PNAS RESEARCH ARTICLE | NEUROSCIENCE

Random noise promotes slow heterogeneous synaptic dynamics important for robust working memory computation

Nuttida Rungratsameetaweemana[a,b,1], Robert Kim[b,c,1], Thiparat Chotibut[d,2], and Terrence J. Sejnowski[b,e,f,2]

Affiliations are included on p. 11.

Edited by Tirin Moore, Stanford University School of Medicine, Stanford, CA; received September 26, 2023; accepted December 6, 2024

Recurrent neural networks (RNNs) based on model neurons that communicate via continuous signals have been widely used to study how cortical neural circuits perform cognitive tasks. Training such networks to perform tasks that require information maintenance over a brief period (i.e., working memory tasks) remains a challenge. Inspired by the robust information maintenance observed in higher cortical areas such as the prefrontal cortex, despite substantial inherent noise, we investigated the effects of random noise on RNNs across different cognitive functions, including working memory. Our findings reveal that random noise not only speeds up training but also enhances the stability and performance of RNNs on working memory tasks. Importantly, this robust working memory performance induced by random noise during training is attributed to an increase in synaptic decay time constants of inhibitory units, resulting in slower decay of stimulus-specific activity critical for memory maintenance. Our study reveals the critical role of noise in shaping neural dynamics and cognitive functions, suggesting that inherent variability may be a fundamental feature driving the specialization of inhibitory neurons to support stable information processing in higher cortical regions.

recurrent neural network | working memory | neural dynamics

The brain is hierarchically organized, with higher cortical areas responsible for complex cognitive functions and lower areas managing more basic sensory processes. Previous studies suggest that this hierarchical organization reflects a corresponding gradation in neural processing timescales (1–4). For instance, higher cortical regions, such as the prefrontal cortex, exhibit slower synaptic dynamics, facilitating sustained information processing crucial for tasks that involve working memory and decision-making. These regions can maintain information over extended periods, enabling the integration of complex cognitive processes. In contrast, lower cortical areas operate with faster dynamics, allowing for rapid sensory processing. This gradient in synaptic dynamics across the cortical hierarchy supports a seamless flow of information.

Prior studies suggest that seemingly noisy activities and neuronal variability tend to increase along the cortical hierarchy (5, 6). In particular, neuronal responses in the prefrontal cortex demonstrate significant modulation of variability, driven by task demands (6). Despite such increased variability and pervasive “noisy” processes from various sources, higher cortical areas are still able to maintain information robustly. Whether the slower synaptic dynamics associated with these higher cortical regions observed in previous studies act as a compensatory mechanism for the inherently higher variability present in these areas remains an open question. This relationship between synaptic dynamics and neuronal variability may hold key insights into the specialized functions of higher cortical areas.

There is growing evidence from computational and modeling studies that introducing noise during the training process can lead to improved stability and robustness of neural networks. Specifically, several studies have demonstrated that injecting Gaussian noise during the training process of multilayer perceptron (MLP) and recurrent neural networks (RNNs) can improve their performance (7–9). For example, ref. 9 examined the impact of injecting noise into the hidden states of vanilla RNNs and found that it contributed to stochastic stabilization through implicit regularization (10). Additionally, ref. 8 studied the regularization effects induced by Gaussian noise in MLPs and showed that the explicit regularization provided several benefits, including increased robustness to perturbations. Despite the demonstrated benefits of noise injection in vanilla RNNs and MLPs, it is not yet clear whether these findings extend to more biologically plausible RNNs that incorporate neuronal firing rate and synaptic dynamics. It is also unclear whether

Significance

Our study addresses a fundamental challenge in the field of neural network modeling, offering critical insights into the training of recurrent neural networks (RNNs) that simulate cognitive processes. Specifically, we demonstrate that by introducing random noise during training, we not only expedite the learning process but also establish robust models capable of maintaining information necessary for working memory tasks. Further analyses revealed that the introduction of noise selectively increased the synaptic decay time constants of inhibitory units, leading to a sustained stimulus-specific activity crucial for memory maintenance. Our findings not only shed light on the optimization of RNN training methods but also hold profound implications for understanding how higher cortical areas evolved to compensate for inherent noise to maintain information.

Author contributions: N.R. and R.K. designed research; N.R., R.K., and T.C. performed research; N.R., R.K., and T.C. analyzed data; and N.R., R.K., T.C., and T.J.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] N.R. and R.K. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: thiparatc@gmail.com or terry@salk.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2316745122/-/DCSupplemental.

Published January 16, 2025.

introducing noise consistently results in slower synaptic dynamics or if this phenomenon is specific to cognitive demands.

In this study, we propose a systematic approach to address these questions. Specifically, we investigate the impact of noise during training of firing-rate RNNs to perform tasks that require different cognitive functions, such as decision-making and working memory. We show that the introduction of noise during training significantly enhances the RNN performance on tasks that specifically require working memory. By dissecting the networks trained with noise and employing stability analysis methods, we further show that noise induces slow dynamics in inhibitory units and forces these units to be more active, resulting in more stable memory maintenance. These findings aligned with recent experimental and theoretical studies that place specific subtypes of inhibitory neurons at the center of working memory computations (11–15). Therefore, our study illustrates how seemingly random noise could give rise to slow dynamics specific to working memory, elucidating that the enhanced stability of memory maintenance associated with higher cortical areas could be the result of increased "noise" inherent to these regions.

Results

Biologically Plausible RNN Model and Task Overview. Even though recent advances in deep learning and AI have greatly increased the functionality and capability of artificial neural network models, it is still challenging to train a network of model neurons to perform cognitive tasks that require memory maintenance. Models based on RNNs of continuous-variable firing rate units have been widely used to reproduce previously observed experimental findings and to explore neural dynamics associated with cognitive functions including working memory, an ability to maintain information over a brief period (16–19).

We study the RNN model composed of excitatory and inhibitory rate units governed by Eq. **1**.

$$\tau_i \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} w_{ij}\phi(x_j(t)) + \sum_{j=1}^{C} w_{ij}^{(\text{noise})}\psi_j(t) + \sum_{j=1}^{U} w_{ij}^{(\text{in})}u_j(t) + \xi_i(t) \qquad \textbf{[1]}$$

In Eq. **1**, τ_i and x_i refer to the synaptic decay time-constant and synaptic current variable, respectively, for unit i. The synaptic current variable is converted to the firing-rate estimate via a nonlinear transfer function ($\phi(\cdot)$). Throughout this study, we employed the standard sigmoid function for ϕ. w_{ij} is the synaptic strength from unit j to unit i, and $\boldsymbol{u}(t)$ is the task-specific input data given to the network via $\boldsymbol{w}^{(\text{in})}$ (*Materials and Methods*). Each neuron in the model received external noise ($\boldsymbol{\xi}(t)$). In contrast to conventional rate-based RNN models, our model incorporated what we referred to as inherent noise, where a set of independent noise signals sampled from a standard Gaussian distribution uncorrelated in time ($\boldsymbol{\psi}(t)$) were introduced to the network through $\boldsymbol{w}^{(\text{noise})}$ (*Materials and Methods*). The trainable parameters of the model included $\boldsymbol{w}$, $\boldsymbol{w}^{(\text{noise})}$, $\boldsymbol{\tau}$, $\boldsymbol{w}^{(\text{out})}$, and b. It is important to note that the noise input weights ($\boldsymbol{w}^{(\text{noise})}$) were replaced by a standard Gaussian random matrix during testing (*Materials and Methods*).

The above firing-rate RNN model was trained using backpropagation through time (BPTT; 20) to perform a task that involves maintaining information over a brief period (i.e., working memory task). The task is a delayed match-to-sample (DMS) task that requires the model to match the signs of the two sequential input stimuli (Fig. 1*A*; see *Materials and Methods*). While the model has shown success in various cognitive tasks (16–19), training the model with important biological constraints to perform the DMS task with a long delay period between the two input stimuli remains challenging. Notably, the training time increases exponentially as a function of the delay duration. As shown in Fig. 1*B*, the model required more trials to achieve successful training on the DMS task as the delay interval increased from 50 ms to 250 ms (all Ps < 0.001, two-sided Wilcoxon rank-sum test). Moreover, when the synaptic decay time constants ($\boldsymbol{\tau}$) for all the units in the model were fixed at a small constant (20 ms), the training process failed to converge.

Noise Improves Learning and Enhances Network Resilience on Working Memory Tasks. In order to study the effects of noise on the dynamics of the firing-rate RNNs and their performance on the DMS task, we introduced noise in the form of random Gaussian currents injected into the units during the training process (Fig. 1*C*; see *Materials and Methods*). For each noise level (C; see *Materials and Methods*), we trained 50 RNNs to perform the DMS task with a delay interval of 250 ms. Specifically, there were four stimulus conditions ($s \in \{(+1,+1),(+1,-1),(-1,+1),(-1,-1)\}$). For the matched cases (stimulus conditions 1 and 4), the model had to generate an output signal approaching $+1$. For stimulus conditions 2 and 3 where the signs of the two sequential stimuli were opposite, the model had to produce an output signal approaching -1. As shown in Fig. 1*D*, the training success rate for the baseline model (i.e., no inherent noise; $C = 0$) was 66% (33 out of 50 RNNs were trained within the first 20,000 trials). As the number of the noise channels (C) increased, the training success rate also increased (Fig. 1*D*). When $C = 10$, all 50 RNNs were successfully trained to perform the task (dark green in Fig. 1*D*). For the networks successfully trained, we did not see any significant difference in the number of training trials/epochs required among the four different noise conditions (Fig. 1*E*). We observed a similar trend for a DMS task involving two delay intervals (*Materials and Methods* and *SI Appendix*, Fig. S1). Varying the number of noise signals and their variance values revealed that adding more noise signals with smaller variance was more beneficial than adding fewer noise channels with higher variance (*SI Appendix*, *Comparison of Different Structures of Inherent Noise* and Fig. S2).

