recurrent neural network (RNN) model to perform a WM task, and compared the emerging timescales with the timescales derived from the prefrontal cortex of rhesus monkeys trained to perform similar WM tasks. Here, we show that both the primate PFC and our RNN model utilize units with long timescales to sustain stimulus information. By analyzing and dissecting the RNN model, we illustrate that inhibitory-to-inhibitory synapses incorporated into a disinhibitory microcircuit tightly control both neuronal timescales and WM task performance. Finally, we show that the primate PFC exhibits signs that it is already equipped with strong inhibitory connectivity even before learning the WM task, implying that a gradient of recurrent inhibition could naturally result in functional specialization in the cortex. We confirm this with our model and show that the task performance of RNNs with short timescales can be enhanced via increased recurrent inhibitory signals. Overall, our work offers timely insight into the role of diverse inhibitory signaling in WM and provides a circuit mechanism that can explain previously observed experimental findings.

[1]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, USA. [2]Neurosciences Graduate Program, University of California San Diego, La Jolla, CA, USA. [3]Medical Scientist Training Program, University of California San Diego, La Jolla, CA, USA. [4]Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA. [5]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA.
✉e-mail: rkim@salk.edu; terry@salk.edu

Results

Spiking RNN model. To study how stable temporal dynamics associated with WM emerge, we trained a spiking RNN model to perform a WM task. The model used in the present study is composed of leaky integrate-and-fire (LIF) units recurrently connected to one another (Methods).

The WM task we used to train the spiking RNNs was a delayed match-to-sample (DMS) task (Fig. 1a; Methods). The task began with a 1 s long fixation period (that is, no external input) followed by two sequential input stimuli (each stimulus lasting for 0.25 s) separated by a delay period (0.75 s). The input signal was set to either −1 or +1 during the stimulus window. If the two sequential stimuli had the same sign (−1/−1 or +1/+1), the network was trained to produce an output signal approaching +1 after the offset of the second stimulus. If the stimuli had opposite signs (−1/+1 or +1/−1), the network produced an output signal approaching −1.

Using a method that we had developed previously, we configured the recurrent connections and synaptic decay time constants ($\boldsymbol{\tau}^{d}$) required for the spiking model to perform the task[18]. Briefly, we trained continuous-variable rate RNNs to perform the task using a gradient-descent algorithm, and the trained networks were then mapped to LIF networks. In total, we 'trained' 40 LIF RNNs of 200 units (80% excitatory and 20% inhibitory units) to perform the task with high accuracy (accuracy > 95%; Methods).

Experimental data. To ensure that our spiking model is a biologically valid one for probing neuronal timescales observed in the cortex, we also analyzed a publicly available dataset containing extracellular spike trains recorded from the dorsolateral prefrontal cortex (dlPFC) of four rhesus monkeys[19–21]. The monkeys were trained on spatial and feature DMS tasks. A trial for both task types began with a fixation period (1 s in duration) during which the monkeys were required to maintain their gaze at a fixation target. For a spatial DMS trial, the monkeys were trained to report if two sequential stimuli separated by a delay period (1.5 s) matched in spatial location (Fig. 1b). More details regarding the dataset and the tasks can be found in the Methods and in Qi et al.[19] and Meyer et al.[20].

Long neuronal timescales in both RNN model and experimental data. Previous studies demonstrated that higher cortical areas consist of neurons with long, heterogeneous timescales using the spike-count autocorrelation decay time constant as a measure of a neuron's timescale[1,7,8]. Here, we sought to confirm that our spiking RNNs trained on the DMS task and the neural data were also composed of units with predominantly long timescales. For each unit from our RNNs and the dlPFC, we computed the autocorrelation decay time constant (τ) of its spike-count during the 1 s fixation period (sMethods)[1]. The baseline activities (average firing rates during the fixation period) of the units that satisfied the inclusion criteria were comparable between the dlPFC data and our model (Fig. 2a; Methods). Both data contained units with slow temporal dynamics (that is, long τ values) and short τ units whose autocorrelation function decayed fast (Fig. 2b). Furthermore, the distribution of the timescales was heavily left-skewed for both data (Fig. 2c,d, left and middle panels) underscoring overall slow temporal properties associated with WM. On the other hand, the RNNs before training (that is, sparse, random Gaussian connectivity weights; Methods) were dominated by units with extremely short timescales (Fig. 2c,d, right panels), suggesting that the long τ units observed in the trained RNNs were the result of the supervised training.

Long neuronal timescales are essential for stable coding of stimuli. Next, we investigated whether units with longer τ values were involved with more stable coding compared with short τ units using cross-temporal decoding analysis[8,22,23]. We performed the cross-temporal decoding analysis on short and long neuronal timescale subgroups from the neural data and the RNN model. A unit was assigned to the short τ group if its timescale was smaller than the lower quartile value. The upper quartile was used to identify units with large τ values. There were 64 units in each subgroup for the experimental data. For the RNN model, there were 230 units in each subgroup.

The cross-temporal discriminability analysis revealed that stronger cue-specific differences (that is, higher discriminability) across the delay period were present in the long τ subgroup compared with the short τ subgroup for both data (Fig. 3a). The significant decodability during the delay period for the dlPFC dataset stemmed mainly from the spatial task dataset (Supplementary Fig. 1). The within-delay discriminability (that is, taking the diagonal values of the cross-temporal decoding matrices) for the long τ group was significantly higher than the discriminability observed from the short τ group throughout the delay period for the RNN model (Fig. 3b). For the dlPFC dataset, we observed a significant correlation between the τ values and the average fixation firing rate (Supplementary Fig. 2), but stratifying the dataset to remove the relationship and repeating the cross-temporal discriminability analysis led to qualitatively similar results (Methods; Supplementary Fig. 3). Although significant within-delay discriminability was not observed for the dlPFC data (Fig. 3b, top), Wasmuht et al.[8] reported significant within-delay decodability during the delay period in the primate lateral prefrontal cortex, consistent with our model findings.

Strong inhibitory connections give rise to task-specific temporal receptive fields. Neuronal timescales extracted from cortical areas have been shown to closely track the anatomical and functional organization of the primate cortex[1,2]. To investigate if such functional specialization also emerges in our spiking model, we trained another group of spiking RNNs ($n = 40$ RNNs) on a simpler task that did not require WM. The non-WM task, which we refer to as a two-alternative forced choice (AFC) task, required the RNNs to respond immediately after the cue stimulus: output approaching −1 for the '−1' cue and +1 for the '+1' cue (Fig. 4a; Methods). Apart from the task paradigm, all the other model parameters were identical to the parameters used for the DMS RNNs.

Because the AFC task paradigm did not require the RNNs to store information related to the cue stimulus, we expected that these networks would exhibit faster timescales compared with the DMS RNNs. Consistent with this hypothesis, the AFC RNNs did not contain as many long τ units as the DMS RNNs (Fig. 4b), and the timescales averaged by network were also significantly faster for the AFC RNNs (Fig. 4c). To demonstrate that the timescale hierarchy we observed in Fig. 4c is not largely driven by the synaptic decay time constants ($\boldsymbol{\tau}^{d}$) we optimized, we trained additional RNNs without optimizing $\boldsymbol{\tau}^{d}$ (fixed to a constant value) for each task model. Fixing $\boldsymbol{\tau}^{d}$ did not disrupt the timescale hierarchy and resulted in moderate, yet significant, changes in neuronal timescales (Supplementary Fig. 4).

To gain insight into the circuit mechanisms underlying the difference in the timescale distributions of the AFC and DMS RNN models, we compared the recurrent connectivity patterns between these two models. Interestingly, mean excitatory and inhibitory synaptic strength was significantly greater for the DMS RNNs (Fig. 4d). To identify which connections led to the long timescales observed in the DMS model, we randomly rewired all the connections belonging to each of the four synaptic types (I → I, I → E, E → I, E → E) and computed the timescales again (Methods). Of the four conditions, only rewiring I → I synapses resulted in significantly shorter timescales than the timescales from the intact DMS model (Fig. 4e), and the distribution of the timescales pooled from all 40 RNNs with I → I connections shuffled resembled the distribution obtained from the AFC model (Supplementary Fig. 5). In addition, the amount of cue-specific information maintained during the delay period