As shown in Fig. 1 *D* and *E*, the noise condition of $C = 10$ yielded the highest training efficiency. Importantly, the RNNs trained with this optimal noise structure were more robust to perturbations of internal dynamics compared to the RNNs trained without any injection of inherent noise. Specifically, the RNNs trained with noise exhibited robust performance in the DMS task even when subjected to randomly generated inherent noise introduced via randomly generated $\boldsymbol{w}^{(\text{noise})}$, as opposed to the optimized $\boldsymbol{w}^{(\text{noise})}$ used during training (Fig. 1*F*). In addition, the networks trained with noise demonstrated superior performance in the DMS task with a longer delay duration compared to the networks trained without noise (Fig. 1*F*). Interestingly, making $\boldsymbol{w}^{(\text{noise})}$ nontrainable led to similar results: All 50 RNNs were successfully trained to perform the DMS task, and the number of trials required was not significantly different from the trials required to train RNNs with tunable

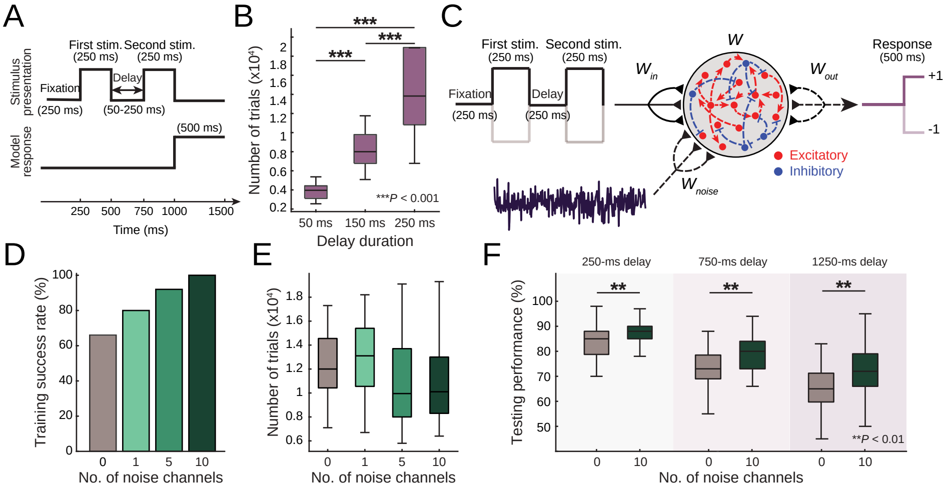

Fig. 1. Delayed match-to-sample (DMS) task and model schematic. (*A*) A schematic diagram of a delayed match-to-sample (DMS) task with two sequential stimuli separated by a delay interval (*Top*). Timeline illustrating the stimulus presentation and model response in a DMS task with a delay interval of 250 ms (*Bottom*). (*B*) The number of trials/epochs needed to train continuous-variable RNNs increases exponentially as the delay interval increases. For each delay duration condition, we trained 50 firing-rate RNNs to perform the DMS task shown in (*A*). The maximum number of trials/epochs was set to 20,000 trials for computational efficiency (all Ps < 0.001, two-sided Wilcoxon rank-sum test). (*C*) A schematic diagram illustrates the paradigm used to train our RNN model on the DMS task in which one delay was present. We introduced and systematically varied the amount of noise in the RNN to study the effects of noise on memory maintenance in a biologically constrained neural network model. The model contained excitatory (red circles) and inhibitory (blue circles) units. The dashed lines represent connections that were optimized using backpropagation. (*D*) Training performance of the RNN models on the DMS task. RNN models with varying amount of noise (i.e., 0, 1, 5, and 10 noise channels) were trained to perform this task. Training success rate was measured as the number of successfully trained RNNs (out of 50 RNNs). (*E*) The average number of trials required to reach the training criteria. (*F*) Testing performance of the RNN models on the DMS task. RNNs successfully trained either without noise (0 noise channels; $n = 33$) or with 10 noise channels ($n = 50$) were tested on the DMS task in which both inherent noise and noisy input signals (external noise) were introduced. We also varied the delay duration of these testing trials to range from 250 ms, 750 ms, and 1,250 ms. For each testing condition, average accuracy of the trained RNN models is shown. Across all conditions, RNNs trained with no noise had lower accuracy than those trained with 10 noise channels (all Ps < 0.01, two-sided Wilcoxon rank-sum test). Boxplot: Central lines, median; *Bottom* and *Top* edges, *Lower* and *Upper* quartiles; whiskers, 1.5× interquartile range; outliers are not plotted.

$\boldsymbol{w}^{(\text{noise})}$ (10,747 ± 3,122 trials for 50 DMS RNNs with tunable $\boldsymbol{w}^{(\text{noise})}$ vs. 11,425 ± 2,436 trials for 50 DMS RNNs with fixed $\boldsymbol{w}^{(\text{noise})}$, mean ± SD; no significant difference by the two-sided Wilcoxon rank-sum test). These results suggest that the injected noise facilitated contextualized sensory encoding and led to a more robust representation of the input stimuli.

To further investigate the impact of inherent noise on the RNN dynamics, we applied the Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE; 21) to the internal state trajectories of one example RNN realization from the baseline ($C = 0$) and noise ($C = 10$) conditions (*Materials and Methods*). When only external noise ($\boldsymbol{\xi}$ in Eq. **1**) was present, both models performed the DMS task with a delay of 250 ms equally well (Fig. 2 *A* and *B*). However, applying PHATE to these two models revealed distinct differences in the dynamics and representations of the four stimulus conditions (Fig. 2 *C* and *D*). In the RNN trained without noise, the neural representations of distinct stimulus conditions were found to intermingle in the lower-dimensional embedding space (Fig. 2*C*). However, in the RNN trained with noise (Fig. 2*D*), the dynamical structures corresponding to the four conditions were clearly demarcated, indicating a more distinct representation of the stimuli. Notably, these neural trajectories exhibited meaningful and informative bifurcations that were driven by the temporal structure of the DMS task (as indicated by the black arrows in Fig. 2*D*). Specifically, the first bifurcation occurred after presentation of the first stimulus (at 250 ms), followed by a second bifurcation at the onset of the second stimulus (at 750 ms). These distinct bifurcations observed in the trajectories over time highlight the role of inherent noise in facilitating contextualized sensory encoding and working memory computation.

Noise Modulates Cell-Type Specific Dynamics Underlying Working Memory Computation. Next, we investigated how the noise facilitated stable maintenance of stimulus information by examining the optimized model parameters. Given the previous studies highlighting the importance of inhibitory connections for information maintenance (11, 13–15), we hypothesized that the inherent noise enhances working memory dynamics by selectively modulating inhibitory signaling. To test this, we first compared the inhibitory recurrent connection weights of the RNNs across different noise conditions ($C = 0, 1, 5, 10$). We did not observe any significant differences in the inhibitory weights (*SI Appendix*, Fig. S4). Similarly, the excitatory recurrent weights were also comparable across the noise conditions (*SI Appendix*, Fig. S4).

As we did not observe any noticeable changes in the recurrent weight structure induced by the noise, we next analyzed the distribution of the optimized synaptic decay time constants ($\boldsymbol{\tau}$). First, we extended our analyses to include a wider range of C by training 50 RNNs for the DMS task with $C = 20$ and $C = 50$. For the $C = 20$ case, 48 out of 50 RNNs were successfully trained within the first 20,000 trials (12,844.7 ± 2,878.9 trials, mean ± SD). In contrast, for the $C = 50$ case, only 8 out of 50 RNNs were successfully trained (16,763.5 ± 1,952.2 trials, mean ± SD). Investigating the synaptic decay time constants across the 6 noise levels ($C \in \{0, 1, 5, 10, 20, 50\}$) revealed that the inhibitory synaptic decay time constant values ($\boldsymbol{\tau}_{\text{inh}}$) were strongly modulated by the number of the noise channels (C): Increasing C led to slower inhibitory synaptic dynamics (Fig. 3). We also observed that the average excitatory synaptic constants (τ_{exc}) started to increase notably when $C > 10$ (Fig. 3*A*). Plotting the difference between the inhibitory and excitatory time constants ($\boldsymbol{\tau}_{\text{inh}} - \boldsymbol{\tau}_{\text{exc}}$) revealed that the difference was maximized when $C = 10$ (Fig. 3*C*). Introducing a greater number of noise