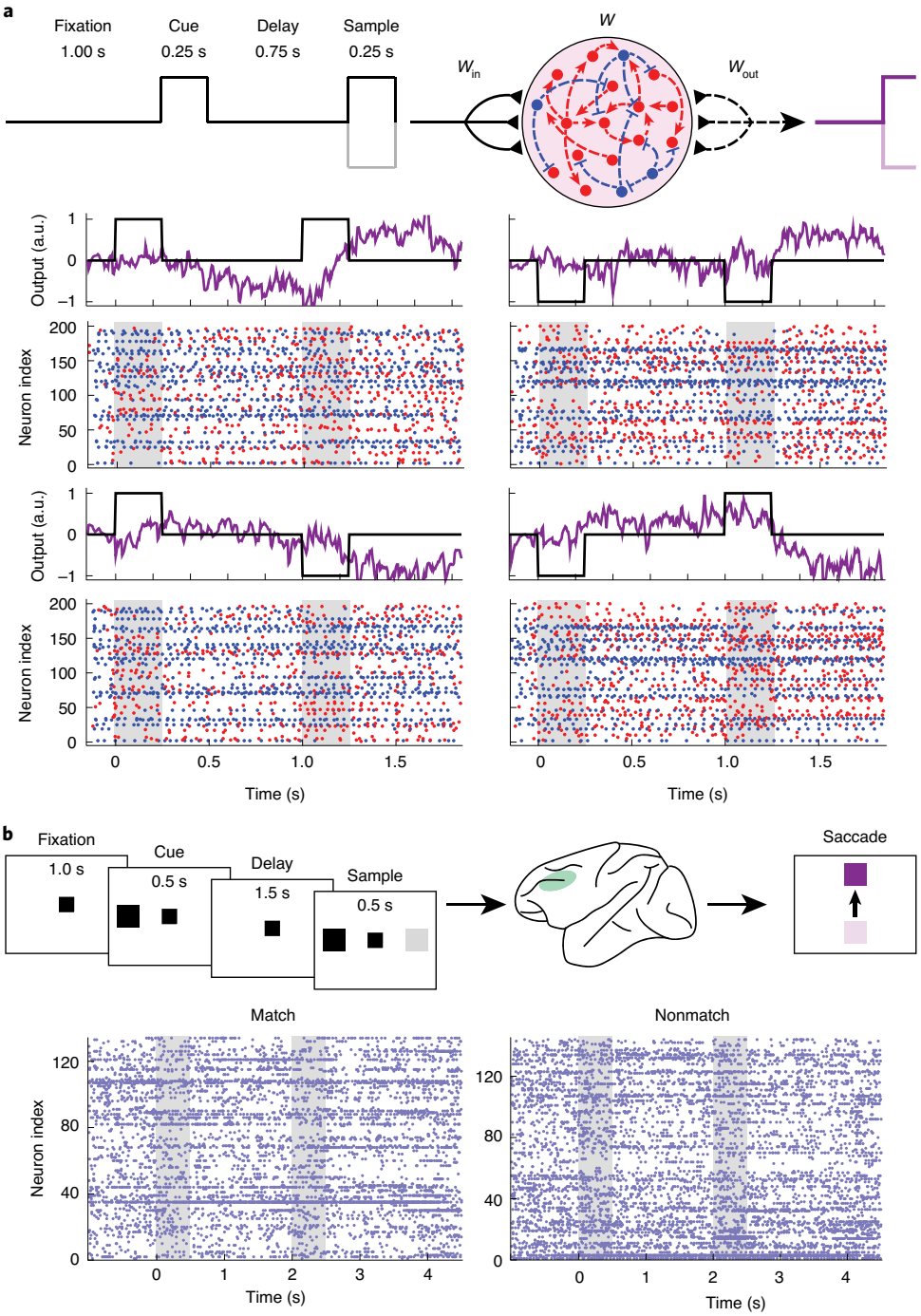

Fig. 1 | RNN model and experimental data. a, Spiking RNN model contained excitatory (red circles) and inhibitory (blue circles) units recurrently connected to one another. The model was trained to perform a DMS task. Each RNN contained 200 units (80% excitatory and 20% inhibitory), and 40 RNNs were trained to perform the DMS task. The dashed lines (recurrent connections and readout weights) were optimized via a supervised learning method. Example output signals along with the corresponding spike raster plots from a trained RNN are shown. Gray shading, stimulus window. **b**, Spatial DMS task paradigm used by Constantinidis et al.[21] to train four rhesus monkeys. Extracellular recordings from the dorsolateral prefrontal cortex (green area) were analyzed. Example spiking raster plots from randomly chosen match (left; 134 neurons) and nonmatch (right; 144 neurons) trials shown. Gray shading, stimulus window.

(as measured by the within-delay decoding time courses) was lowest for the I → I rewired condition (Fig. 4f), suggesting that shuffling I → I synapses was detrimental to memory maintenance.

Inhibitory-to-inhibitory connections regulate both neuronal timescales and task performance. We next investigated if I → I synapses could be manipulated to provide more stable temporal receptive fields and to improve WM maintenance.

Recent studies revealed that optogenetically stimulating SST or VIP interneurons that specifically inhibit PV interneurons could improve memory retrieval[10–12]. Based on these experimental observations, we expected that strengthening I → I synapses would

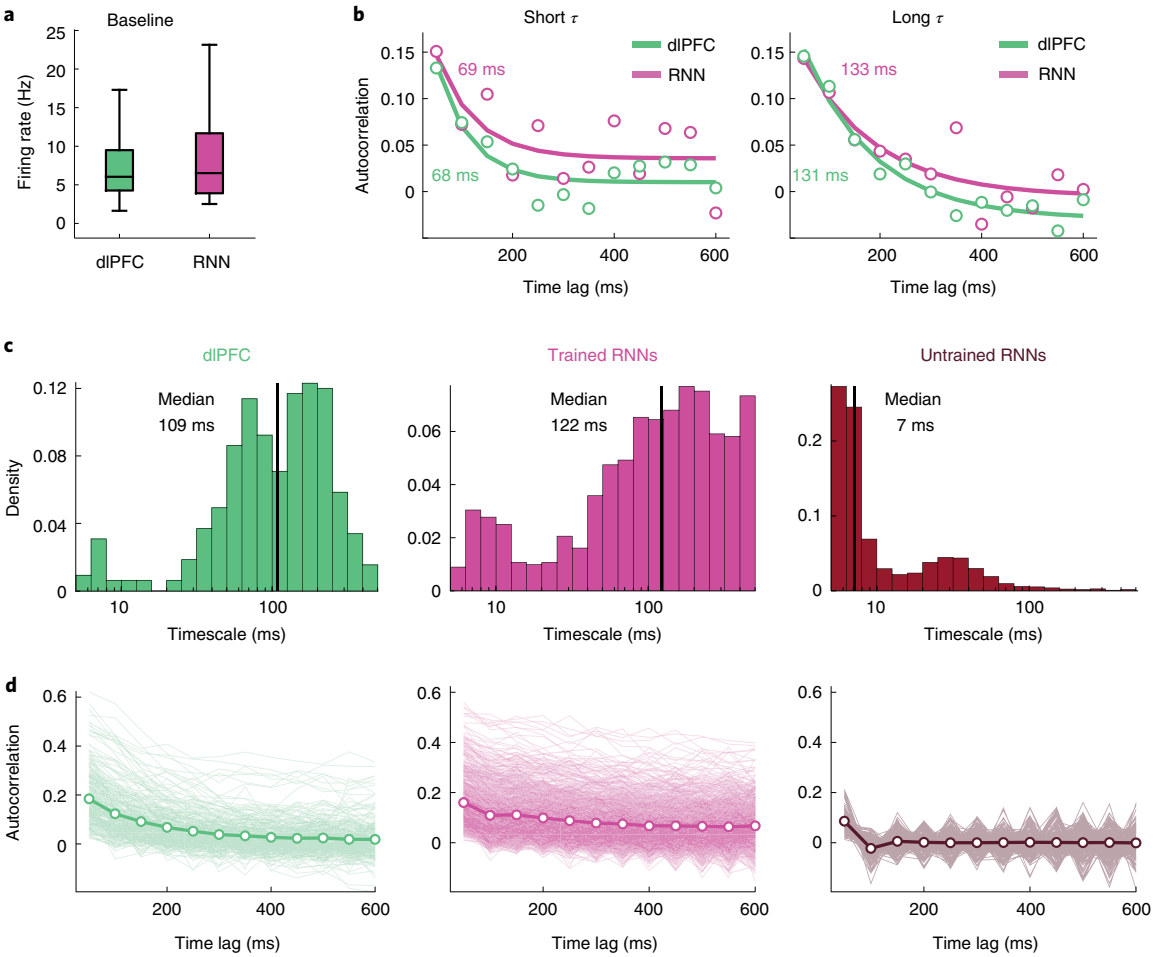

Fig. 2 | RNN model trained on the DMS task and the dlPFC data contain units with long timescales. a, Distribution of the firing rates during the fixation period was not significantly different between the experimental data ($n = 325$ neurons) and the RNN model ($n = 931$ units; $P < 0.70$, two-sided Wilcoxon rank-sum test). **b**, Autocorrelation decay curves from example units with short (left) and long (right) timescale values. **c**, Histograms of the distribution of the timescales from the experimental data ($n = 325$; green), trained RNNs ($n = 931$; magenta) and untrained RNNs ($n = 3{,}963$; dark brown). Solid vertical lines represent median $\log(\tau)$. **d**, Autocorrelation decay curves from single units (light) and the population average autocorrelation (bold) for the dlPFC data, trained RNNs and random RNNs. For the random RNNs, only 20% of the total single unit traces are shown. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers are not plotted.

increase neuronal timescales and task performance of the DMS RNNs. To test this hypothesis, we identified a group of RNNs with poor DMS task performance (26 RNNs; mean accuracy ± s.e.m., 71.77 ± 1.43%). This group of RNNs allowed us to observe the effects of synaptic manipulations on memory maintenance more easily than in the group of RNNs used in the previous section, which performed the task with very high accuracy. Next, we modeled the effects of optogenetic manipulation of VIP, SST and PV neurons by either decreasing or increasing I → I synaptic strength ($W_{I \to I}$) in each network by 30% (Methods). Decreasing the connection strength led to significantly shorter timescales compared with the RNNs without any modification (Fig. 5a, left). Strengthening $W_{I \to I}$ resulted in a moderate but significant increase in neuronal timescale (Fig. 5a, left). The average within-delay decodability measure during the delay period (Methods) of the RNNs followed the same pattern: decreasing $W_{I \to I}$ severely impaired WM maintenance, whereas increasing $W_{I \to I}$ significantly improved task performance (Fig. 5a, right). For I → E connections, enhancing only $W_{I \to E}$ resulted in significant changes in both timescale and within-delay decodability (Fig. 5b). Manipulating E → I synapses did not affect the within-delay discriminability, but decreasing $W_{E \to I}$ significantly shortened the timescales (Fig. 5c). Altering the excitatory-to-excitatory connections did not produce any significant changes (Fig. 5d). Consistent with these observations, RNNs with only I → I connections trainable were able to learn the task (26 out of 40 rate RNNs trained successfully), while RNNs with plastic I → E, E → I, or E → E connections could not be trained to perform the DMS task (data not shown). Overall, these findings suggest that I → I synapses tightly mediate both temporal stability and WM maintenance. The findings also indicate that the main downstream effect of I → I connections is to disinhibit excitatory units.

Unique inhibitory-to-inhibitory circuitry for WM maintenance. Here, we dissect the DMS RNN model to elucidate how specific and strong I → I connections lead to stable memory retention.

Focusing on inhibitory units only, we first characterized the cue stimulus selectivity from each inhibitory unit in an example DMS network (Methods). Analyzing the selectivity index values revealed two distinct subgroups of inhibitory units in the network: one group of units favoring the positive cue stimulus and the other group selective for the negative stimulus (Fig. 6a, top). The input weights (W_{in}) that project to these units closely followed the selectivity pattern (Fig. 6a, bottom). Similar selectivity patterns were observed in the excitatory population (Supplementary Fig. 6).