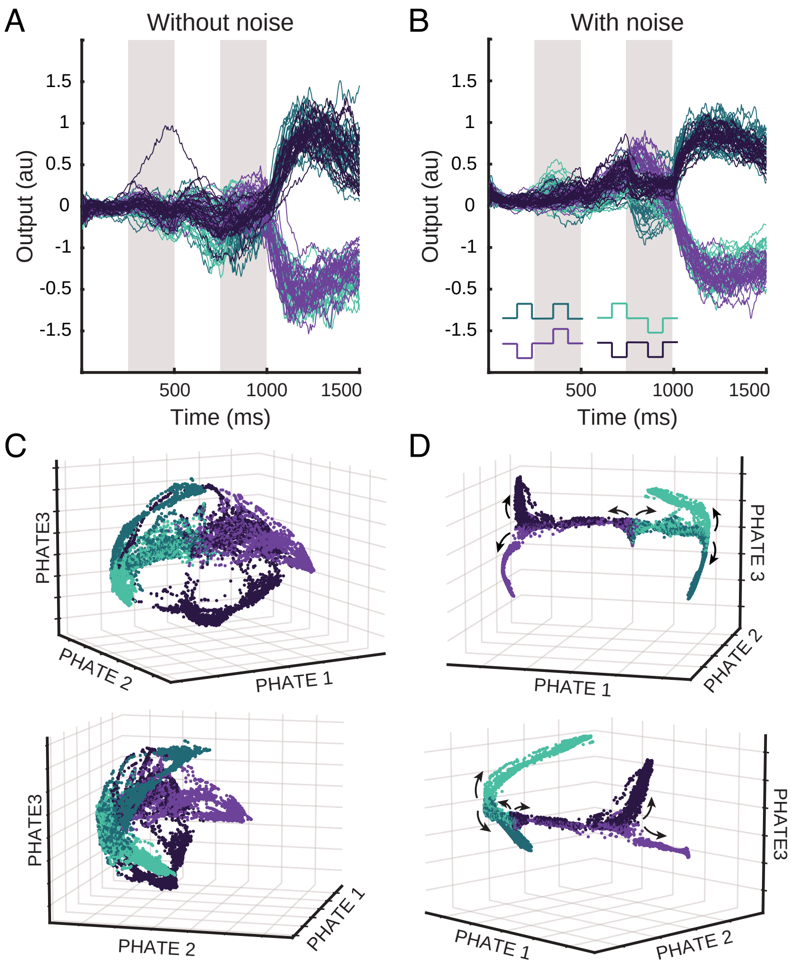

Fig. 2. Neural representations of each stimulus condition on the DMS task. (*A*) Network output of a sample RNN model successfully trained without noise to perform the DMS task with a delay of 250 ms. For match cases ($s \in \{(+1, +1), (-1, -1)\}$), the network accurately generated an output signal approaching +1 (dark green and dark purple tracings). For nonmatch cases ($s \in \{(+1, -1), (-1, +1)\}$), the network produced an output signal approaching −1 (light green and light purple). (*B*) Network output of a sample RNN model trained with noise to perform the same DMS task as (*A*). A schematic of the four stimulus conditions used in the DMS task shown in the *Bottom Left* corner. (*C*) PHATE-embedding of the network activity (from the onset of the first stimulus window) from the RNN shown in (*A*). (*D*) PHATE-embedding derived from the network activity of the RNN trained with noise (same network as *B*). Black arrows indicate temporal progression of the PHATE trajectories over the trial duration.

channels with smaller variance was more effective in maximizing $\boldsymbol{\tau}_{\text{inh}} - \boldsymbol{\tau}_{\text{exc}}$ than using fewer noise channels with larger variance (*SI Appendix*, Fig. S3).

Based on these results, we then hypothesized that the decreased training success rate observed for $C = 20$ and $C = 50$ was due to the large increase in $\boldsymbol{\tau}_{\text{exc}}$. To test this hypothesis, we trained an additional set of 50 RNNs for the $C = 50$ case, with $\boldsymbol{\tau}_{\text{exc}}$ fixed to their initial values and not adjustable during training. Imposing the constraint of fixing $\boldsymbol{\tau}_{\text{exc}}$ resulted in a significant improvement in the training success rate (40 out of 50 RNNs successfully trained), indicating that the noise-induced prolongation of $\boldsymbol{\tau}_{\text{exc}}$ impaired memory maintenance. In addition, we observed similar findings when the number of inhibitory units was increased to match the number of excitatory units (*SI Appendix*, Fig. S5).

Studying the average firing rate activities ($\boldsymbol{r}$) of the excitatory and inhibitory units from the trained DMS RNNs revealed that the inhibitory units, on average, had significantly higher firing rates than the excitatory units (*SI Appendix*, Fig. S6). This significant difference in $\boldsymbol{r}$, along with the strong outgoing inhibitory connections (*SI Appendix*, Fig. S4), likely led to the preferential prolongation of $\boldsymbol{\tau}_{\text{inh}}$ during learning. These findings are in line with recent modeling studies that emphasized the importance of slow inhibitory dynamics in maintaining information (15) and underscore the importance of incorporating inherent noise during training to shape learned dynamics to robustly perform working memory computations.

Noise Pushes Model Neurons with Slow Synaptic Dynamics Toward the Edge of Instability. We next focused on understanding the role of slow inhibitory signaling in the networks trained with noise. For an RNN to perform well on the DMS task, it is plausible that RNN persistently maintains information during the delay window. This condition can be achieved when each unit in the network maintains relatively stable synaptic current activity throughout the delay window, i.e., $\boldsymbol{x}(t) \approx \boldsymbol{x}^*$ at a given time point t during the delay period, where $\boldsymbol{x}^*$ is the delay period steady state. For both models (RNNs trained without and with noise), the synaptic current activity during the delay period exhibited stability (*SI Appendix*, Fig. S7). We then performed the linear stability analysis around $\boldsymbol{x}^*$, revealing the role of slow inhibitory signaling as follows.

For each first stimulus condition, $s_1 \in \{-1, +1\}$, we studied the impact of a small instantaneous perturbation around the stimulus-specific delay period steady state ($\boldsymbol{x}^*_{s_1}$). In the absence of

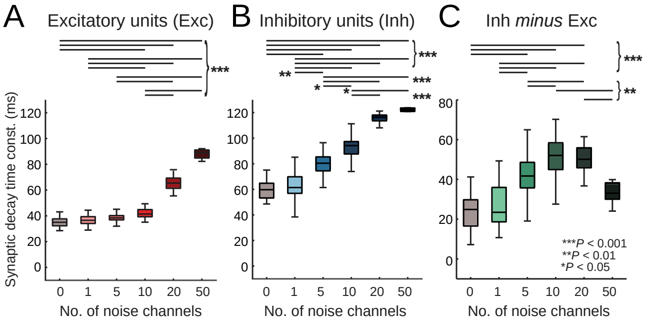

Fig. 3. Influence of noise on cell-type specific temporal dynamics. Comparison of synaptic decay time constants of RNN models trained on the DMS task with varying amount of noise. (*A*) For each noise condition, synaptic decay time constants of successfully trained models are reported for excitatory units ($n = 33$, 40, 46, 50, 48, 8 for the noise conditions of 0, 1, 5, 10, 20, and 50 channels, respectively). Overall, injection of random noise during training increased synaptic decay time constants averaged across excitatory units in the networks ($P\text{s} < 0.001$, $H = 156.4$; Kruskal–Wallis test with Dunn's post hoc test). (*B*) Comparison of synaptic decay time constants for inhibitory units of the trained RNN models ($P\text{s} < 0.05$, $H = 194.2$; Kruskal–Wallis test with Dunn's post hoc test). (*C*) Difference in the trained synaptic decay time constants between the inhibitory (Inh) and excitatory units (Exc) ($P\text{s} < 0.01$, $H = 127.1$; Kruskal–Wallis test with Dunn's post hoc test). Boxplot: Central lines, median; *Bottom* and *Top* edges, *Lower* and *Upper* quartiles; whiskers, 1.5× interquartile range; outliers are not plotted.

an input stimulus and noise, the synaptic current activities evolve according to (modified from Eq. **1**):

$$\frac{dx_i}{dt} = \frac{1}{\tau_i}\left(-x_i + \sum_{j=1}^{N} w_{ij}\sigma(x_j)\right) \equiv F_i(\boldsymbol{x}). \qquad \textbf{[2]}$$

Perturbing $\boldsymbol{x}^*_{s_1}$ by $\delta\boldsymbol{x}_{s_1}$ would lead to

$$\left.\frac{d\boldsymbol{x}}{dt}\right|_{\boldsymbol{x}^*_{s_1}+\delta\boldsymbol{x}_{s_1}} = \boldsymbol{F}(\boldsymbol{x}^*_{s_1}) + J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1} + O(\delta\boldsymbol{x}^2_{s_1}), \qquad \textbf{[3]}$$

where $J(\boldsymbol{x}^*_{s_1})$ is the Jacobian matrix (*Materials and Methods*). Since $\boldsymbol{F}(\boldsymbol{x}^*_{s_1}) \approx \boldsymbol{0}$, the perturbed dynamics (Eq. **3**) can be rewritten as

$$\frac{d\delta\boldsymbol{x}_{s_1}}{dt} \approx J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1} \qquad \textbf{[4]}$$

with the Jacobian matrix written explicitly as

$$J_{ij}(\boldsymbol{x}^*_{s_1}) = \frac{1}{\tau_i}\left[-\delta_{ij} + w_{ij}\sigma(x_j)(1-\sigma(x_j))\right]\Big|_{\boldsymbol{x}=\boldsymbol{x}^*_{s_1}}. \qquad \textbf{[5]}$$

Performing spectral decomposition on J and calculating the eigenvalues (λ) of the example RNN models employed in Fig. 2 revealed that all eigenvalues of J exhibited negative real parts, indicating that the steady states ($\boldsymbol{x}^*_{s_1}$) are indeed stable against mild instantaneous perturbations (Fig. 4 *A*–*D*; see *Materials and Methods*). Interestingly, the RNN model trained with noise contained more slowly relaxing modes with oscillatory behaviors compared to the network trained without noise (i.e., eigenvalues with nonzero imaginary components shifted toward zero along the real axis in Fig. 4 *C* and *D*). Furthermore, these slowly relaxing modes require spatially extended perturbations to trigger a neural response, as evidenced by their low (left) Inverse Participation Ratio (IPR) values (greener dots in Fig. 4 *C* and *D*, and comparison of the average IPR values between the two RNNs shown in Fig. 4*E*; see *Materials and Methods*). Throughout this work, the IPR refers to the IPR of the *left* eigenvectors, as it directly reflects the sensitivity to perturbations. Specifically, a larger IPR indicates a more localized perturbation that affects a smaller number of units is sufficient to trigger a neural response, while a smaller IPR means a more delocalized perturbation affecting a larger number of units is required to stimulate a neural response (*Materials and Methods*). In other words, RNNs trained with noise are more robust compared to the RNNs trained without noise, as they require sustained perturbations to a larger number of units for the steady states to be destabilized.

In order to further characterize the slow relaxation modes observed in the RNN trained with noise, we first identified the units involved in the left eigenvectors corresponding to the top 50 eigenvalues (i.e., 50 least negative eigenvalues) for each RNN model (*Materials and Methods*). We categorize the units with nonzero amplitudes in the top 50 eigenvectors as dominant units (perturbation on these units could more dominantly influence the RNN to destabilize), while the units with zero amplitudes are referred to as nondominant units. Notably, in both RNN models (trained without and with noise), the dominant units were associated with significantly larger synaptic decay time constants compared to the nondominant units (Fig. 4 *F* and *G*). Furthermore, the synaptic decay dynamics of the dominant units in the RNNs trained with noise were significantly slower than the dynamics of the dominant units in the networks trained without noise ($P < 0.001$, two-sided Wilcoxon rank-sum test). Interestingly, the top 50 left eigenmodes from the RNNs trained with noise contained a significantly larger number of inhibitory units than the top eigenmodes from the networks trained without inherent noise (Fig. 4*H*).

These findings suggest that injection of the inherent noise during training resulted in an increased proportion of units exhibiting slower synaptic dynamics (i.e., dominant units). In addition, this noise-induced effect pushed the top left eigenmodes composed of these units closer to the edge of instability (critical boundary between stable and unstable behavior).

To investigate the impact of these factors on working memory, our analysis focused on the sustained maintenance of $s_1 = +1$ during a long delay period (1,250 ms) in both models. Since the networks consisted of units that were selective to each of the first stimulus conditions ($s_1 \in \{+1, -1\}$), the successful maintenance of $s_1 = +1$ during the delay period relied on two key conditions: persistent excitation of the units tuned to $s_1 = +1$ and persistent inhibition of the units tuned to $s_1 = -1$. As shown in Fig. 5*A*, the average normalized firing rate timecourses (normalized by subtracting the average baseline activity; see *Materials and Methods*) of the dominant units preferring $s_1 = +1$ in the top 50 left eigenmodes of the RNNs trained without noise demonstrated higher firing rates during the stimulus presentation of $+1$ compared to the nondominant units selective for $s_1 = +1$. Throughout the delay period, the average normalized firing rate activity of the dominant units exhibited a rapid decay (Fig. 5*A*). Repeating the above analysis on the RNNs trained with noise revealed that the dominant units maintained $s_1 = +1$ at a significantly higher rate throughout the delay window than the dominant units from the networks trained without noise (Fig. 5*B*; $P < 0.001$, two-sided Wilcoxon rank-sum test), consistent with the slow synaptic dynamics seen in Fig. 4.

Next, we directed our attention to the units in the corresponding right eigenmodes to investigate the impact of the slow dynamics observed in the top left eigenvectors on the dynamics of the network response (*Materials and Methods*). We hypothesized that the slow decay and persistent activity observed in the units

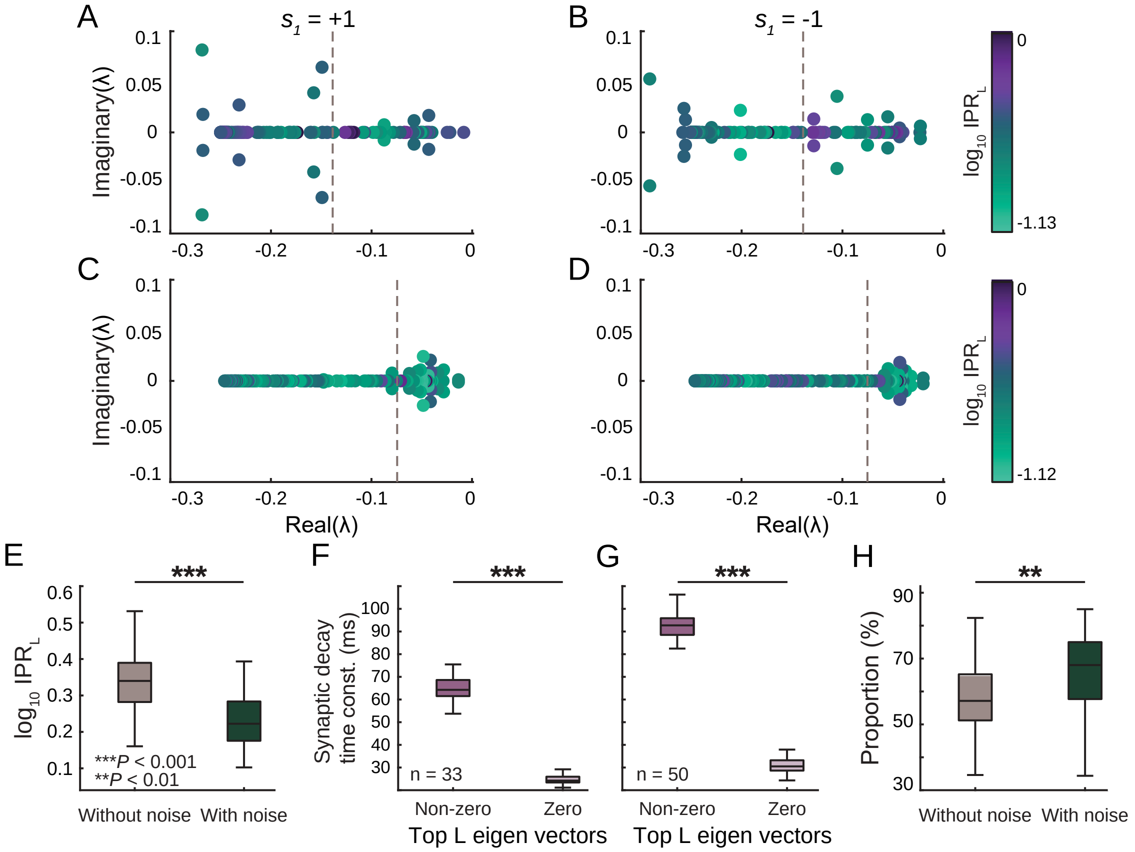

Fig. 4. Spectral characteristics of network stability during the delay period. Spectra of the Jacobian (*J*) extracted from the network activity during the delay window. (*A* and *B*) Spectra of a sample RNN model trained without noise (same RNN as Fig. 2*A*) during the delay period following the first stimulus presentation ($s_1 \in \{+1, -1\}$). (*C* and *D*) Spectra of a sample RNN model trained with noise ($C = 10$; same network as Fig. 2*B*) during the delay period following the first stimulus presentation ($s_1 \in \{+1, -1\}$). For both conditions, we observed stable steady states $\boldsymbol{x}^*_{s_1}$ as evident from the real parts of all the eigenvalues being negative. For the RNN trained with noise, the eigenvalues with nonzero imaginary parts shifted to the right (toward zero along the real axis) and were associated with lower Inverse Participation Ratio (IPR) values (*C* and *D*). Vertical dashed lines represent the cutoff for the top 50 real eigenvalues. (*E*) Average IPR values from the RNN trained without noise were significantly higher (i.e., more localized) than those from the model trained with noise. (*F*) Average synaptic decay time constants of the dominant (nonzero elements in the top 50 eigenvectors) and nondominant (zero elements in the top 50 eigenvectors) units from all the RNNs trained without noise. (*G*) Average synaptic decay time constants of the dominant and nondominant units from all the RNNs trained with noise. (*H*) Proportion of inhibitory units among the dominant units in the RNNs trained with noise was significantly higher compared to the RNNs trained without noise. Boxplot: Central lines, median; *Bottom* and *Top* edges, *Lower* and *Upper* quartiles; whiskers, 1.5× interquartile range; outliers are not plotted. Two-sided Wilcoxon rank-sum tests were performed.

of the top left eigenmodes would confer similar properties to the units in the corresponding right eigenvectors during the delay period. Additionally, we posited that the units tuned for $+1$ and -1 in the right eigenmodes would exhibit persistent excitation and inhibition, respectively. As shown in Fig. 5 *C* and *D*, the units tuned for $+1$ in the right eigenmodes of the RNNs trained with noise demonstrated significantly higher activity during the delay period compared to the $+1$-preferring units in the right eigenmodes of the networks trained without noise. Furthermore, the units tuned for -1 in the right eigenmodes of the RNNs trained with noise exhibited significant suppression throughout the delay period compared to the units tuned for -1 in the right eigenmodes of the networks trained without noise (Fig. 5 *E* and *F*). These results suggest that the networks trained with noise exhibit greater robustness to perturbations compared to the RNNs trained without noise, and the noise-induced increase in synaptic decay time constants of the inhibitory units near the edge of instability facilitated maintenance of stimulus-specific information for an extended duration.