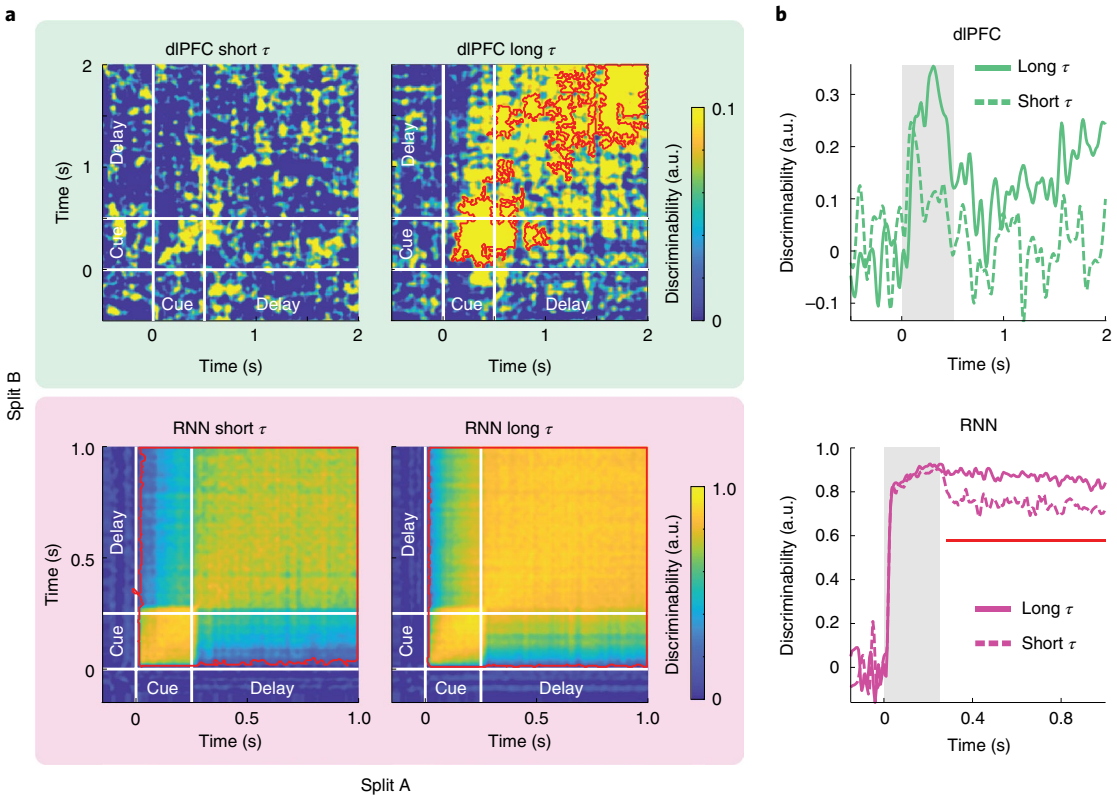

Fig. 3 | Long τ units maintain cue stimulus information during the delay period robustly. a, Cross-temporal discriminability matrices for the dlPFC data (top row) and the RNN model (bottom row). Red contours indicate significant decodability ($P < 0.05$ by one-sided cluster-based permutation test; Methods). **b**, Within-delay discriminability time courses from the short (dashed) and long (solid) τ groups for the dlPFC data and the RNN model. Gray shading, cue stimulus window. Red lines indicate significant differences in decoding between the short and long τ groups ($P < 0.05$ by one-sided cluster-based permutation test; Methods). a.u., arbitrary units.

Given these two subgroups with distinct selectivity patterns, we next hypothesized that mutual inhibition between these two groups (across-group inhibition) was stronger than within-group inhibition. Indeed, inhibition between the oppositely tuned inhibitory populations was significantly greater (both in synaptic strength and number of connections) than inhibition within each subgroup across all RNNs (Fig. 6b). To confirm that the behavioral improvement we observed with I→I enhancement in Fig. 5a was due largely to the strengthened across-group inhibition, we increased across-group and within-group I→I connections separately (Methods). The maintenance of the cue stimulus improved following enhancement of the across-group inhibition, whereas increasing the within-group inhibition impaired maintenance (Fig. 6c). In addition, across-group I→I enhancement resulted in a significant increase in neuronal timescale (Fig. 6d).

In summary, these findings imply that robust inhibition of oppositely tuned inhibitory subpopulations is critical for memory maintenance in our RNN model. For example, a positive cue stimulus activates the inhibitory and excitatory subgroups selective for that stimulus and deactivates the negative stimulus subgroups (Fig. 6e). During the delay period, the inhibition strength between these two inhibitory subgroups dictates the stability of the cue-specific activity patterns generated during the stimulus window (Fig. 6f). The positive feedback provided by the similarly tuned excitatory neurons sustains the stimulus-specific activity of the inhibitory subgroups (Supplementary Fig. 7). The circuit diagram shown in Fig. 6f is further validated by repeating the analyses performed in Fig. 5 to cue-selective inhibitory and excitatory subgroups (Supplementary Fig. 8).

Circuit mechanism for WM generates units with long neuronal timescales. The circuit mechanism (Fig. 6e,f) explains why enhancing I→I connections results in improved WM performance, but it is still not clear how this same mechanism also produces units with long timescales.

Here, we first demonstrate that a high trial-to-trial spike-count variability during the fixation period could give rise to slow decay of the spike-count autocorrelation function. If a neuron exhibits highly variable activity patterns across trials such that it is highly active (that is, persistent firing) in some trials and relatively silent in other trials, the Pearson correlation between any two time bins within the fixation window could be large (Fig. 7a). On the other hand, firing activities with a low trial-to-trial variability could result in a weak correlation between two time bins. To directly test this positive relationship between trial-to-trial variability and neuronal timescales, we computed spike-count Fano factors (spike-count variance divided by spike-count mean across trials; Methods) for the short and long τ subgroups in both neural and model data. The Fano factor values for the short-timescale subgroup were significantly smaller than the values obtained from the long τ group for both data (Fig. 7b). There was also a significant positive correlation between the spike-count Fano factors and neuronal timescales across all the units in both data (Spearman rank correlation, $r = 0.25$, $P < 0.0001$ for dlPFC; $r = 0.28$, $P < 0.0001$ for RNN; Supplementary Fig. 9).

Manipulating each of the four synaptic types (decreasing or increasing synaptic strength by 30%) in our DMS RNN model revealed that I→I connections strongly modulated the spike-count Fano factors (Fig. 7c). Enhancing I→I synaptic strength led to units with more variable spiking patterns across trials, whereas reducing

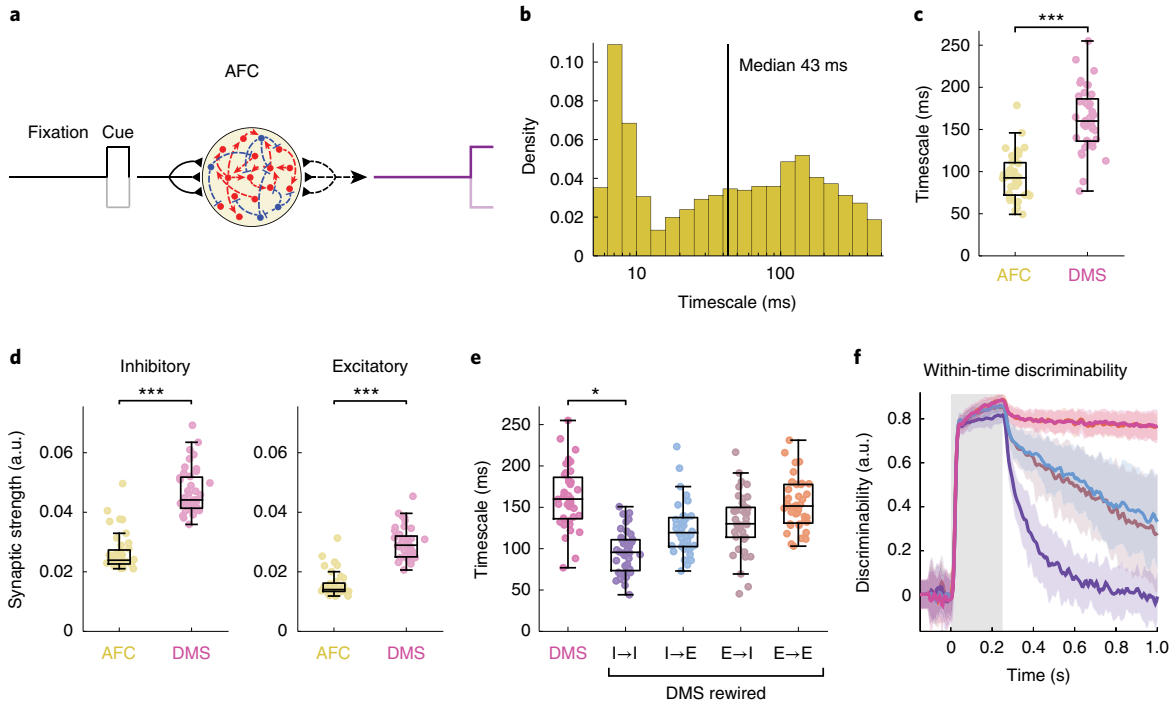

Fig. 4 | Inhibitory synaptic weights lead to task-specific timescales. a, Task paradigm for the AFC task. **b**, Distribution of the neuronal timescales extracted from 40 RNNs trained on the AFC task. Solid vertical line represents median log(τ). **c**, Average timescale values from the AFC ($n=40$ RNNs) and DMS ($n=40$ RNNs) models. Each circle represents the average value from one RNN. $^{***}P=2.92\times10^{-14}$ by two-sided Wilcoxon rank-sum test. **d**, Average recurrent inhibitory (left) and excitatory (right) synaptic strengths from the AFC ($n=40$ RNNs) and DMS ($n=40$ RNNs) models. $^{***}P=1.51\times10^{-18}$ (left) and $P=1.01\times10^{-17}$ (right) by two-sided Wilcoxon rank-sum test. **e**, Average timescales from the DMS RNNs ($n=40$), with each synaptic type rewired randomly (Friedman test, $F=74.19$, $P=2.95\times10^{-15}$). $^{*}P=0.0089$ by Dunn's multiple comparison test. **f**, Within-delay discriminability time courses averaged across all the DMS RNNs for each rewiring condition. Color scheme as in **e**. The bold line indicates the mean timecourse averaged across 40 RNNs (and all units). Colored shading, ±s.d. Gray shading, cue stimulus window. Within-delay discriminability timecourse for the E → E condition shown behind the timecourse for the intact model. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers not plotted.

the strength resulted in smaller Fano factors (see example shown in Supplementary Fig. 10).

In our RNN model, strong I → I synapses give rise to both excitatory and inhibitory units behaving in a highly variable manner during the fixation period (Fig. 7d). For instance, an inhibitory unit selective for the positive stimulus could be partially activated in some trials by chance (that is, via random noise during the fixation period), and this, in turn, could silence a portion of the negative stimulus inhibitory population (light blue circle in Fig. 7d). This leads to variable firing activities across trials in inhibitory units. Furthermore, the dynamic activity of the inhibitory population could be transferred to the excitatory population via disinhibition. Therefore, I → I connections play a central role in conferring the network with highly dynamic baseline firing patterns, which then translate to high τ values.

Strong I → I is an intrinsic property of prefrontal cortex. Cognitive flexibility is one of the hallmarks of the prefrontal cortex[24,25]. If higher-order areas are indeed wired with specific and robust I → I synapses that give rise to stable temporal receptive fields, then what would happen to these connections during learning? Would learning a new task disrupt the existing I → I connectivity structure, thereby abolishing the previously established timescale distribution? To answer these questions, we analyzed neuronal timescales from the same monkeys before they learned the DMS task. For the pretraining condition, the monkeys were trained on a passive task (Fig. 8a): they were trained to maintain their gaze at a central fixation point throughout the trial regardless of the stimuli presented around the fixation point[26].

Surprisingly, the timescales from the spike-train data from the dlPFC of the same four monkeys that learned the passive task were similar to the timescales obtained after the monkeys learned the DMS task (Fig. 8b). In addition, the cue-specific information maintenance during the delay period by long τ units was largely abolished, and the within-delay decoding was similar between long τ and short τ neurons (Fig. 8c). These findings suggest that the primate dlPFC was already equipped with stable temporal receptive fields and that learning the DMS task resulted in long τ neurons carrying more information during the delay period while preserving the network temporal dynamic architecture.

Based on these findings, we reasoned that prefrontal cortical areas and other higher cognitive areas are endowed with strong I → I connections, the connectivity patterns of which do not undergo significant plastic changes during learning. Instead, learning-related changes occur to the connections stemming from upstream networks that project to these areas. To test this, we asked if we could optimize only the upstream connections (that is, input weights; W_{in}) of the good performance DMS RNNs ($n=40$) to perform a passive version of the DMS task (Supplementary Fig. 11; Methods). By freezing the recurrent connections (W), we ensured that the previously observed distribution of the timescales (Fig. 2b) was preserved. The readout weights (W_{out}) were frozen to ensure that they were not simply set to 0 to do the passive task. Repeating the cross-temporal discriminability analysis on the retrained RNNs showed that the cue

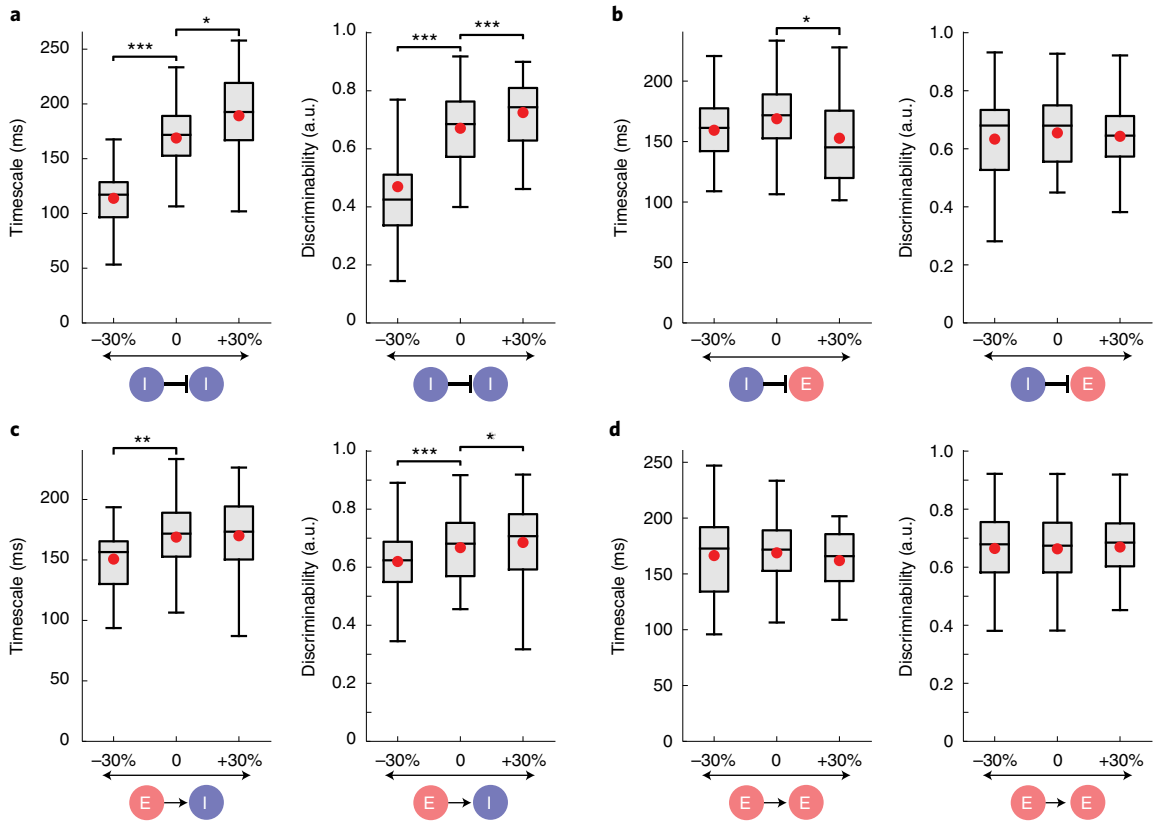

Fig. 5 | I → I connectivity strength strongly mediates both neuronal timescales and task performance. a–d, Timescales and within-delay discriminability changes observed in the poor-performance DMS model ($n = 26$) when I → I (**a**), I → E (**b**), E → I (**c**) or E → E (**d**) connection strength was either decreased or increased by 30%. For **a:** left, $^{*}P = 0.0236$ and $^{***}P = 7.54 \times 10^{-6}$**;** right, $^{***}P = 2.27 \times 10^{-5}$ and $^{***}P = 3.76 \times 10^{-5}$ (from left to right). For **b**, $^{*}P = 0.0236$. For **c:** left, $^{**}P = 0.0079$**;** right, $^{***}P = 2.48 \times 10^{-4}$ and $^{*}P = 0.01$. Two-sided Wilcoxon signed-rank test was used. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers not plotted.

stimulus information during the delay period was not maintained as robustly by long τ units (Supplementary Fig. 11). Retuning the recurrent connections instead of the input weights for the passive task disrupted the existing timescale structure and resulted in significantly faster timescales (Supplementary Fig. 12).

The above results from the experimental data and our model suggest strongly that higher cortical areas might have intrinsically diverse and robust inhibitory signaling. This innate property, in turn, would give rise to long neuronal timescales, and the incoming connections to these areas could undergo plastic changes to support various higher cognitive functions that require integration of information on a slower timescale. Along this line of thought, we wanted to probe if the AFC RNNs, which do not have strong inhibitory-to-inhibitory signaling, are flexible enough to perform other tasks. Previous modeling studies have demonstrated the importance of disinhibitory circuitry for gating incoming stimuli and decision-making[27,28]. We first investigated if disinhibitory circuitry without strong I → I synapses was sufficient for flexible decision-making by retraining the AFC RNNs (with the recurrent architecture, W, frozen) to perform a task that requires flexible input gating. The new task is modeled after the design used by the previous studies[18,29] and required selective gating of incoming stimuli (Fig. 8d; Methods). Both AFC (39 out of 40 RNNs) and DMS models (40 out of 40 RNNs) were successfully retrained to perform the new task (Fig. 8e). Next, we retrained both models to perform a different WM task, DNMS task (Methods). As shown in Fig. 8f, none of the AFC RNNs could be trained to perform the DNMS task. When we repeated the retraining procedure with the I → I recurrent connections strengthened (Methods) and the performance of the AFC RNNs improved significantly (Fig. 8f right). On the other hand, the input weights of the DMS RNNs could be tuned to perform the DNMS task (Fig. 8g). Taken together, these results suggest that strong I → I connections might not be necessary for selective attention and integrating incoming stimuli, but these connections become important for carrying out WM computations.

Discussion

In this study, we provide a computational model that gives rise to task-specific spontaneous temporal dynamics, reminiscent of the hierarchy of neuronal timescales observed across primate cortical areas[1]. When trained on a WM task, our RNN model was composed of units with long timescales, the distribution of which was surprisingly similar to that obtained from the primate dlPFC. In addition, the long-timescale units encoded and maintained WM-related information more robustly than the short-timescale units during the delay period. By analyzing the connectivity structure of the model, we showed that a unique circuit motif that incorporates strong I → I synapses is an integral component of WM computations and slow baseline temporal properties. Interestingly, I → I synaptic weights could be manipulated to control both memory maintenance and neuronal timescales tightly. Our work also provides mechanistic insight into how I → I connectivity supports the memory storage and dynamic baseline activity patterns crucial for long neuronal timescales. Lastly, we propose that the microcircuitry we identified is intrinsic to higher-order cortical areas, enabling them to perform cognitive tasks that require steady integration of information.

Relating specific baseline spiking activities to the underlying circuit mechanisms has been challenging, due partly to the lack

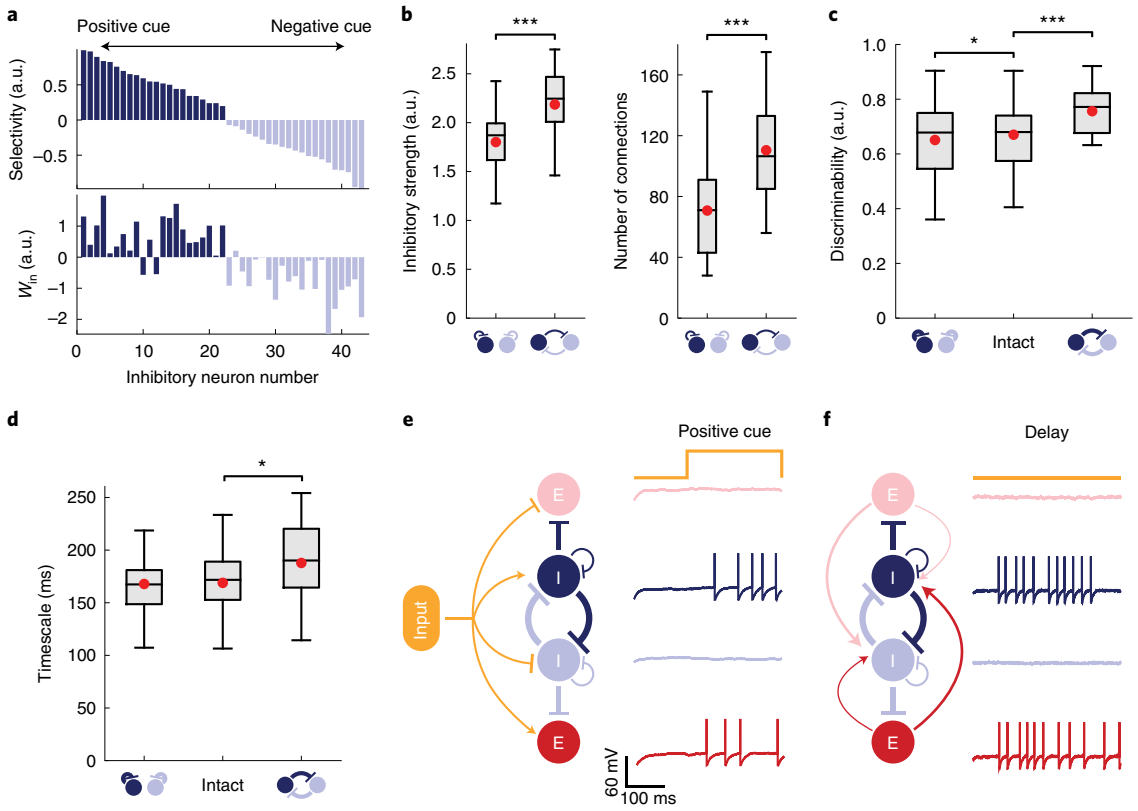

Fig. 6 | Two oppositely tuned inhibitory subgroups mutually inhibit each other for WM maintenance. a, Cue preference selectivity (top) and input weights (W_{in}; bottom) from inhibitory units of an example DMS RNN. The selectivity index values are sorted in descending order. **b**, Average inhibitory strengths (left) and number of inhibitory connections (right) within and across two oppositely tuned inhibitory subgroups from the poor-performance RNNs ($n=26$). $^{***}P=4.17\times10^{-7}$ (left) and $^{***}P=2.98\times10^{-8}$ (right) by two-sided Wilcoxon signed-rank test. **c**, Average within-delay discriminability of the poor-performance DMS model ($n=26$) when the within-group or across-group inhibition was increased by 30%. $^{*}P=0.049$. and $^{***}P=2.98\times10^{-7}$ by two-sided Wilcoxon signed-rank test. **d**, Average neuronal timescales of the DMS RNNs ($n=26$) when the within-group or across-group inhibition was increased by 30%. $^{*}P=0.01$ by two-sided Wilcoxon signed-rank test. **e**,**f**, Schematic illustration of the circuit mechanism employed by the DMS RNN model during the cue stimulus window (**e**) and delay period (**f**). The positive cue stimulus was used as an example, and membrane voltage tracings from example units are shown. Dark blue and dark red units indicate units that prefer the positive cue stimulus, while the light blue and light red units favor the negative cue. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers not plotted.