Robustness and Increased Efficiency Due to Inherent Noise Are Specific to Working Memory Computations. Finally, we asked whether the modulatory effects of noise during training were specific to working memory dynamics. To address this question, we employed two cognitive tasks that do not require maintenance of sensory information over time, namely sensory detection or go/no-go (GNG) task and context-dependent sensory integration (CTX) task (*Materials and Methods*). In the GNG task (Fig. 6*A*), the RNN model had to generate an output signal that indicated whether a target sensory signal was present. The CTX task is a more challenging variant of the GNG task, where the model was trained to produce an output that corresponded to one of the two input modalities as determined by a context signal (16) (Fig. 6*B*). As these task paradigms do not involve any delay interval, the model only requires minimal information maintenance, if any, to perform well on these tasks.

Our findings demonstrated that the RNN models were able to perform these non–working memory tasks well without any noise and that adding noise during training did not further improve training efficiency. In fact, it took longer for models to reach successful training criteria when noise ($C > 10$) was added during training compared to the no noise condition for both sensory detection and context-dependent sensory integration tasks (Ps < 0.001 for both tasks, Kruskal–Wallis test with Dunn's post hoc test). To investigate whether noise modulated the temporal dynamics on these tasks, we

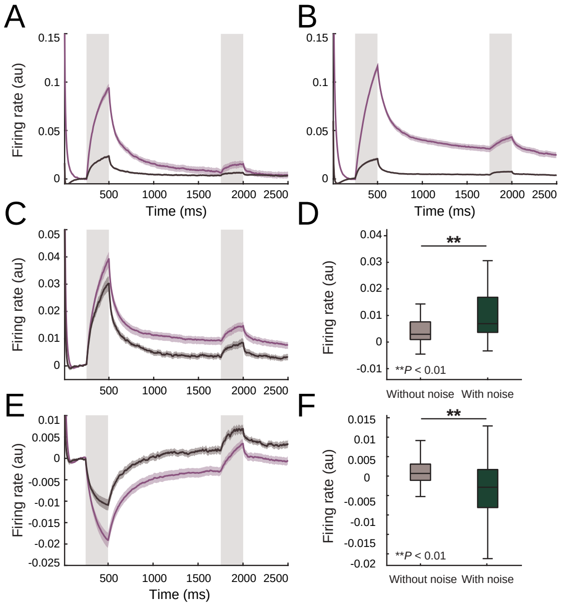

Fig. 5. Persistent activity of dominant units from RNNs trained with noise. (*A*) Average normalized firing rate timecourses of the dominant units selective to +1 (purple) and nondominant units preferring +1 (dark gray) in the top 50 left eigenmodes of the RNNs trained without noise. The DMS task was modified to have a delay duration of 1,250 ms, and 50 trials with the first stimulus of +1 were used to extract the timecourses. (*B*) Average normalized firing rate timecourses of the dominant units selective to +1 (purple) and nondominant units preferring +1 (dark gray) in the top 50 left eigenmodes of the RNNs trained with noise. (*C*) Average normalized firing rate timecourses of the units preferring +1 in the corresponding right eigenvectors from the RNNs trained without noise (dark gray) and with noise (purple). (*D*) Average normalized firing rate of the units shown in (*C*) during the late delay period (last 750 ms) was significantly higher for the RNNs trained with noise compared to the networks trained without noise. (*E*) Average normalized firing rate timecourses of the units preferring −1 in the corresponding right eigenvectors from the RNNs trained without noise (dark gray) and with noise (purple). (*F*) Average normalized firing rate of the units shown in (*E*) during the late delay period (last 750 ms) was significantly lower for the RNNs trained with noise compared to the networks trained without noise. Mean ± SE shown. Boxplot: Central lines, median; *Bottom* and *Top* edges, *Lower* and *Upper* quartiles; whiskers, 1.5× interquartile range; outliers are not plotted. Two-sided Wilcoxon rank-sum tests were performed.

analyzed the synaptic decay time constants of the excitatory and inhibitory units. Our results revealed that the difference in the inhibitory and excitatory time constants ($\tau_{\text{inh}} - \tau_{\text{exc}}$) was not strongly modulated by C for both tasks (Fig. 6 *C* and *D* and *SI Appendix*, Fig. S8).

As shown in *SI Appendix*, Fig. S8, the inhibitory time constants from both GNG and CTX RNNs were significantly larger than the values from the DMS RNNs. Given that the GNG and CTX tasks were considerably easier to learn compared to the DMS task (547 ± 97 trials across 50 GNG RNNs, 1,521 ± 243 trials across 50 CTX RNNs, 12,225 ± 2,891 trials across 33 DMS RNNs, mean ± SD), we trained 50 RNNs each for the GNG and CTX tasks with a minimum of 12,000 trials (matching the average number of training trials used for the DMS RNNs) to ensure a more equivalent comparison. As shown in *SI Appendix*, Fig. S9, training the RNNs for a longer duration resulted in significantly decreased inhibitory synaptic time constants compared to those from the RNNs without the minimum trial constraint for both GNG and CTX tasks. Given the significant changes in the synaptic decay time constants induced by lengthening the training duration, we repeated our analyses from Fig. 6 *C* and *D* with the minimal number of training trial set to 12,000. For conciseness, we focused on comparing the no-noise case with the 10 noise channels ($C = 10$) case. Imposing the minimum training trial constraint led to quantitatively similar results: Adding noise during training did not result in dramatic increases in the inhibitory synaptic time constants for both GNG and CTX tasks. More importantly, we did not observe notable increases in $\boldsymbol{\tau}_{\text{inh}} - \boldsymbol{\tau}_{\text{exc}}$ (*SI Appendix*, Fig. S10).

These findings suggest that the slow synaptic decay dynamics induced by noise are specific to working memory functioning where robust information maintenance is needed to ensure successful performance.

Discussion

In this study, we demonstrated that introducing random noise into firing-rate RNNs allowed the networks to achieve efficient and stable memory maintenance critical for performing working memory tasks. We also showed that the models trained with noise were able to generalize to sustain stimulus-related information longer than the delay period used during training. Further analyses uncovered that the introduction of noise led to the emergence of inhibitory units with slow synaptic decay dynamics, which were predominantly associated with dominant eigenmodes situated near the edge of instability. These eigenmodes were critical for maintaining information during the delay period of the working memory task. Specifically, the network should exhibit stability to prevent minor noisy perturbations from causing substantial alterations in its dynamics and compromising the information of the stimulus. However, the network should not be overly stable, as this would result in rapid decay of the information associated with the stimulus, as the network quickly returns to its stable configuration. Hence, these eigenmodes crucial for robust memory maintenance emerge near the edge of instability. In addition, these effects were specific to the models trained to perform working memory task, suggesting that noise-induced changes were specific to working memory.

Previous studies have demonstrated that neuronal variability increases along the cortical hierarchy (1–3, 5) and that higher cortical areas, including the prefrontal cortex, tend to exhibit high trial-to-trial variability (4, 6, 15, 22). The high variability associated with these higher cortical areas is thought to be due to their involvement in integrating and processing complex, abstract information (4, 23, 24). This increased variability may also result from the integration of bottom–up and top–down processes, a complexity that low-level areas do not typically handle (1, 25). Whether such high variability and other sources of noise intrinsic to the higher-order cortical regions could engender stable network dynamics for supporting working memory is an important open question. By providing an easy-to-use framework for understanding how inherent noise influences information maintenance and learning dynamics when performing memory-dependent cognitive tasks, we aimed to address this question in our current work.

One limitation of the present study is the lack of comparisons with RNNs trained with learning algorithms that are not based on gradient-descent optimization. One such algorithm is First-Order Reduced and Controlled Error (FORCE) learning which has been employed to train rate and spiking RNNs (26, 27). Due to the nature of the method, it is currently not possible to train the synaptic decay time constant term using FORCE training, making the comparison with our models difficult. Reinforcement learning is another learning algorithm that can be employed to train biologically realistic RNNs (17).