of computational models capable of both performing cognitive tasks and capturing temporal dynamics derived from experiments. Bouchacourt et al.[30] employed Poisson spiking neurons wired randomly to present a flexible WM model, whereas Mongillo et al.[17] used LIF RNNs constrained by experimental measurements to underscore the importance of inhibitory connectivity in WM. These studies provide biologically plausible models that can explain several experimental and behavioral aspects of WM, but it is unclear whether units with stable baseline temporal dynamics are recruited for performing WM maintenance in these models. It is also possible to study neuronal timescales using continuous rate (that is nonspiking) RNNs, which have been used widely to uncover neural mechanisms behind cognitive processes[29,31–34]. Although spontaneous firing rate estimates could be used in place of spike counts to compute the autocorrelation decay time constants, our spiking RNN model allowed us to (1) use the same experimental procedures previously used to estimate neuronal timescales, (2) easily interpret and compare our model results with experimental findings, and (3) uncover spiking statistics (spike-count Fano factors) associated with long neuronal timescales.

Our work revealed that strong I → I connections are critical for long neuronal timescales, and we investigated the functional implication of such connections in WM-related behavior. Despite the fact that excitatory pyramidal cells make up the majority of neurons in cortical areas, inhibitory interneurons have been shown to exert greater influence at the local network level[35,36]. Furthermore, different subtypes of interneurons play functionally distinct roles in cortical computations[9,14]. In agreement with these observations, recent studies uncovered the importance of disinhibitory gating imposed by VIP interneurons[10,13,37,38]. Surprisingly, optogenetically activating VIP neurons in the PFC of mice trained to perform a WM task significantly enhanced their task performance, highlighting that disinhibitory signaling is vital for memory formation and recall[10]. Similar to VIP neurons, SST interneurons have also been shown to disinhibit excitatory cells for fear memory[12,13]. Intriguingly, the connectivity structures of the RNNs we trained on a WM task using supervised learning also centered around disinhibitory circuitry with strong I → I synapses (Fig. 6). The strength of the I → I connections was coupled tightly to the task performance of the RNNs. Thus, our work suggests that microcircuitry with robust I → I synapses could be a common substrate in higher-order cortical areas that require short-term memory maintenance.

Most notably, our results shed light on exactly how robust I → I connections maintain stable memory storage and long neuronal timescales. By dissecting our WM RNN model, we found that strong mutual inhibition between two oppositely tuned inhibitory

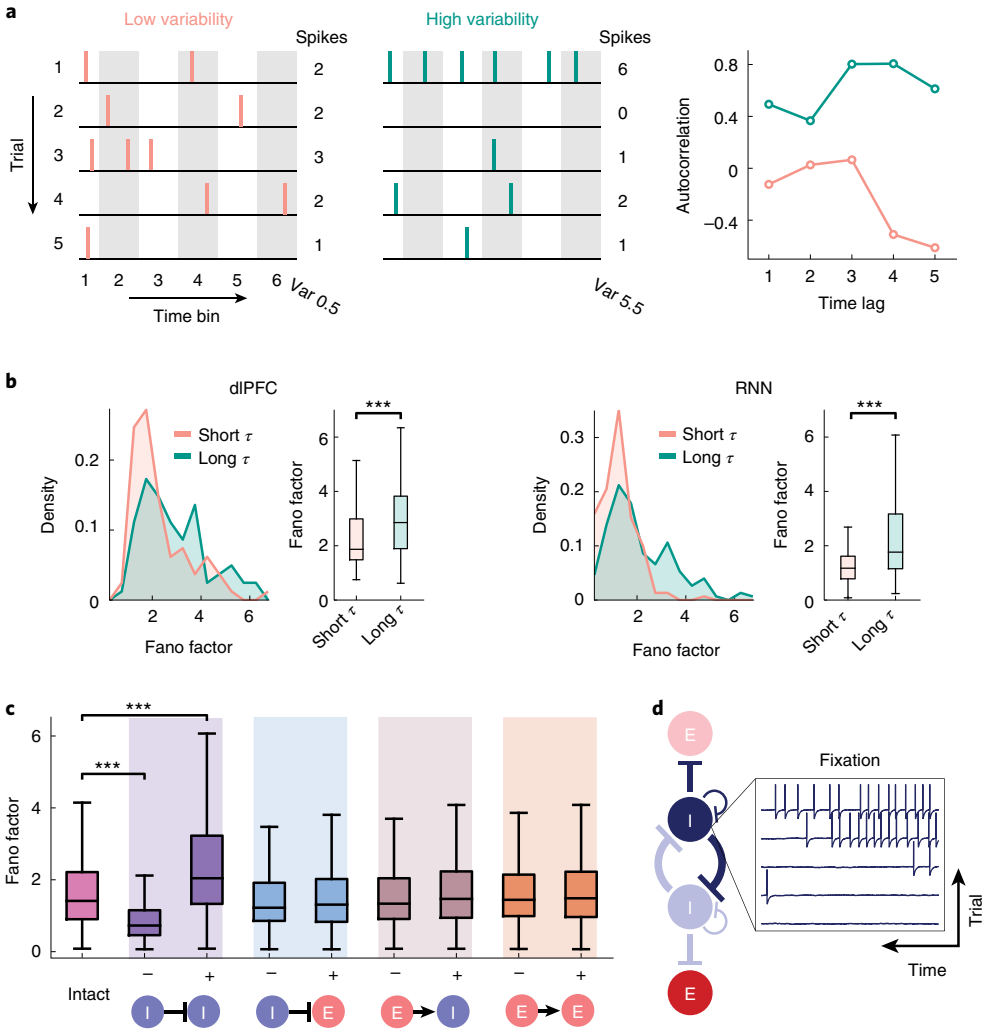

Fig. 7 | High trial-to-trial spike-count variability during fixation corresponds to long neuronal timescale. a, Schematic illustrating how high spike-count variability (Var) across multiple trials can result in slow decay of the autocorrelation function. **b**, Comparison of the spike-count Fano factors from the short and long τ groups in the neural data ($n=81$ neurons in each group; left) and the DMS RNN model ($n=151$ units in each group; right). **c**, Average Fano factors from the DMS model with each of the synaptic type either decreased (−) or increased (+) by 30% (Kruskal–Wallis test, $H=665.2$, $P<0.0001$). Intact, $n=741$ units; I → I decreased, $n=562$ units; I → I increased, $n=754$ units; I → E decreased, $n=1149$ units; I → E increased, $n=571$ units; E → I decreased, $n=786$ units; E → I increased, $n=722$ units; E → E decreased, $n=733$ units; E → E increased, $n=719$ units. **d**, Spiking activity of an example inhibitory unit during the fixation period across five trials. The trials were sorted by the number of spikes. Units that were strongly modulated by the disinhibitory circuit mechanism showed highly dynamic baseline firing patterns across trials. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers not plotted. ***$P<0.0001$ by two-sided Wilcoxon rank-sum test (**b**) or Dunn's multiple comparisons test (**c**).

subgroups was necessary for maintaining stimulus-specific information during the delay period (Fig. 6). This emerging circuit mechanism of mutual inhibition is similar to previous decision-making models where feedback inhibition was utilized to produce winner-take-all competition[39]. In the current study, we demonstrated that a winner-take-all motif without strong I → I synapses is sufficient for selective gating of information and decision-making (Fig. 8e) and that strong mutual inhibition can confer the disinhibitory circuit with WM capability (Fig. 8f,g). We also illustrated that our model units, which were strongly modulated by I → I synapses, displayed highly dynamic baseline activities, leading to both large trial-to-trial Fano factors and long neuronal timescales (Fig. 7). Our findings suggest that baseline trial-to-trial spike-count variability and neuronal timescales are reliable indicators of the underlying circuit mechanisms: neurons with asynchronously occurring synchronous firing patterns (that is, high variability) could make up WM-related microcircuits. Furthermore, we propose that these signatures are area-specific and do not undergo significant changes during learning.

One of the testable hypotheses that our modeling work provides is that strength of I → I connections defines the cortical hierarchy: higher cortical areas contain stronger and more diverse inhibitory signaling than lower cortical regions. This hypothesis is supported strongly by a large-scale experimental study quantifying the density of SST and PV interneurons across cortical and subcortical regions in mice[40]. The study found the density of SST interneurons to closely parallel the hierarchical organization of the cortex: PV interneurons were predominant in sensory-motor areas, while SST interneurons were prevalent in association areas. On a smaller scale, Medalla et al. discovered that inhibitory signaling strength and diversity were higher in the anterior cingulate cortex than the prefrontal cortex[16,41]. This observation is consistent with the neuronal timescale hierarchy

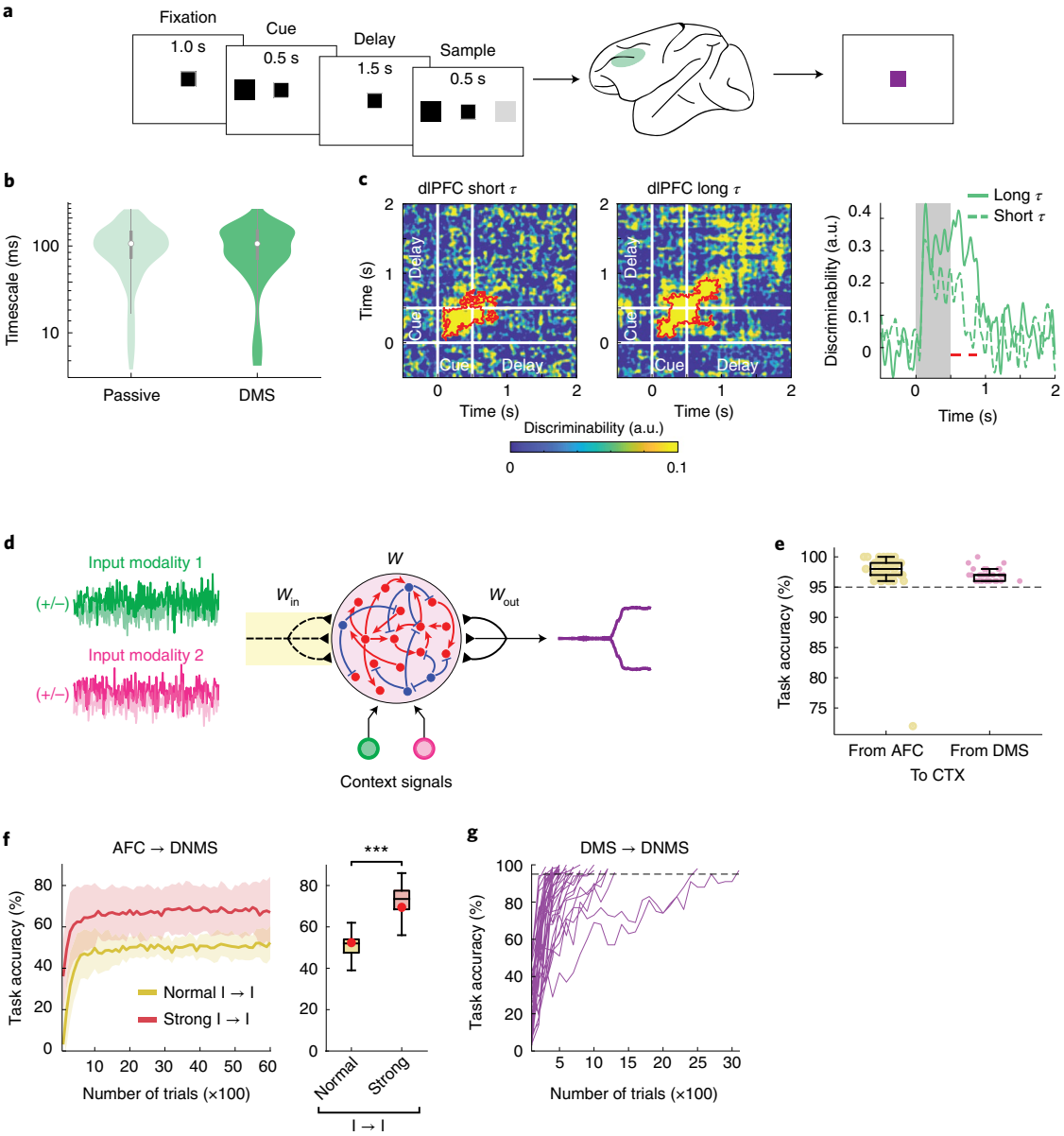

Fig. 8 | Strong I → I connections might be intrinsic to prefrontal cortex. a, Passive task paradigm used by Constantinidis et al.[21] to train the same four monkeys before they learned the DMS tasks (Fig. 1b). **b**, Distribution of the neuronal timescales from the monkeys before (that is, passive; $n = 434$ neurons) and after they learned the DMS tasks ($n = 325$ neurons). **c**, Cross-temporal decoding matrices and within-delay decoding time courses from the short and long τ subgroups. **d**, Task paradigm for the CTX. **e**, Average task performance of the AFC and DMS rate RNNs ($n = 40$ in each group) after being retrained to perform the CTX task. Dashed line, task performance threshold (95%). **f**, Task performance during retraining of the AFC rate RNNs ($n = 40$) to perform the DNMS task (left) and average performance at the end of training (right). The task performance increased significantly when I → I connections were strengthened (orange; Methods). Shaded area, ±s.d. **g**, Task performance during retraining of the DMS rate RNNs to perform the DNMS task. Task performance during retraining for individual networks shown. Boxplot: central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers not plotted. Red contours indicate significant discriminability (cluster-based permutation test, $P < 0.05$; Methods). Red lines indicate significant differences in decoding between the short and long τ groups (cluster-based permutation test, $P < 0.05$; Methods). ***$P < 0.0001$ by Wilcoxon signed-rank test.

that Murray et al. have reported previously[1]. Thus, our findings and the proposed hypothesis that I → I connections become stronger and more prevalent along the cortical hierarchy are supported strongly by previous experimental observations.

Although our model can capture several experimental findings, a few interesting questions remain for future studies. For example, our spiking RNN model utilizes connectivity patterns derived from a gradient-descent approach, which is not biologically plausible. It will be important to characterize whether more biologically valid learning mechanisms, such as reinforcement learning or Hebbian learning, also generate spiking networks with heterogeneous neuronal timescales. In addition, our training method did not allow for robust training of RNNs on a DMS task with a long delay window. This was circumvented by training RNNs on a DMS task with a short delay period and identifying the networks that could perform the 750-ms delay DMS task (Methods). It will be important in the future to study if our method can be modified to be more generalizable. Another unexplored aspect is the working memory capacity of our model.

Although the WM task design employed here involves only one WM item, that is, the identity of the cue stimulus (−1 or +1), our proposed circuit mechanism can be extended to store more than one item at a time (Supplementary Fig. 13). Lastly, we have not investigated how our proposed model could be modified to maintain nonbinary stimuli. One possible method would be to connect multiple disinhibitory motifs with overlapping but distinct receptive fields (Supplementary Fig. 14), similar to the bump attractor model proposed previously[42]. Future work will investigate if such a mechanism is also employed in the cortex to sustain nonbinary items in WM. In summary, we have explored a neural circuit mechanism that performs logical computations over time with stable temporal receptive fields.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-00753-w.

Received: 17 February 2020; Accepted: 5 November 2020;
Published online: 07 December 2020

References

1. Murray, J. D. et al. A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).
2. Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H. & Wang, X.-J. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron* **88**, 419–431 (2015).
3. Cavanagh, S. E., Wallis, J. D., Kennerley, S. W. & Hunt, L. T. Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. *eLife* **5**, e18937 (2016).
4. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
5. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
6. Fascianelli, V., Tsujimoto, S., Marcos, E. & Genovesio, A. Autocorrelation structure in the macaque dorsolateral, but not orbital or polar, prefrontal cortex predicts response-coding strength in a visually cued strategy task. *Cereb. Cortex* **29**, 230–241 (2017).
7. Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat. Commun.* **9**, 3498 (2018).
8. Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K. & Stokes, M. G. Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nat. Commun.* **9**, 3499 (2018).
9. Kim, D. et al. Distinct roles of parvalbumin- and somatostatin-expressing interneurons in working memory. *Neuron* **92**, 902–915 (2016).
10. Kamigaki, T. & Dan, Y. Delay activity of specific prefrontal interneuron subtypes modulates memory-guided behavior. *Nat. Neurosci.* **20**, 854–863 (2017).
11. Xu, H. et al. A disinhibitory microcircuit mediates conditioned social fear in the prefrontal cortex. *Neuron* **102**, 668–682 (2019).
12. Cummings, K. A. & Clem, R. L. Prefrontal somatostatin interneurons encode fear memory. *Nat. Neurosci.* **23**, 61–74 (2019).
13. Krabbe, S. et al. Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nat. Neurosci.* **22**, 1834–1843 (2019).
14. Pfeffer, C. K., Xue, M., He, M., Huang, Z. J. & Scanziani, M. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.* **16**, 1068–1076 (2013).
15. Tremblay, R., Lee, S. & Rudy, B. GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron* **91**, 260–292 (2016).
16. Medalla, M., Gilman, J. P., Wang, J.-Y. & Luebke, J. I. Strength and diversity of inhibitory signaling differentiates primate anterior cingulate from lateral prefrontal cortex. *J. Neurosci.* **37**, 4717–4734 (2017).
17. Mongillo, G., Rumpel, S. & Loewenstein, Y. Inhibitory connectivity defines the realm of excitatory plasticity. *Nat. Neurosci.* **21**, 1463–1470 (2018).
18. Kim, R., Li, Y. & Sejnowski, T. J. Simple framework for constructing functional spiking recurrent neural networks. *Proc. Natl Acad. Sci. USA* **116**, 22811–22820 (2019).
19. Qi, X.-L., Meyer, T., Stanford, T. R. & Constantinidis, C. Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. *Cereb. Cortex* **21**, 2722–2732 (2011).
20. Meyer, T., Qi, X.-L., Stanford, T. R. & Constantinidis, C. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *J. Neurosci.* **31**, 6266–6276 (2011).
21. Constantinidis, C., Qi, X.-L. & Meyer, T. Single-neuron spike train recordings from macaque prefrontal cortex during a visual working memory task before and after training. *CRCNS* https://doi.org/10.6080/K0ZW1HVD (2016).
22. Stokes, M. G. et al. Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
23. Spaak, E., Watanabe, K., Funahashi, S. & Stokes, M. G. Stable and dynamic coding for working memory in primate prefrontal cortex. *J. Neurosci.* **37**, 6503–6516 (2017).
24. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
25. Goldman-Rakic, P. S. in *Comprehensive Physiology* (ed. Terjung, R. L.) 373–417 (American Cancer Society, 2011).
26. Meyer, T., Qi, X.-L. & Constantinidis, C. Persistent discharges in the prefrontal cortex of monkeys naïve to working memory tasks. *Cereb. Cortex* **17**, i70–i76 (2007).
27. Yang, G. R., Murray, J. D. & Wang, X.-J. A dendritic disinhibitory circuit mechanism for pathway-specific gating. *Nat. Commun.* **7**, 12815 (2016).
28. Wang, X.-J. & Yang, G. R. A disinhibitory circuit motif and flexible information routing in the brain. *Curr. Opin. Neurobiol.* **49**, 75–83 (2018).
29. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
30. Bouchacourt, F. & Buschman, T. J. A flexible model of working memory. *Neuron* **103**, 147–160 (2019).
31. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS Comput. Biol.* **12**, e1004792 (2016).
32. Miconi, T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6**, e20899 (2017).
33. Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).
34. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
35. Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* **505**, 318–326 (2014).
36. Batista-Brito, R. et al. Developmental dysfunction of VIP interneurons impairs cortical circuits. *Neuron* **95**, 884–895.e9 (2017).
37. Pi, H.-J. et al. Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).
38. Karnani, M. M. et al. Opening holes in the blanket of inhibition: localized lateral disinhibition by VIP interneurons. *J. Neurosci.* **36**, 3471–3480 (2016).
39. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).
40. Kim, Y. et al. Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* **171**, 456–469.e22 (2017).
41. Medalla, M. & Barbas, H. Synapses with inhibitory neurons differentiate anterior cingulate from dorsolateral prefrontal pathways associated with cognitive control. *Neuron* **61**, 609–620 (2009).
42. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).