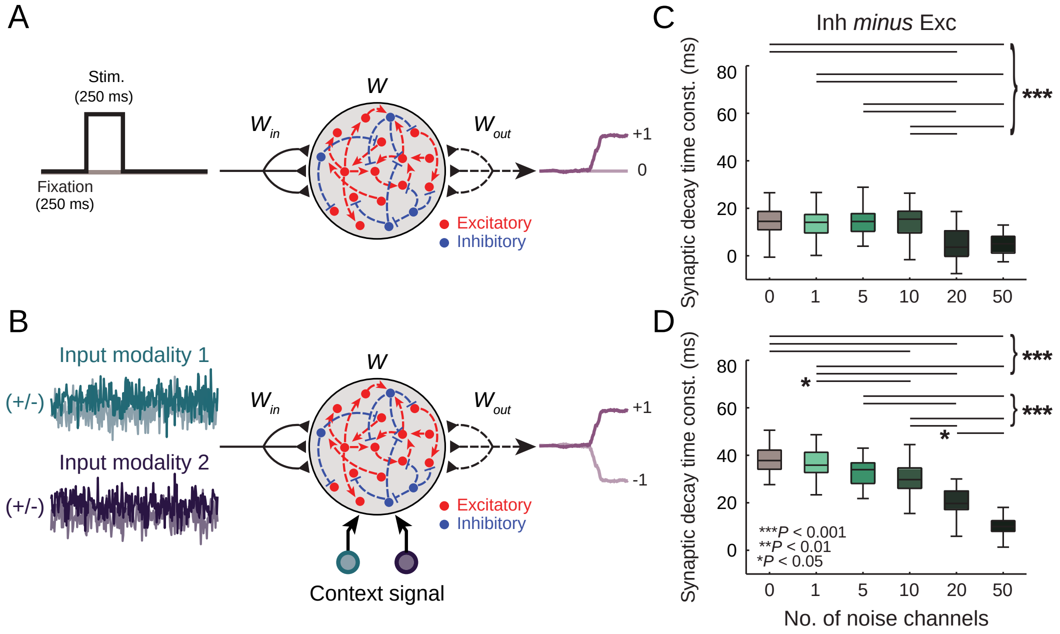

Fig. 6. Network functional motifs underlying working memory-independent computation. Schematics diagrams illustrating working memory-independent tasks and the corresponding network dynamics of the RNN models successfully trained on these tasks. (*A*) Sensory detection or go/no-go (GNG) task, in which the RNN modes were trained to produce an output indicating the presence of a brief input pulse. (*B*) Context-dependent sensory integration (CTX) task, where the RNN models were trained to generate an output based on the identity of a sensory stimulus whose relevance was determined by an explicit context cue. (*C*) Difference in the synaptic decay time constants between the inhibitory (Inh) and excitatory units (Exc) in the RNN models trained on the GNG task (Ps < 0.001, $H = 113.4$; Kruskal–Wallis test with Dunn's post hoc test). (*D*) Difference in the synaptic decay time constants between the inhibitory (Inh) and excitatory units (Exc) in the RNN models trained on the CTX task (Ps < 0.05, $H = 209.0$; Kruskal–Wallis test with Dunn's post hoc test). Boxplot: Central lines, median; *Bottom* and *Top* edges, *Lower* and *Upper* quartiles; whiskers, 1.5× interquartile range; outliers are not plotted.

Even though we showed that increasing the number of inherent noise channels could lead to heterogeneous synaptic decay time constants, the theoretical basis behind the preferential tuning of the inhibitory synaptic decay constants associated with working memory is not clear. As briefly mentioned in Results, strong inhibitory signaling and the resulting strong suppression of excitatory units could possibly explain why gradient updates prefer inhibitory units. Future work will focus on better understanding the theoretical and computational basis for the emergence of slow inhibitory synaptic dynamics.

Relatedly, recent findings (28) provide critical insights into the mechanisms underlying the emergence of long behavioral timescales during temporal tasks involving complex sensory signals and high cognitive demands. Their work demonstrates that these long timescales can originate either from the intrinsic properties of single neurons or from recurrent network interactions, with the latter being a more optimal mechanism for flexibly supporting working memory computations across multiple levels of difficulty. As the inherent noise used in our current work modulates not only the individual neuronal synaptic time constants but also recurrent connections, our findings further support the notion that enhancing network dynamics, rather than solely relying on the intrinsic properties of individual neurons, provides a more resilient and efficient mechanism for handling complex cognitive tasks. However, a systematic approach to tease out the contributions of individual neuron properties and network interactions is warranted to fully understand their respective roles in supporting working memory.

Furthermore, while our findings highlight the benefits of noise-induced modulation of synaptic time constants in working memory tasks, these dynamics may also be advantageous in scenarios requiring prolonged input integration for decision-making due to sensory ambiguity or sequential presentation. This suggests that the mechanisms identified in the present study could significantly enhance the performance of neural networks across a wide range of cognitive tasks involving complex, naturalistic signals, extending well beyond the confines of working memory computation. Future research should investigate this possibility by examining cognitive tasks with varied sensory complexities and integration demands. Such studies would not only confirm the generalizability of the present findings but also deepen our understanding of how noise-enhanced dynamics can improve the functionality of biological networks, particularly in environments characterized by complex sensory signals.

By interpreting the concept of noise within the context of biology, the present study proposes a general framework that bridges recent advances in machine intelligence with empirical findings in neuroscience. Our approach involves introducing inherent noise into a biologically plausible artificial neural network model during training to simulate aspects of cortical noise and systematically evaluating its effects on model dynamics and performance under various testing conditions. This approach outlines a theoretical framework that aims to bridge computational findings with biological observations. While our model serves as an initial step toward understanding how noise might influence neural dynamics, it is important to acknowledge that these results are derived from a computational perspective using simplified network models. Therefore, further research is essential to establish more robust links between behaviors of artificial networks and true biological functions. Ultimately, the present study underscores the potential of computational models as tools for advancing our understanding of how noise influences cognitive functions and its implications for clinical applications.

Materials and Methods

Continuous-Rate RNN Model. We constructed our biologically realistic RNN model based on Eq. **1**. All the units in the network are governed by Eqs. **1**, **6**, and **7**.

$$r_i(t) = \sigma(x_i(t)) = \frac{1}{1 + \exp(-x_i(t))} \qquad [6]$$

$$o(t) = \boldsymbol{w}^{(\mathrm{out})}\boldsymbol{r}(t) + b \qquad [7]$$

τ_i is the synaptic decay time constant of unit i, x_i is the synaptic current variable of unit i, w_{ij} is the synaptic weight from unit j to unit i, and r_i is the firing rate estimate of unit i (estimated by using the sigmoid transfer function in Eq. **6**). Each model contains 200 units. To adhere to previous empirical observations regarding the proportion of excitatory and inhibitory units in the brain, we constructed each RNN with a composition of 80% excitatory and 20% inhibitory units [i.e., E-I ratio of 80/20; (29–31)]. The model receives time-varying input composed of U channels of signals over T time steps ($\boldsymbol{u} \in \mathbb{R}^{U \times T}$) via the input weight matrix, $\boldsymbol{w}^{(\mathrm{in})} \in \mathbb{R}^{N \times U}$. For the DMS task, $\boldsymbol{u}$ contained two streams of input signals (i.e., $U = 2$). The network also receives random noise via $\boldsymbol{w}^{(\mathrm{noise})} \in \mathbb{R}^{N \times C}$, where C is the number of independent noise signals in $\boldsymbol{\psi} \in \mathbb{R}^{C \times T}$. Each signal in $\boldsymbol{\psi}$ was independently drawn from the standard Gaussian distribution with zero mean and unit variance. We considered $C \in \{0, 1, 5, 10, 20, 50\}$ in this study. The external noise ($\boldsymbol{\xi} \in \mathbb{R}^{N \times T}$; uncorrelated in time) was generated from a Gaussian distribution with zero mean and a variance of 0.01. The output (o) of the network was computed as a weighted average of the activities of the units via the readout weights ($\boldsymbol{w}^{(\mathrm{out})}$) and the constant bias term (b).

We numerically simulate the following discretized dynamics obtained from the first-order approximation method (correct up to the noise amplitude) with the step size (Δt) of 5 ms (Eq. **8**).

$$\boldsymbol{x}_t = \left(\boldsymbol{1} - \frac{\Delta t}{\boldsymbol{\tau}}\right)\boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau}}\left(\boldsymbol{w}\boldsymbol{r}_{t-1} + \boldsymbol{w}^{(\mathrm{noise})}\boldsymbol{\psi}_{t-1} + \boldsymbol{w}^{(\mathrm{in})}\boldsymbol{u}_{t-1}\right) + \xi_{t-1} \qquad [8]$$

$1/\boldsymbol{\tau}$ denotes a diagonal matrix whose ith diagonal element is $1/\tau_i$. Note that, to ensure mathematically precise conversion of continuous-time variances σ_ψ^2 and σ_ξ^2 of the Gaussian inherent noise $\psi_j(t)$ and of the Gaussian external noise $\xi_i(t)$ of Eq. **1** into their discrete-time counterparts, the noise amplitudes of Eq. **8** must be rescaled using Euler–Maruyama discretization. Specifically, the noise amplitude in the discrete-time description must be scaled by a factor of $\sqrt{\sigma_\psi^2/\Delta t}$ for the inherent noise and $\sqrt{\sigma_\xi^2/\Delta t}$ for the external noise, respectively. While converting discrete-time parameters back to their continuous-time equivalents is necessary to recover the exact continuous-time noise variances in Eq. **1**, this variance rescaling does not affect the overall findings of this work. For simplicity, our simulation results are performed using Eq. **8**. The network was trained using BPTT. The trainable parameters of the model included $\boldsymbol{w}$, $\boldsymbol{w}^{(\mathrm{noise})}$, $\boldsymbol{\tau}$, $\boldsymbol{w}^{(\mathrm{out})}$, and b. During testing, $\boldsymbol{w}^{(\mathrm{noise})}$ was replaced by a standard Gaussian random matrix.

To further impose biological constraints, we incorporated Dale's principle (separate populations for excitatory and inhibitory units) using methods similar to those implemented in previous studies (17, 32).

Instead of fixing the synaptic decay constant ($\boldsymbol{\tau}$) to a fixed value for all the units, we optimized the parameter for each unit using a similar algorithm similar to the method described in ref. 32. The parameter was trained to range from 20 ms to 125 ms to model heterogeneous synaptic dynamics of different receptors in the cortex (33, 34). We initialized the synaptic decay time constant parameter ($\boldsymbol{\tau}$) using

$$\tau_i = \sigma(\mathcal{N}(0, 1))\tau_{\mathrm{step}} + \tau_{\mathrm{min}}, \qquad [9]$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathcal{N}(0, 1)$ refers to the standard normal distribution. $\tau_{\mathrm{min}} = 20$ ms and $\tau_{\mathrm{step}} = 105$ ms were used to constrain the parameter to range from 20 ms to 125 ms. The gradient of the cost function with respect to the synaptic decay term is derived in *SI Appendix*.

The schematic diagram of the model is shown in Fig. 1*C*. All the models were implemented with TensorFlow 1.10.0 and trained on NVIDIA GPUs (Quadro P4000 and Quadro RTX 4000).

DMS Task. Two delayed match-to-sample (DMS) tasks were used to train our RNN model and assess how the noise influenced the robustness of memory maintenance in the network. Both tasks involved two sequential stimuli (each lasting 250 ms) separated by a delay interval of 250 ms. The first stimulus was presented after a fixation period of 250 ms. During the stimulus window, the input signal ($\boldsymbol{u}$) was set to either -1 or $+1$ (Fig. 1*A*). If the signs of the two sequential stimuli matched (i.e., stimulus condition 1: $s = (+1, +1)$; stimulus condition 4: $s = (-1, -1)$; Fig. 3*A*), the model was trained to produce an output signal approaching $+1$. When the signs were opposite (i.e., stimulus condition 2: $s = (+1, -1)$; stimulus condition 3: $s = (-1, +1)$; Fig. 3*A*), the model had to produce an output signal approaching -1. For the first task, the model had to respond immediately after the second stimulus (Fig. 1 *A* and *C*). A second delay period of 250 ms was added after the second stimulus for the second task (*SI Appendix*, Fig. S1*A*). Due to the two delay periods, the second DMS task is considered a more challenging working memory task than the first task. The primary focus of the present study is the one-delay DMS task, and all the DMS findings presented in the main text are exclusively derived from this specific paradigm. The results for the two-delay DMS task are shown in *SI Appendix*, Fig. S1.

Training Protocol. Our model training was deemed successful if the following two criteria were satisfied within the first 20,000 epochs:

- Loss value (defined as the root mean squared error between the network output and target signals) <7
- Task performance (defined as the average accuracy of the network output over 100 randomly generated testing trials) >95%

If the network did not meet the criteria within the first 20,000 epochs, the training was terminated. For each task and each value of $C \in \{0, 1, 5, 10, 20, 50\}$, we trained 50 RNNs using the above strategy. We considered the RNNs trained with $C = 0$ (i.e., without any noise) as the baseline model.

Testing Protocol. To evaluate the robustness and stability of the trained RNNs, we devised a series of testing conditions where different aspects of the one-delay DMS task (Fig. 1*F*) were systematically manipulated. During testing, inherent noise and noisy input signals were introduced to the trained networks. For each successfully trained RNN, we generated $\boldsymbol{w}^{(\mathrm{noise})}$ and $\boldsymbol{\psi}$ as identically distributed Gaussian random variables to deliver random noise during testing.

For the noisy input signal, white-noise signals (drawn from the standard normal distribution) were added to the sensory signals ($\boldsymbol{u}$) to mimic stimulus-related noise. Additionally, we also varied the duration of the delay interval to range from 250 ms to 1,250 ms (with a 500-ms increment) to assess the stability of memory maintenance (Fig. 1*F*).

Working Memory-Independent Tasks. In addition to the DMS tasks that require memory maintenance over time, we designed two additional cognitive tasks that do not involve working memory computation. By comparing the dynamics of the RNNs between the DMS tasks and these working memory-independent tasks, we were able to identify the specific network dynamics associated with working memory computation.

For the sensory detection or GNG task, our RNN model was trained to produce an output signal approaching $+1$ when a stimulus was presented (250 ms in duration), following a fixation period of 250 ms. For a trial where a stimulus was not presented, the model had to maintain the output signal close to 0 (Fig. 6*A*). For the CTX task, the model received two streams of noisy stimulus signals (input modality 1 and input modality 2; Fig. 6*B*) along with a constant-valued, context signal which informed the model which sensory input modality was relevant on each trial. A random Gaussian time series signal with zero mean and unit variance was used to simulate a noisy sensory input signal. Each time series signal was then shifted by a positive or negative constant offset value to encode sensory evidence toward either the positive or negative choice, respectively. The magnitude of the offset value determined the degree of evidence for the specific choice (positive/negative) represented in the relevant noisy input signal. The network had to generate an output signal approaching $+1$ or -1 in response to the cued input signal with a positive or negative mean, respectively. Thus, if the cued input signal was generated with a positive offset value, the network was expected to produce an output that approached $+1$ irrespective of the mean of the irrelevant input signal. For both the GNG and CTX tasks, the training termination criteria were similar to those used for the DMS (*Training Protocol*).

Visualization of Network Dynamics. To visualize the neural dynamics of working memory computation as a function of injected inherent noise during training, we employed the PHATE algorithm (21). This dimensionality reduction technique is a manifold learning algorithm that enables faithful visualization of high-dimensional data while best preserving the global data structure. Two example RNN models successfully trained either without ($C = 0$) or with noise ($C = 10$) were presented with a simulation of 100 DMS test trials (25 from each of the four stimulus conditions). The delay interval was fixed at 250 ms, such that the temporal structure of the testing phase mirrored that of the training environment (Fig. 1*C*).

We then used the resulting neural activity data from each model type during this testing phase as input data for PHATE in order to compute the low-dimensional embedding corresponding to the neural activity of the RNNs trained with and without noise. Specifically, for each of the RNNs trained under each noise condition (without or with noise), the diffusion operator matrix was first calculated using pairwise similarities among individual points in the input network activity time series (downsampled by a factor of 5). This matrix was raised to a power exponent to amplify the local structure while preserving the global structure of the input data. The resulting matrix was then used to generate the low-dimensional embedding that captures the neural dynamics of the input data.

To characterize potential topological patterns within the neural dynamics associated with each RNN, clustering was performed on this PHATE-generated embedding. Specifically, a K-means clustering algorithm was used to partition the data into distinct groups based on their spatial proximity in the low-dimensional space. For visualization purposes, a 3-dimensional PHATE embedding of a sample model from each noise condition (i.e., without noise and with noise; Fig. 2 *C* and *D*) was plotted and colored by stimulus conditions. Black arrows were also included to indicate the temporal evolution of the neural trajectories over the trial duration. These embeddings provided insights into the temporal structure underlying working memory computation associated with the network dynamics that resulted from the incorporation of inherent noise during training.

Network Stability Analysis During the Delay Interval. To investigate the neural dynamics associated with memory maintenance, we employed linear stability analysis. Specifically, we performed this analysis on the synaptic currents of the RNNs successfully trained without or with noise during the delay period in the DMS task (i.e., from the offset of the first stimulus to the onset of the second stimulus (Fig. 1*C*). Throughout this window, the network activities exhibited consistent steady-state patterns, as illustrated in *SI Appendix*, Fig. S7.

For each first stimulus condition $s_1 \in \{-1, +1\}$, we defined the steady-state synaptic current variable ($\boldsymbol{x}^*_{s_1}$) by first averaging $\boldsymbol{x}_{s_1}(t)$ across time within the delay window and then averaging across multiple trials (50 trials per each first stimulus condition). The impact of a small instantaneous perturbation around the delay period steady state $\boldsymbol{x}^*_{s_1}$ on the synaptic current patterns is determined by the deterministic dynamics of Eq. **1** in the absence of an input stimulus and is shown in Eq. **2**.

For a weak perturbation $\delta\boldsymbol{x}_{s_1}$ around $\boldsymbol{x}^*_{s_1}$, the linearized approximation of the perturbed dynamics is $\frac{d\boldsymbol{x}}{dt}\Big|_{\boldsymbol{x}^*_{s_1}+\delta\boldsymbol{x}_{s_1}} = \boldsymbol{F}(\boldsymbol{x}^*_{s_1}) + J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1} + O(\delta\boldsymbol{x}^2_{s_1})$, where $J(\boldsymbol{x}^*_{s_1})$ is the Jacobian matrix $J_{ij}(\boldsymbol{x}^*_{s_1}) = \frac{\partial F_i}{\partial x_j}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*_{s_1}}$. By the assumption of the steady state $\boldsymbol{x}^*_{s_1}$, which is also consistent with the numerical results, we have $\boldsymbol{F}(\boldsymbol{x}^*_{s_1}) \approx \boldsymbol{0}$. Thus, the linearized dynamics of the perturbation $\delta\boldsymbol{x}_{s_1}$ can be written as Eqs. **4** and **5**.

Network responses to weak perturbations around the steady states can now be systematically explored by the spectral analysis (eigenvalues and eigenvectors) of the Jacobian in Eq. **5**.

For brevity, we will add the subscript *s* only when the stimuli-specific statement is needed. Also, J will denote the Jacobian evaluated at the steady state of interest. In this notation, given the linearized perturbed dynamics of Eq. **4**, the initial perturbation $\delta\boldsymbol{x}_0$ will evolve into the response at time t, $\delta\boldsymbol{x}(t)$, that can be studied via the spectral decomposition of J (35) as

$$\delta\boldsymbol{x}(t) = \sum_{n=1}^{N} e^{\lambda_n t}\boldsymbol{v}^R_n\left(\boldsymbol{v}^L_n\delta\boldsymbol{x}_0\right), \qquad [\mathbf{10}]$$

where $\boldsymbol{v}^L_n$ and $\boldsymbol{v}^R_n$ are, respectively, the left and the right eigenvector of J with the eigenvalue λ_n. Notably, our trained RNNs exhibit highly asymmetric $\boldsymbol{w}$ such that the Jacobian Eq. **5** is non-Hermitian, leading to distinct left and right eigenvectors.

Eq. **10** states that an initial perturbation $\delta\boldsymbol{x}_0$ via $\boldsymbol{v}^L_n$ will contribute to a response $\boldsymbol{v}^R_n$, such that the response will grow (decay) exponentially on the timescale of $\left|1/\mathrm{Re}\left(\lambda_n\right)\right|$ when $\mathrm{Re}\left(\lambda_n\right) > 0$ ($\mathrm{Re}\left(\lambda_n\right) < 0$).

Since the dominant responses to a perturbation depend on the overlap between the perturbation and the top-most left eigenvectors $\left(\boldsymbol{v}^L_n\delta\boldsymbol{x}_0\right)$, the non-zero elements of the top-most left eigenvectors determine the spatial extent of perturbation required to significantly influence the system's response. Along this line, the larger the number of non-zero elements in the top-most left eigenvectors, the larger the number of units that need to be perturbed to destabilize the steady states.

We employ the IPR, a measure commonly used in the study of localization phenomena in statistical physics (36), to reflect the number of units participating in the perturbation. The IPR provides valuable insights into the localization of perturbations by indicating the number of units involved in the perturbation process. In particular,

$$\mathrm{IPR}(\lambda_n) = \frac{\sum_{i=1}^{N}\left|(\boldsymbol{v}_n)_i\right|^4}{\left(\sum_{i=1}^{N}\left|(\boldsymbol{v}_n)_i\right|^2\right)^2}. \qquad [\mathbf{11}]$$

The IPR of the left and the right eigenvector will be denoted by IPR_L and IPR_R respectively, though we will focus on IPR_L as we are interested in the size of the neural subpopulations required to be perturbed to initiate a neural response. Note that the maximum and the minimum values of IPR_L are attained at, respectively, 1 when only a single neuron is non-zero, and $1/N$ when all the units are uniformly activated. A larger or a smaller value of IPR_L indicates that the perturbation is localized around a smaller number of units, or extended over a larger number of units, respectively.

Stimulus Selective Units. To identify units selectively tuned for each of the first stimulus condition ($s_1 \in \{+1, -1\}$), we first generated 50 trials for each stimulus condition and computed average firing rates ($\boldsymbol{r}$) during the first stimulus presentation window. Next, we performed a one-sided Wilcoxon rank-sum test for each unit to determine its selectivity.

Firing Rate Normalization. For the firing rate timecourses (Fig. 5), we normalized the trial-averaged firing rate of each unit by subtracting its corresponding baseline trial-averaged firing rate. The baseline activity was determined by considering the window preceding the onset of the first stimulus.

Statistical Analyses. All the RNNs trained in the present study were randomly initialized (with random seeds) before training. Throughout this study, we employed non-parametric statistical methods to assess statistically significant differences between groups. For comparing differences between two groups (e.g., the $\log_{10}$ $\mathrm{IPR_L}$ of RNNs trained with or without noise), we used the two-sided Wilcoxon rank-sum or signed-rank test. For comparing more than two groups (e.g., the synaptic decay time constants associated with RNNs trained with varying degree of noise), we used the Kruskal–Wallis test with Dunn's post hoc test to correct for multiple comparisons.

Data, Materials, and Software Availability. MATLAB-formatted data have been deposited in Noisy RNNs Project (https://osf.io/dqy3g/) (37).

ACKNOWLEDGMENTS. We would like to thank Jet Li, Leon Dietrich Gruber, and James Yu for their insightful comments. We thank Jorge Aldana and the Computational Neurobiology Laboratory at the Salk Institute for assistance with computing resources. This work was supported by the Swartz Foundation (N.R.), the Kavli Institute for Brain and Mind at the University of California San Diego (N.R.), the National Institute of Biomedical Imaging and Bioengineering R01EB026899-01 (T.J.S.), National Institute of Neurological Disorders and Stroke R01NS104368 (T.J.S.), Mission funding from the cooperative agreement under the United States Army Research Laboratory W911NF-17-S-0003-03 (N.R.), the funding from the National Research Council of Thailand #1187111 on the fiscal year 2023 (T.C.), and from Thailand Science Research and Innovation Fund Chulalongkorn University (IND66230005) (T.C.). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of The U.S. Army Combat Capabilities Development Command, Army Research Laboratory or the United States Government.

Author affiliations: [a]Department of Biomedical Engineering, Columbia University, New York, NY 10027; [b]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; [c]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA 90048; [d]Department of Physics, Chula Intelligent and Complex Systems, Chulalongkorn University, Bangkok 10330, Thailand; [e]Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093; and [f]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

1. J. D. Murray *et al.*, A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).
2. N. M. Dotson, S. J. Hoffman, B. Goodell, C. M. Gray, Feature-based visual short-term memory is widely distributed and hierarchically organized. *Neuron* **99**, 215–226 (2018).
3. C. A. Runyan, E. Piasini, S. Panzeri, C. D. Harvey, Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
4. R. Gao, R. L. Van den Brink, T. Pfeffer, B. Voytek, Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife* **9**, e61277 (2020).
5. R. L. Goris, J. A. Movshon, E. P. Simoncelli, Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014).
6. C. Hussar, T. Pasternak, Trial-to-trial variability of the prefrontal neurons reveals the nature of their engagement in a motion discrimination task. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21842–21847 (2010).
7. A. B. Dieng, J. Altosaar, R. Ranganath, D. M. Blei, Noise-based regularizers for recurrent neural networks. *OpenReview* (2018). https://openreview.net/pdf?id=ryk77mbRZ.
8. A. Camuto, M. Willetts, U. Simsekli, S. J. Roberts, C. C. Holmes, Explicit regularisation in gaussian noise injections. *Adv. Neural Inf. Proces. Syst.* **33**, 16603–16614 (2020).
9. S. H. Lim, N. B. Erichson, L. Hodgkinson, M. W. Mahoney, Noisy recurrent neural networks. *Adv. Neural Inf. Proces. Syst.* **34**, 5124–5137 (2021).
10. G. Blanc, N. Gupta, G. Valiant, P. Valiant, "Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process" in *Conference on Learning Theory*, J. Abernethy, S. Agarwal, Eds. (PMLR, 2020), pp. 483–513.
11. S. Krabbe *et al.*, Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nat. Neurosci.* **22**, 1834–1843 (2019).
12. K. A. Cummings, R. L. Clem, Prefrontal somatostatin interneurons encode fear memory. *Nat. Neurosci.* **23**, 61–74 (2019).
13. G. Mongillo, S. Rumpel, Y. Loewenstein, Inhibitory connectivity defines the realm of excitatory plasticity. *Nat. Neurosci.* **21**, 1463–1470 (2018).
14. H. Xu *et al.*, A disinhibitory microcircuit mediates conditioned social fear in the prefrontal cortex. *Neuron* **102**, 668–682 (2019).
15. R. Kim, T. J. Sejnowski, Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nat. Neurosci.* **24**, 129–139 (2021).
16. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
17. H. F. Song, G. R. Yang, X. J. Wang, Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS Comput. Biol.* **12**, e1004792 (2016).
18. T. Miconi, Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6**, e20899 (2017).
19. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X. J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
20. P. J. Werbos, Backpropagation through time: What it does and how to do it. *Proc. IEEE* **78**, 1550–1560 (1990).
21. K. R. Moon *et al.*, Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
22. M. N. Shadlen, W. T. Newsome, The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–3896 (1998).
23. C. J. Honey *et al.*, Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
24. C. Baldassano *et al.*, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).
25. Y. Lerner, C. J. Honey, L. J. Silbert, U. Hasson, Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
26. D. Sussillo, L. Abbott, Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
27. W. Nicola, C. Clopath, Supervised learning in spiking neural networks with force training. *Nat. Commun.* **8**, 2208 (2017).
28. S. Khajehabdollahi *et al.*, Emergent mechanisms for long timescales depend on training curriculum and affect performance in memory tasks. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2309.12927 (Accessed 29 June 2024).
29. S. H. Hendry, H. Schwark, E. Jones, J. Yan, Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex. *J. Neurosci.* **7**, 1503–1519 (1987).
30. A. Alreja, I. Nemenman, C. J. Rozell, Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices. *PLoS Comput. Biol.* **18**, e1009642 (2022).
31. C. C. Sherwood *et al.*, Scaling of inhibitory interneurons in areas v1 and v2 of anthropoid primates as revealed by calcium-binding protein immunohistochemistry. *Brain Behav. Evol.* **69**, 176–195 (2007).
32. R. Kim, Y. Li, T. J. Sejnowski, Simple framework for constructing functional spiking recurrent neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22811–22820 (2019).
33. R. Duarte, A. Seeholzer, K. Zilles, A. Morrison, Synaptic patterning and the timescales of cortical dynamics. *Curr. Opin. Neurobiol.* **43**, 156–165 (2017).
34. A. M. Zador, L. E. Dobrunz, Dynamic synapses in the cortex. *Neuron* **19**, 1–4 (1997).
35. F. L. Metz, I. Neri, T. Rogers, Spectral theory of sparse non-hermitian random matrices. *J. Phys. A Math. Theor.* **52**, 434003 (2019).
36. E. Abrahams, *50 Years of Anderson Localization* (World Scientific, 2010).
37. N. Rungratsameetaweemana, R. Kim, T. Chotibut, T. J. Sejnowski, Data from "Random noise promotes slow heterogeneous synaptic dynamics important for robust working memory computation." OSF. https://osf.io/dqy3g. Deposited 31 December 2024.