Methods

Continuous rate RNN model. The spiking RNNs used in the main text were generated by first training their counterpart continuous-variable rate RNNs using a gradient-descent algorithm. After training, the continuous RNNs were converted to LIF RNNs using the method that we developed previously[18]. The continuous RNN model contained $N=200$ recurrently connected units that were governed by

$$\boldsymbol{\tau}^{\mathrm{d}} \cdot \frac{\mathrm{d}\mathbf{x}}{\mathrm{dt}} = -\mathbf{x} + W^{\mathrm{rate}}\mathbf{r}^{\mathrm{rate}} + \mathbf{I}_{\mathrm{ext}} \quad (1)$$

$$\mathbf{r}^{\mathrm{rate}} = \frac{1}{1+\exp(-\mathbf{x})}$$

where $20\,\mathrm{ms} \le \tau^{\mathrm{d}} \le 125\,\mathrm{ms}$ corresponds to the synaptic decay time constants for the N units in the network, $\mathbf{x} \in \mathrm{R}^{1\times N}$ is the synaptic current variable, and $\mathbf{r}^{\mathrm{rate}} \in \mathrm{R}^{1\times N}$ refers to the firing rate estimates of the units. A standard logistic sigmoid function was used to estimate a firing rate of a neuron from its synaptic current (x). The synaptic connectivity matrix ($W^{\mathrm{rate}} \in \mathrm{R}^{N\times N}$) is initialized as a random, sparse matrix drawn from a normal distribution with zero mean and s.d. of $1.5/\sqrt{N\times P_c}$, where $P_c = 0.20$ is the initial connectivity probability.

For the synaptic decay time constants ($\boldsymbol{\tau}^{\mathbf{d}}$) for all the units in a network, we first initialized the constants with random values ranging between 20 ms and 125 ms:

$$\boldsymbol{\tau}^{\mathrm{d}} = \sigma(\mathcal{N}(0,1))\tau_{\mathrm{step}} + \tau^{\mathrm{d}}_{\mathrm{min}}$$

where $\sigma(\cdot)$ is the sigmoid function, $\tau^{\mathrm{d}}_{\mathrm{min}}$ is the minimum time constant (that is, 20 ms), and τ_{step} was used to set the maximum constant value (that is, $\tau^{\mathrm{d}}_{\mathrm{max}} = \tau_{\mathrm{step}} + \tau^{\mathrm{d}}_{\mathrm{min}} = 125$ ms). Backpropagation was then used to optimize the time constants along with the recurrent connections and the readout weights.

The external currents ($\boldsymbol{I}_{\mathrm{ext}}$) include task-specific input stimulus signals (Training details) along with a Gaussian white-noise variable:

$$\mathbf{I}_{\mathrm{ext}} = W_{\mathrm{in}}u + \mathcal{N}(0, 0.01)$$

where the time-varying, task-specific stimulus signals $\left(u \in \mathbb{R}^{N_{\mathrm{in}}\times 1}\right)$ are given to the network via $W_{\mathrm{in}} \in \mathrm{R}^{N\times N_{\mathrm{in}}}$, a Gaussian random matrix with zero mean and unit variance. N_{in} corresponds to the number of input signals associated with a specific task, and $\mathcal{N}(0, 0.01) \in \mathbb{R}^{N\times 1}$ represents a Gaussian random noise with zero mean and variance of 0.01.

A linear readout of the population activity was used to define the output of the rate network:

$$o^{\mathrm{rate}}(t) = W^{\mathrm{rate}}_{\mathrm{out}}\mathbf{r}^{\mathrm{rate}}(t)$$

where $W^{\mathrm{rate}}_{\mathrm{out}} \in \mathbb{R}^{1\times N}$ refers to the readout weights.

Equation (1) is discretized using the first-order Euler approximation method:

$$\boldsymbol{x}_t = \left(\mathbf{1} - \frac{\Delta t}{\boldsymbol{\tau}^{\mathrm{d}}}\right)\cdot \boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau}^{\mathrm{d}}}\cdot\left(W^{\mathrm{rate}}\boldsymbol{r}^{\mathrm{rate}}_{t-1} + W_{\mathrm{in}}u_{t-1}\right) + \mathcal{N}(0, 0.01)$$

where $\Delta t = 5$ ms is the discretization time step size used throughout this study.

Training details. Adam (adaptive moment estimation), a stochastic gradient-descent algorithm, was used to update the synaptic decay variable ($\boldsymbol{\tau}^{\mathrm{d}}$), recurrent connections (W^{rate}) and readout weights $\left(W^{\mathrm{rate}}_{\mathrm{out}}\right)$. The learning rate was set to 0.01, and the TensorFlow default values were used for the first and second moment decay rates. In addition, Dale's principle (that is, separate excitatory and inhibitory populations) was imposed using the method previously proposed[31]. For retraining previously trained RNNs (Fig. 8), only the input weights (W_{in}) were trainable, and the recurrent weights and the readout weights were fixed to their trained values.

Two LIF RNN models were employed in this study by training rate RNNs on two different tasks: DMS and AFC tasks.

DMS RNNs. For the DMS RNN model, the input matrix ($u \in \mathrm{R}^{2\times 500}$) contained two input channels for two sequential stimuli (over 500 time steps with 5 ms step size). The task began with a 1-s (200 time steps) fixation period during which the input matrix was set to 0. During the fixation, stimulus and delay windows, the RNN was required to maintain its output close to 0. The first channel delivered the first stimulus (250 ms in duration) after 1 s (200 time steps) of fixation, while the second channel modeled the second stimulus (250 ms in duration), which began 50 ms after the offset of the first stimulus. The short delay (50 ms) allowed rate RNNs to learn the task efficiently, and the delay duration was increased after training (see below). During each stimulus window, the corresponding input channel was set to either −1 or +1. If the two sequential stimuli had the same sign (−1/−1 or +1/+1), the network was trained to produce an output signal approaching +1 after the offset of the second stimulus. If the stimuli had opposite signs (−1/+1 or +1/−1), then the network produced an output signal approaching −1. The training was stopped when the loss function fell below 7 and the task performance was greater than 95%. After the rate RNNs were trained successfully and converted to LIF networks, a subgroup of LIF RNNs that performed the actual DMS paradigm used in the main text (that is, delay duration set to 750 ms) with accuracy greater than 95% were identified and analyzed. We trained 142 rate RNNs, and a subset of the trained RNNs (41 out of 142 RNNs) were converted successfully to spiking RNNs that could perform the 750-ms delay DMS task. For Figs. 5–7, a group of LIF RNNs that performed the DMS task with accuracy between 60% and 80% was used.

AFC RNNs. The input matrix ($u \in \mathrm{R}^{1\times 350}$) for the AFC paradigm was set to 0 for the first 200 time steps (that is, 1 s fixation). A short stimulus (125 ms in duration) of either −1 or +1 was given after the fixation period. After the stimulus offset, the network was trained to produce an output signal approaching −1 for the '−1' stimulus and +1 for the '+1' stimulus. The training termination criteria were the same as those used for the DMS model above.

Spiking RNN model. For our spiking RNN model, we considered a network of LIF units recurrently connected to one another. These units are governed by:

$$\tau_{\mathrm{m}}\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = -v_i(t) + (x_i(t) + I_{\mathrm{ext}}(t))R \quad (2)$$

where τ_{m} is the membrane time constant (10 ms), $v_i(t)$ is the membrane voltage of unit i at time t, $x_i(t)$ is the synaptic input current that unit i receives at time t, $\boldsymbol{I}_{\mathrm{ext}}$ is the external input current and R is the leak resistance (set to 1). The synaptic input current (x) is modeled using a double-exponential synaptic filter applied to the presynaptic spike trains:

$$x_i = \sum_{j=1}^{N} W^{\mathrm{spk}}_{ij} r^{\mathrm{spk}}_j$$
$$\frac{\mathrm{d}r^{\mathrm{spk}}_i}{\mathrm{dt}} = -\frac{r^{\mathrm{spk}}_i}{\tau^{\mathrm{d}}_i} + s_i$$
$$\frac{\mathrm{d}s_i}{\mathrm{dt}} = -\frac{s_i}{\tau_r} + \frac{1}{\tau_r\tau^{\mathrm{d}}_i}\sum_{t^k_i < t}\delta\left(t - t^k_i\right)$$

where W^{spk}_{ij} is the recurrent connection strength from unit j to unit i, $\tau_r = 2$ ms is the synaptic rise time and τ^{d}_i refers to the synaptic decay time for unit i. The synaptic decay time constant values and the recurrent connectivity matrix were transferred from the trained rate RNNs (more details described in Kim et al.[18]). The spike train produced by unit i is represented as a sum of Dirac δ functions, and t^k_i refers to the kth spike emitted by unit i.

The external current input ($\boldsymbol{I}_{\mathrm{ext}}$) contained task-specific input values along with a constant background current set near the action potential threshold. The output of our spiking model at time t is given by

$$o^{\mathrm{spk}}(t) = W^{\mathrm{spk}}_{\mathrm{out}}\boldsymbol{r}^{\mathrm{spk}}(t)$$

where the readout weights $\left(W^{\mathrm{spk}}_{\mathrm{out}}\right)$ are also transferred from the trained rate RNN model.

Other LIF model parameters included the action potential threshold (−40 mV), the reset potential (−65 mV), the absolute refractory period (2 ms), and the constant bias current (−40 pA). Equation (2) was discretized using a first-order Euler method with $\Delta t = 0.05$ ms.

Electrophysiological recordings. Extracellular recordings, previously published and described in detail[19–21], were analyzed to validate our RNN model. The dataset contained spike-train recordings from four rhesus macaque monkeys before and after they learned two DMS tasks. Briefly, for the pretraining condition, the monkeys were rewarded for maintaining fixation on the center of the screen regardless of the visual stimuli shown throughout the trial (Fig. 8a). For the post-training condition, the monkeys were trained on two DMS tasks: spatial and feature DMS tasks. For the spatial task (Fig. 1b), the monkeys were trained to report if two sequential stimuli matched in their spatial locations. For the feature task, they had to distinguish if two sequential stimuli matched in their shapes. The dataset included spike times from single neurons in the dorsal and ventral PFC, but only the units from the dorsal PFC were analyzed for this study.

Estimation of neuronal timescales. To estimate neuronal timescales, we computed the decay time constant of the spike-count autocorrelation function for each unit during the fixation period[1]. For each unit, we first binned its spike trains during the fixation period over multiple trials using a nonoverlapping 50-ms moving window. Since the fixation duration was 1 s for the experimental data and our model, this resulted in a [Number of Trials × 20] spike-count matrix for each unit. For the experimental data, the minimum number of trials required for a neuron to be considered for analysis was 11 trials. The average number of trials from all the neurons from the post-training condition was 86.8 ± 35.1 (mean ± s.d.) trials. For the pretraining condition, the average number of trials was 95.4 ± 344.4. For the RNN model, we generated 50 trials for each unit.

Next, Pearson's correlation coefficient (ρ) was computed between two time bins (that is, two columns in the spike-count matrix) separated by a lag (Δ). The coefficient was calculated for all possible pairs with a maximum lag of 600 ms. The coefficients were averaged for each lag value, and an exponential decay function

was fitted across the average coefficient values ($\bar{\rho}$) using the Levenberg-Marquardt nonlinear least-squares method:

$$\bar{\rho}(\Delta) = A\left(\exp\left(-\frac{\Delta}{\tau}\right) + B\right) \quad (3)$$

where A and B are the amplitude and the offset of the fit, respectively. The timescale (τ) defines how fast the autocorrelation decays and was used to estimate each neuron's timescale.

The following inclusion criteria (commonly used in previous experimental studies) were applied to the RNN model and the experimental data: (1) minimum average firing rate of 1 Hz during the fixation period for the experimental data and 2.5 Hz for the RNN model, (2) $0 < \tau \leq 500$ ms, (3) $A > 0$ and (4) a first decrease in ρ earlier than $\Delta = 150$ ms. In addition, the fitting was started after the first decrease in autocorrelation. For the experimental dataset, 325 dlPFC units from the post-training condition and 434 units from the pretraining condition satisfied the above criteria. For the DMS RNN model, 931 units from 40 good performance RNNs and 604 units from 26 poor performance RNNs met the criteria. For the AFC model, 1138 units from 40 RNNs satisfied the criteria.

Cross-temporal decoding analysis. The amount of information encoded by each unit was estimated using cross-temporal decoding analysis[8,22,23]. For both experimental and model data, a Gaussian kernel (s.d. = 50 ms) was first applied to the spike trains to obtain the firing rate estimates over time. For each cue stimulus identity, each neuron's firing rate time courses were divided into two splits (even vs. odd trials) and averaged across trials within each split. There were nine cue conditions (that is, nine spatial locations) for the spatial DMS task and eight cue conditions (that is, eight shapes) for the feature DMS task. Within each task, all possible pairwise differences in mean firing rates between any two cue conditions for each neuron in each split were computed. Next, Pearson's correlation coefficient was determined for each pairwise difference condition between the two splits (at each time point across neurons). The correlation coefficients from both tasks (36 pairwise difference conditions for the spatial task and 28 conditions for the feature task) at each time point were averaged after applying the Fisher's z-transformation, resulting in a single measure we refer to as a discriminability or decodability score. The within-delay discriminability scores were computed from the correlation coefficients at $t_1 = t_2$ where t_1 and t_2 refer to the time points used for the two splits. To estimate the stability of the cue stimulus maintenance during the delay window, we averaged the within-delay discriminability scores across the delay period for each RNN. Cross-temporal decoding matrices and within-delay decoding time courses for the dlPFC data (Figs. 3 and 8) were smoothed for better visualization, but all statistical tests were performed on unsmoothed data.

Connectivity rewiring method. For Fig. 4e, we characterized which connection type contributed the most to the long neuronal timescales observed in the DMS RNN model by randomly shuffling connections belonging to each type (I → I, I → E, E → I, or E → E) while preserving the original distribution of the connection types. For the I → I type, all the outward connections from each inhibitory unit to other inhibitory units were first identified. These connections were then rewired randomly in a manner that preserved their connection identity (that is, I → I). This procedure was repeated for the other three synaptic types. For Fig. 5, all the synaptic weights corresponding to each connection type were either decreased or increased by 30% without rewiring.

To quantify the amount of cue-specific information maintained during the delay period in each of the four shuffling conditions (Fig. 4f), we performed the within-delay decoding analysis (see above) for all the units in each RNN per shuffling condition. This resulted in 40 within-delay decoding time courses (one for each RNN) for each rewiring condition.

Cue stimulus selectivity. To identify inhibitory units selective for each of the two cue stimuli (−1 or +1), we computed a cue preference index (θ) for each unit using:

$$\theta_i = \frac{r_{i,+1} - r_{i,-1}}{r_{i,+1} + r_{i,-1}}$$

where $r_{i,+1}$ refers to the average firing rate of unit i across positive cue stimulus trials (50 trials) during the cue stimulus window, while $r_{i,-1}$ indicates the average activity across negative cue stimulus trials (50 trials). Thus, $\theta_i > 0$ indicates that unit i prefers the positive cue stimulus over the negative stimulus. Based on this selectivity measure, two subgroups of inhibitory units (one for $\theta > 0$ and the other for $\theta < 0$) were identified for each DMS RNN.

Spike-count Fano factors. The relationship between spike-count variability and neuronal timescales was investigated by computing trial-to-trial spike-count Fano factors during the fixation period (Fig. 7). For each unit included in the timescale analysis, the variance of the total number of spikes within the 1-s fixation window across trials was first computed. The Fano factor was then calculated by dividing the variance by the mean spike count. The trials used for computing the Fano factors were identical to those used for estimating the neuronal timescales for both neural and RNN data.

Reconfiguring pretrained RNNs. In Fig. 8e–g, the continuous-variable rate RNNs trained to perform the AFC and DMS tasks were used. For Fig. 8e, the input weights (W_{in}) of the AFC and DMS RNNs were retrained via the same gradient-descent algorithm to perform the CTX task (see below). For Fig. 8f, the input weights (W_{in}) of the AFC RNNs were retrained to perform the DNMS task (see below). The I → I connections were either unaltered (yellow in Fig. 8f) or increased by 200% (orange in Fig. 8f). In Fig. 8g, only the input weights for the DMS RNNs were reconfigured to perform the DNMS task. The maximum number of training trials was set to 6,000 trials for computational efficiency.

Context-dependent input integration (CTX) task. The implementation of the context-dependent input integration (CTX) task was identical to the one previously studied[18]. Briefly, the input stimuli contained four streams of signals where the first two channels corresponded to noisy input signals from modality 1 and modality 2, respectively. The last two channels (context signals in Fig. 8d) were used to instruct the network which modality to pay attention to. For example, the third channel was turned on (that is, set to 1 throughout the trial) and the fourth channel was set to 0 to instruct the network to integrate the modality 1 input signal. Each modality input signal was modeled as white-noise signal (sampled form the standard normal distribution) with constant offset bias terms. More details on the implementation of the task paradigm are described in Kim and Sejnowski[18].

Delayed-non-match-to-sample (DNMS) task. The DNMS task paradigm was similar to the DMS task paradigm. The network was trained to produce an output signal approaching +1 if the two sequential input stimuli had opposite signs. If the two input stimuli had the same sign, the network was trained to produce an output signal approaching −1.

Statistical analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[19,20,34]. All the RNNs trained in the study were randomly initialized (with random seeds) before training. Our RNNs were retrained three times and the main findings presented in this study were replicated each time. Blinding was not performed for the RNN model analysis since the data were simulated by the authors. Data points that did not meet the inclusion criteria described above (Estimation of neuronal timescales) were excluded.

Throughout this study, we employed nonparametric statistical methods. For all the figures utilizing boxplots, we used two-sided Wilcoxon rank-sum or signed-rank method to determine statistically significant difference between two groups. For comparing more than two groups, we used either Friedman (Fig. 4e) or Kruskal–Wallis test (Fig. 7c) with Dunn's post hoc test to correct for multiple comparisons. In addition, we employed a nonparametric cluster-based permutation test to account for multiple comparisons and to determine significant discriminability (Fig. 3a) and differences in discriminability between short and long τ subgroups (Figs. 3 and 8) (ref. [43]). To identify significant clusters in the cross-temporal matrices (Figs. 3a and 8c), cue stimulus condition labels were randomly shuffled 1,000 times within each split to construct the null distribution. A point was considered significant if its value exceeded the 95th percentile of the null distribution, and the largest cluster size (that is, number of contiguous points that were significant) from the data was compared against the null distribution of the largest cluster size values to correct for multiple comparisons. To determine if within-delay decoding time courses were significantly different between long and short τ groups (Figs. 3b and 8c), τ group labels were shuffled randomly 1,000 times within each split and each task. Again, a time point was considered significant if it was greater than the 95th percentile of the null distribution. Similar multiple comparison correction, as described above, was applied. More information can be found in the Nature Research Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The trained RNN models used in the present study are deposited as MATLAB-formatted data in Open Science Framework, https://osf.io/md4wg. The experimental data used in the study can be obtained from Constantinidis et al.[21].

Code availability

The code for the analyses performed in this work is available at https://github.com/rkim35/wmRNN.

References

43. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG and MEG data. *J. Neurosci. Methods* **164**, 177–190 (2007).

Acknowledgements

We are grateful to B. Tsuda, Y. Chen and J. Fleischer for helpful discussions and feedback on the manuscript. We also thank J. Aldana for assistance with computing resources. This work was funded by the National Institute of Mental Health (grant no. F30MH115605-01A1 to R.K.). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 graphics processing unit used for this research.

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

R.K. and T.J.S. designed the study and wrote the manuscript. R.K. performed the analyses and simulations.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-00753-w.

Correspondence and requests for materials should be addressed to R.K. or T.J.S.

Peer review information *Nature Neuroscience* thanks Dean Buonomano, Timothy Buschman, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Robert Kim, Terrence J. Sejnowski

Last updated by author(s): Oct 16, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code written in Python (TensorFlow 1.10.0, numpy 1.16.4, scipy 1.3.1) to generate the network models presented in the manuscript is publicly accessible on GitHub (https://github.com/rkim35/wmRNN). The gradient descent optimization algorithm included in TensorFlow 1.10.0 was used throughout the study.
Data analysis	The code for the main analyses performed in this work is available at https://github.co m/rkim35/wmRNN.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The trained recurrent neural network models used in the present study are deposited as MATLAB-formatted data in Open Science Framework, https://osf.io/md4wg. The experimental data used in the study can be obtained from ref. 21.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes. For the modeling portion of the study, multiple networks (minimum 20 networks as established by similar, previous modeling studies mentioned in the main text; ref. 34) were trained for statistical significance.
Data exclusions	For both modeling and experimental data, units/neurons that did not meet the following inclusion criteria were excluded: (1) minimum average firing rate of 1 Hz during the fixation period for the experimental data and 2.5 Hz for the RNN model, (2) neuronal timescales smaller than or equal to 500 ms, (3) amplitude of the exponential decay fit > 0, and (4) a first decrease in the autocorrelation earlier than 150 ms. These pre-established criteria were previously employed by experimental studies probing neuronal timescales.
Replication	The recurrent neural networks were retrained three times and the main findings presented in this study were replicated each time. For the experimental data, previously reported main findings in refs. 19, 20 were replicated successfully using the dataset available in ref. 21.
Randomization	All the recurrent neural networks trained in the study were randomly initialized (with random seeds) before training.
Blinding	Blinding was not performed for the RNN model analysis since the data were simulated by the authors. Blinding was not performed for the experimental data as there was only one experimental group.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging