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

Nuttida Rungratsameetaweemana [1,2 *], Robert Kim [2,3 *], Thiparat Chotibut [4], Terrence J. Sejnowski [5,6]

[1] Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA

[2] Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[3] Neurology Department, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[4] Chula Intelligent and Complex Systems, Department of Physics, Chulalongkorn University, Bangkok, Thailand

[5] Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093, USA

[6] Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA

[*] Equal contribution

Correspondence: thiparatc@gmail.com (T.C.) and terry@salk.edu (T.J.S.)

## Abstract

Recurrent neural networks (RNNs) based on model neurons that communicate via continuous signals have been widely used to study how cortical neurons 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. Critically, the training process becomes difficult when the synaptic decay time constant is not fixed to a large constant number for all the model neurons. We hypothesize that the brain utilizes intrinsic cortical noise to generate a reservoir of heterogeneous synaptic decay time constants optimal for maintaining information. Here, we show that introducing random, internal noise to the RNNs not only speeds up the training but also produces stable models that can maintain information longer than the RNNs trained without internal noise. Importantly, this robust working memory performance induced by incorporation of internal noise during training is attributed to an increase in synaptic decay time constants of a distinct subset of inhibitory units. This increase leads to slower decay of stimulus-specific activity, which plays a critical role in memory maintenance.

## Introduction

It is widely acknowledged that the cortex exhibits a high level of spontaneous activity that appears unrelated to task-specific neural codes or behaviors. However, recent works have demonstrated that such "cortical noise" contains information about the environmental context and has a direct impact on downstream behavioral outcomes [1–3]. For instance, Musall et al. [1] showed that the cortical noise in mice contains information about the visual stimulus even in the absence of a task, suggesting that it may play a role in sensory processing. Similarly, Stringer et al. [3] found that the cortical noise in mice contains information about the animal's location and movement speed, which is crucial for navigation. Furthermore, previous studies have also shed light on the significance and relevance of cortical noise to cognitive processes. For example, Caron et al. [4] showed that the random structures of the olfactory system in *Drosophila* optimized the diversity of odor representations in neural circuits. Together, these findings challenge the traditional view of cortical noise as mere "background noise," highlighting its potential role in cognitive functions.

In addition to the experimental findings, 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 multi-layer perceptron (MLP) and recurrent neural networks (RNNs) can improve their performance [5–7]. For example, Lim et al. [7] examined the impact of injecting noise into the hidden states of vanilla RNNs and found that it contributed to stochastic stabilization through implicit regularization [8]. Additionally, Camuto et al. [6] 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 dynamics. It is also unclear if introducing noise can improve the cognitive capabilities of these RNNs. We hypothesize that incorporating noise into such biologically plausible RNNs will give rise to persistent activity, which in turn will be crucial for enhancing working memory performance.

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's 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 [9–13]. Therefore, our study illustrates how seemingly random noise in the cortex could lead to specific changes in synaptic dynamics critical for complex cognitive functioning.

## Results

**Biologically plausible RNN model and task overview.** Even though recent advances in deep learning and artificial intelligence (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 recurrent neural networks (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 [14–17].

We study the following RNN model composed of excitatory and inhibitory rate units:

$$\tau_i \frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} w_{ij}\phi(x_j) + (\boldsymbol{w}_{in})_i \boldsymbol{u} \tag{1}$$

where $\tau_i$ 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 $\phi$. $w_{ij}$ is the synaptic strength from unit $j$ to unit $i$, and $\boldsymbol{u}(t)$ is the task-specific input data given to the network. The input signals are given to neuron $i$ via $(\boldsymbol{w}_{in})_i$.

The above firing-rate RNN model was trained using backpropagation through time (BPTT; [18]) 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 (Figure 1a; see *Methods*). While the model has shown success in various cognitive tasks [14–17], 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 Figure 1b, the model required more trials to achieve successful training on the DMS task as the delay interval increased from 50 ms to 150 ms and 250 ms (all $P$s $< 0.001$, two-sided Wilcoxon rank-sum test).

Moreover, when the synaptic decay time constant ($\tau$) was fixed at a small constant (i.e., fast decay rate), 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 (Figure 1c; see *Methods*). For each noise level ($C$; see *Methods*), we trained 50 RNNs to perform the DMS task with a delay interval of 250 ms. Specifically, there are 4 stimulus conditions ($s = +1/+1$, $s = +1/-1$, $s = -1/+1$, and $s = -1/-1$). For the matched cases (stimulus condition 1 and 4), the model had to generate an output signal approaching $+1$. For stimulus condition 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 Figure 1d, the training success rate for the baseline model (i.e., no internal 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 (see *Methods*), the training success rate also increased (see *Supplementary Materials*). When $C = 10$, all 50 RNNs were successfully trained to perform the task (dark green in Figure 1d). 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 ($C \in \{0, 1, 5, 10\}$; Figure 1e). We observed a similar trend for a DMS task involving two delay intervals (see *Methods*; see *Supplementary Materials*).

As shown in Figure 1d and 1e, the noise condition of $C = 10$ yielded the highest training efficiency. Importantly, the RNNs trained with this optimal noise structure were also more robust to perturbations of noisy input signals and internal dynamics (see *Methods*) and could perform the DMS task with longer delay periods as compared to the RNNs trained without any injection of internal noise (Figure 1f). 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 internal noise on the RNN dynamics, we applied the Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE; [19]) to the internal state trajectories of the RNNs trained with and without noise (see *Methods*). Applying this dimensionality reduction method to one example RNN realization from the baseline ($C = 0$) and noise ($C = 10$) conditions revealed distinct differences in the dynamics and representations of the four stimulus conditions (Figure 2a). In the RNN trained without noise, the neural representations of distinct stimulus conditions were found to intermingle in the lower-dimensional embedding space (Figure 2b). However, in the RNN trained with noise (Figure 2c), the dynamical structures corresponding to the four conditions were clearly demarcated, indicating a more distinct representation of the stimuli. Notably, these neural trajectories exhibit meaningful and informative bifurcations that are driven by the temporal structure of the DMS task (as indicated by the black arrows in Figure 2c). Specifically, the first bifurcation occurs upon 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 injected internal noise in facilitating contextualized sensory encoding and working memory computation, as evidenced by the clear segregation in the trajectory patterns.

**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 [9, 11–13], we hypothesized that the internal 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 (see *Supplementary Materials*). Similarly, the excitatory recurrent weights were also comparable across the noise conditions (see *Supplementary Materials*).

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 ($\boldsymbol{\tau}$). Interestingly, the synaptic decay constant distribution shifted toward the maximum value (125 ms; see *Method*) for the RNNs trained with noise (Figure 3a). Separating the distribution of the inhibitory units from the excitatory units revealed that the change in the decay dynamics was mainly attributable to the shift in the inhibitory synaptic decay dynamics (Figure 3c). In addition, the extent of the shift was correlated with the number of the noise channels ($C$): as $C$ increased, the inhibitory synaptic signals decayed slower (see *Supplementary Materials*). We also observed an increase in the decay time constant in the excitatory population as the level of noise increased (Figure 3b). Notably, when comparing the changes in the population decay time constants between inhibitory and excitatory groups, the noise-induced slowing dynamics were more prominent in the inhibitory subpopulation ($P$s $< 0.001$, $H = 89.3$; Kruskal-Wallis test with Dunn's post hoc test). These findings are in line with recent modeling studies that emphasized the importance of slow inhibitory dynamics in maintaining information [13].

Since the RNNs trained with noise showed an increase in the inhibitory synaptic decay time-constant, we explored whether increasing the inhibitory $\boldsymbol{\tau}$ would enhance the robustness of RNNs trained without noise. To test this hypothesis, we used the example RNN trained without noise (same network as the one shown in Figure 2b). Despite the low-dimensional representations of the stimulus conditions appearing blended (Figure 2b), the network exhibited high accuracy in performing the DMS task (Figure 3d). When $\boldsymbol{\tau}$ for all the units in the network were increased to the maximum value (i.e., 125 ms), the network's performance significantly decreased (Figure 3e). We also observed that increasing the inhibitory $\boldsymbol{\tau}$ to 125 ms, while keeping the excitatory $\boldsymbol{\tau}$ at its original value, impaired the task performance (Figure 3f). Together, these findings underscore the importance of incorporating internal noise during training to shape learned dynamics and enhance the network's capacity to robustly perform working memory computations.

**Noise pushes model neurons with slow synaptic dynamics toward the edge of instability.** Given that artificially increasing the inhibitory synaptic time constants in the RNNs trained without noise did not lead to improved memory maintenance (Figure 3f), we next focused on understanding the role of slow inhibitory signaling in the networks trained with noise. Operating under the assumption that a robust RNN generates stable and persistent activity patterns to maintain information, we performed linear stability analysis around $\boldsymbol{x}(t) \approx \boldsymbol{x}^*$ 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 (see *Supplementary Materials*).

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 an input stimulus, we have the following equation (modified from Equation (1)):

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

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

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

where $J(\boldsymbol{x}^*_{s_1})$ is the Jacobian matrix (see *Methods*). Since $\boldsymbol{F}(\boldsymbol{x}^*_{s_1}) \approx \boldsymbol{0}$, the perturbed dynamics (Equation (3)) can be re-written as

$$\frac{d\delta\boldsymbol{x}_{s_1}}{dt} \approx J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1}, \tag{4}$$

with the Jacobian matrix written explicitly as

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

Performing spectral decomposition on $J$ and calculating the eigenvalues ($\boldsymbol{\lambda}$) of the example RNN models employed in Figure 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 (Figure 4a–h; see *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 non-zero imaginary components shifted toward zero along the real axis in Figure 4e–h). Furthermore, these modes characterized by slow relaxation dynamics were found to exhibit de-localization, as evidenced by their low Inverse Participation Ratio (IPR) values (greener dots in Figure 4e–h and comparison of average IPR values between the two RNNs shown in Figure 4i; see *Methods*). Specifically, a larger IPR indicates a more localized perturbation that affects a smaller number of units, while a smaller IPR corresponds to a more delocalized perturbation affecting a larger number of units. 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 state 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 ten eigenvalues (i.e., ten least negative eigenvalues) for each RNN model (see *Methods*). We categorize the units with non-zero amplitudes in the top ten eigenvectors as dominant units, while the units with zero amplitudes are referred to as non-dominant 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 non-dominant units (Figure 4j and 4k). 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).

These findings suggest that the injection of 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 eigenmodes composed of these units closer to the edge of instability (critical boundary between stable and unstable behavior). Next, we analyzed the firing rate activities of the dominant and non-dominant units in the two models. As shown in Figure 5a, the firing rate timecourses of the dominant units (dark purple) in the RNN trained without noise were not significantly different from those of the non-dominant units (light purple) during the delay period following the first stimulus presentation. In contrast, the dominant units in the RNN example model trained with noise showed elevated firing rates throughout the delay period (Figure 5b), implying that these units sustain the stimulus information through persistent firing. Performing the above analysis on all trained models revealed similar findings (Figure 5c and 5d). By comparing the average delay period firing rate of the dominant units in the two models, we observed that the dominant units in the RNNs trained with noise exhibited significantly higher activity compared to the dominant units in the noise-free RNNs (Figure 5e). No significant differences were observed in the average delay period activity of the non-dominant units between the two models (Figure 5f). These findings strongly suggest that training with noise induced the top eigenmodes to contain units with slow synaptic dynamics conducive for sustaining information for extended periods.

**Robustness and increased efficiency due to intrinsic noise are specific to working memory computations.** Finally, we asked if the modulatory effects of noise during training were specific to working memory dynamics. To address this question, we devised two cognitive tasks that do not require maintenance of sensory information over time, namely two-alternative forced choice (AFC) task and context-dependent sensory integration (CTX) task (see *Methods*). In the AFC task (Figure 6a), 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 AFC 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 [14] (Figure 6c). 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 for either task. In fact, it took longer for models to reach successful training criteria when noise was added during training for both sensory integration and context-dependent sensory integration tasks ($P$s $< 0.001$ for both tasks). To investigate if noise modulated the temporal dynamics on these tasks, we analyzed synaptic decay time constants of all the units as well as separately for excitatory and inhibitory units. Our results revealed no difference in the synaptic decay dynamics in the inhibitory units from the models that trained without noise and those trained with noise (Figure 6b and d). 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. Furthermore, the stability and perturbation analyses of the CTX RNNs revealed that the networks trained with noise were not more robust compared to the models trained without noise (see *Supplementary Materials*).

## 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. 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.

Our findings are closely related to the previous studies that reported the benefits of random neural noise ubiquitous in the cortex in memory recall and associative learning [20, 21]. For example, recent experiments showed that a high level of noise and randomness in the olfactory system (i.e., random and seemingly unstructured networks in the piriform cortex) allows for not only flexible encoding of sensory information but also maintenance of the encoded information [4, 22–24]. Consistent with this line of work, the injected noise in our RNN models during the training helped stabilize the encoding of sensory space and thus enhanced learning efficiency. Taken together, our study provides an easy-to-use framework for understanding how internal noise influences information maintenance and learning dynamics when performing working memory cognitive tasks.

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 [25, 26]. 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 [27].

Even though we showed that increasing the number of noise channels could lead to heterogeneous synaptic decay time constants, it is unclear why only inhibitory synaptic decay constants undergo significant changes for working memory tasks. Future work will focus on better understanding the theoretical and computational basis for the emergence of slow inhibitory synaptic dynamics.

By interpreting the concept of noise in machine learning 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 includes introducing internal noise into a biologically realistic artificial neural network model during training to simulate cortical noise and systematically evaluating its effects on model dynamics and performance under different testing conditions. Elucidating the computational underpinnings of how cortical noise modulates cognitive functions will help us better understand how such processes are disrupted in neuropsychiatric conditions such as schizophrenia and autism spectrum disorder. Finally, our framework has the potential to shed light on the fundamental mechanisms that may give rise to the therapeutic effects of deep brain stimulation (DBS), a neuromodulation technique that entails the targeted delivery of electrical stimulation to specific brain regions.

## Methods

**Continuous-rate recurrent neural network (RNN) model.** We constructed our biologically realistic RNN model based on Equation (1). All the units in the network are governed by the following equations:

$$\tau_i \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} w_{ij} r_j(t) + (\boldsymbol{w}_{noise})_i \boldsymbol{\psi}(t) + (\boldsymbol{w}_{in})_i \boldsymbol{u}(t) + \xi_i(t) \tag{6}$$

$$r_i(t) = \sigma(x_i(t)) = \frac{1}{1 + \exp(-x_i(t))} \tag{7}$$

$$o(t) = \boldsymbol{w}_{out} \boldsymbol{r}(t) + b \tag{8}$$

where $\tau_i$ 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 Equation (7)). 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; [28–30]). 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, $\boldsymbol{w}_{in} \in \mathbb{R}^{N \times U}$ ($(\boldsymbol{w}_{in})_i$ refers to the input weight matrix for neuron $i$). The input signal ($\boldsymbol{u}$) represents task-specific incoming sensory information. The network also receives random noise via $\boldsymbol{w}_{noise} \in \mathbb{R}^{N \times C}$ where $C$ is the number of independent noise signals in $\boldsymbol{\psi} \in \mathbb{R}^{C \times T}$. Each signal in $\boldsymbol{\psi}$ was drawn from the standard normal Gaussian distribution (i.e., zero mean and unit variance). We considered $C \in \{0, 1, 5, 10\}$. The sensory noise ($\boldsymbol{\xi} \in \mathbb{R}^{N \times T}$) was modeled with a Gaussian noise, uncorrelated in time, with zero mean and 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 ($\boldsymbol{w}_{out}$) and the constant term ($b$).

The dynamics were discretized using the first-order Euler approximation method and with the step size ($\Delta t$) of 5 ms:

$$\boldsymbol{x}_t = \left(\mathbf{1} - \frac{\Delta t}{\boldsymbol{\tau}}\right) \boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau}} (\boldsymbol{w} \boldsymbol{r}_{t-1} + \boldsymbol{w}_{noise} \boldsymbol{\psi}_{t-1} + \boldsymbol{w}_{in} \boldsymbol{u}_{t-1}) + \boldsymbol{\xi}_{t-1} \tag{9}$$

where $\boldsymbol{x}_t = \boldsymbol{x}(t)$ and $1/\boldsymbol{\tau}$ denotes a diagonal matrix whose $i^{\text{th}}$ diagonal element is $1/\tau_i$. The network was trained using backpropagation through time (BPTT). The trainable parameters of the model included $\boldsymbol{w}$, $\boldsymbol{w}_{noise}$, $\boldsymbol{\tau}$, $\boldsymbol{w}_{out}$, and $b$. 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 [31, 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 Kim et al. [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

$$\tau_i = \sigma(\mathcal{N}(0,1))\tau_{step} + \tau_{min},$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathcal{N}(0,1)$ refers to the standard normal distribution. $\tau_{min} =$ 20 ms and $\tau_{step} = 105$ 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 *Supplementary Information*.

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

**Delay match-to-sample (DMS) task.** Two 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 (Figure 1a). If the signs of the two sequential stimuli matched (i.e., stimulus condition 1: $s = (+1/+1)$; stimulus condition 4: $s = (-1/-1)$; Figure 3a), 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: $s = (-1/+1)$; Figure 3a), the model had to produce an output signal approaching -1. For the first task, the model had to respond immediately after the second stimulus (Figure 1c). A second delay period of 250 ms was added after the second stimulus for the second task (see *Supplementary Materials*). 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.

**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 $C \in \{0, 1, 5, 10\}$, 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 (Figure 1f) were systematically manipulated. During testing, internal noise and noisy input signals were introduced to the trained networks. For each successfully trained RNN, we generated $\boldsymbol{w}_{noise}$ 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 1250 ms (with a 500-ms increment) to assess the stability of memory maintenance (Figure 1f).

**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 two-alternative forced choice (AFC) 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 (Figure 6a). For the context-dependent sensory integration (CTX) task, the model received two streams of noisy stimulus signals (input modality 1 and input modality 2; (Figure 6c) 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 towards 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 AFC and CTX tasks, the training termination criteria were similar to those used for the DMS (see *Training protocol*).

**Visualization of network dynamics.** To visualize the neural dynamics of working memory computation as a function of injected internal noise during training, we employed the Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) algorithm [19]. 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 (see Figure 1c).

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; Figure 2b-c) was plotted and colored by stimulus conditions (Figure 2a). 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 internal noise during training.

**Network stability analysis.** 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 (see Figure 1c). Throughout this window, the network activities exhibited consistent steady-state patterns, as illustrated in *Supplementary Materials*.

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 Equation (1) in the absence of an input stimulus:

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

For a weak perturbation $\delta\boldsymbol{x}_{s_1}$ around $\boldsymbol{x}^*_{s_1}$, the linearized approximation of the perturbed dynamics is $\left.\frac{d\boldsymbol{x}}{dt}\right|_{\boldsymbol{x}^*_{s_1}+\delta\boldsymbol{x}_{s_1}} = \boldsymbol{F}(\boldsymbol{x}^*_{s_1}) + J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1} + O(\delta\boldsymbol{x}^2_{s_1})$, where $J(\boldsymbol{x}^*_{s_1})$ is the Jacobian matrix $J_{ij}(\boldsymbol{x}^*_{s_1}) = \left.\frac{\partial F_i}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^*_{s_1}}$. By the assumption of the late-time 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

$$\frac{d\delta\boldsymbol{x}_{s_1}}{dt} \approx J(\boldsymbol{x}^*_{s_1})\delta\boldsymbol{x}_{s_1}, \tag{4}$$

with the Jacobian matrix written explicitly as

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

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 (5).

For clarity, 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 (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

$$\delta\boldsymbol{x}(t) = \sum_{n=1}^{N} e^{\lambda_n t}\boldsymbol{\psi}^R_n\left(\boldsymbol{\psi}^L_n \delta\boldsymbol{x}_0\right), \tag{10}$$

where $\boldsymbol{\psi}_n^L$ and $\boldsymbol{\psi}_n^R$ are, respectively, the left and the right eigenvector of $J$ with the eigenvalue $\lambda_n$. Notably, our trained RNNs exhibit highly asymmetric $\boldsymbol{w}$ such that the Jacobian (5) is non-hermitian, leading to distinct left and right eigenvectors.

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

Since the dominant responses to a perturbation depend on the overlap between the perturbation and the top-most left eigenvectors $\left(\boldsymbol{\psi}_n^L \delta \boldsymbol{x}_0\right)$, 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 late-time steady states.

We employ the Inverse Participation Ratio (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,

$$\mathrm{IPR}(\lambda_n) = \frac{\sum_{i=1}^{N} |(\boldsymbol{\psi}_n)_i|^4}{\left(\sum_{i=1}^{N} |(\boldsymbol{\psi}_n)_i|^2\right)^2}. \tag{11}$$

The IPR of the left and the right eigenvector will be denoted by $\mathrm{IPR}_L$ and $\mathrm{IPR}_R$ respectively, though we will focus on $\mathrm{IPR}_L$ as we are interested in the size of the neural subpopulations participating in the perturbation. Note that the maximum and the minimum values of $\mathrm{IPR}_L$ 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 $\mathrm{IPR}_L$ indicates that the perturbation is localized around a smaller number of units, or extended over a larger number of units, respectively.

**Perturbation analysis.** For the example models shown in Figure 2, we first performed the network stability analysis described above. We then ranked the eigenvalues ($\boldsymbol{\lambda}$) based on their real values and identified the corresponding left eigenvectors ($\boldsymbol{\psi}_n^L$) for the top ten eigenvalues. For each of the top ten eigenvalues, we also computed the associated $\mathrm{IPR}_L$ (see *Supplementary Materials*). Next, we perturbed the set of units that contributed to each of the ten left eigenvectors during the response window to assess the network's sensitivity to perturbation (see *Supplementary Materials*).

For each of the ten perturbations, the network's task performance was computed (average task performance shown in *Supplementary Materials*). To determine the task performance per

$IPR_L$ (PIPR), we divided the IPR values by the corresponding perturbed task performance (see *Supplementary Materials*).

**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 two-sided Wilcoxon rank-sum or signed-rank test. For comparing morethan two groups (e.g., the synaptic decay time constants associated with RNNs trained with varying degree of noise), we used Kruskal-Wallis test with Dunn's post hoc test to correct for multiple comparisons.

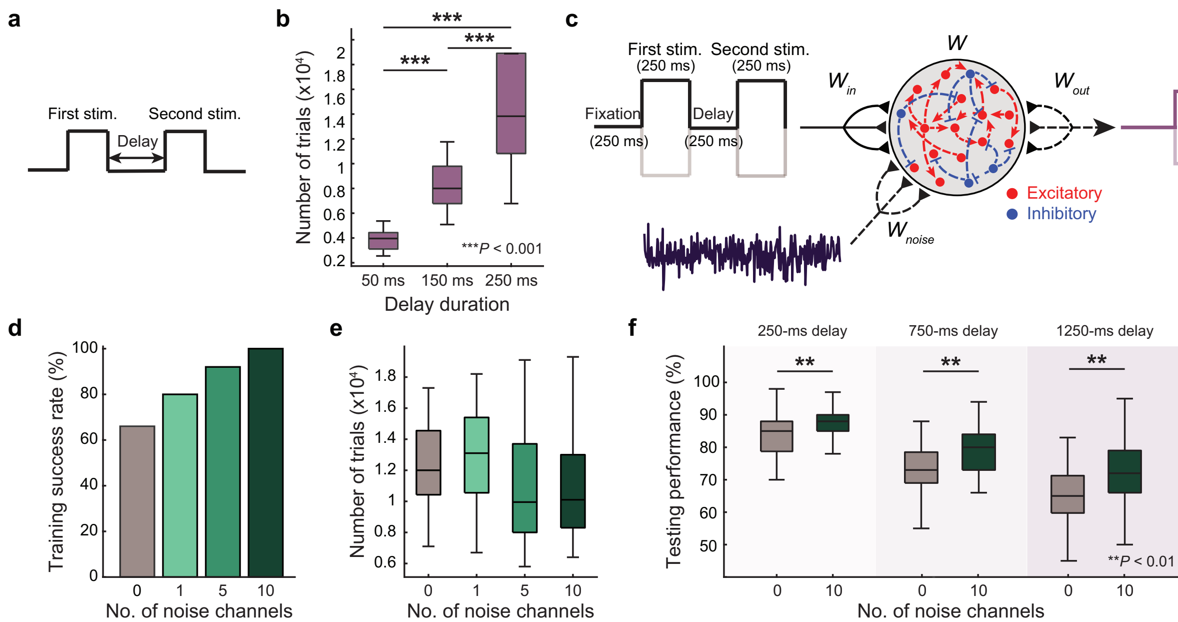


**Fig. 1 | Delayed match-to-sample (DMS) task and model schematic.** **a**, A schematic diagram of a Delayed match-to-sample (DMS) task with two sequential stimuli separated by a delay interval. **b**, The number of trials/epochs needed to train continuous-variable RNNs increases exponentially as the 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 $P$s $< 0.001$, two-sided Wilcoxon rank-sum test). **c**, A schematic diagram illustrates the paradigm used to trained 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 network 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). 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 internal noise and noisy input signals were introduced. We also varied the delay duration of these testing trials to range from 250 ms, 750 ms, and 1250 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 $P$s $< 0.01$, 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.

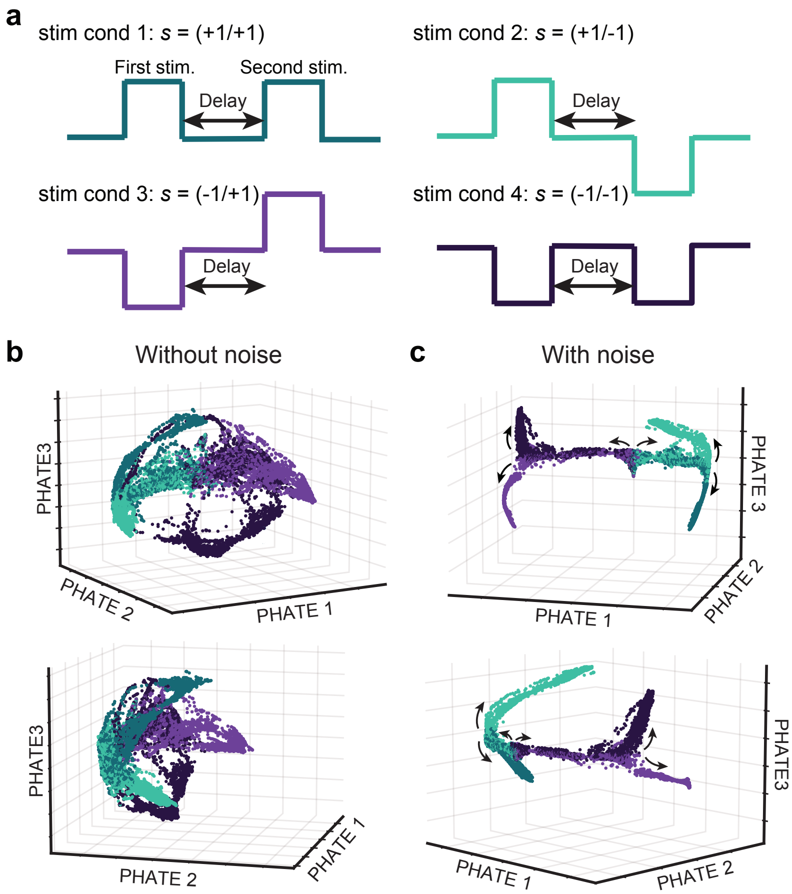


**Fig. 2 | Neural representations of each stimulus condition on the DMS task. a**, A schematic of the four stimulus conditions used in the delayed match-to-sample (DMS) task. For stimulus condition 1 ($s = +1/+1$) and 4 ($s = -1/-1$), the model had to generate an output signal approaching $+1$. For stimulus condition 2 ($s = +1/-1$) and 3 ($s = -1/+1$), the model had to produce an output signal approaching -1. **b**, PHATE-embedding computed from network activity on testing trials (see *Methods*) of an example RNN model trained without noise. The embedding based on network activity from the onset of the first stimulus is plotted. **c**, PHATE-embedding extracted from the network activity during testing of a sample RNN model that was trained with noise ($C = 10$). The embedding based on network activity from the onset of the first stimulus is plotted. Black arrows indicate temporal progression of the PHATE trajectories over the trial duration. Trajectories within the PHATE-embedding are illustrated based on the stimulus conditions from which the data were extracted. While task-based clusters can be clearly observed in the PHATE-embedding of the RNN model trained with noise (**c**), such patterns are not present in the embedding of the model trained without noise (**b**). Importantly, the task-informed clustering associated with the model trained with noise exhibits temporal dynamics that are tightly linked to the onsets of the first and second stimulus such that the first and second branching emerged at the presentation onset of the first and second stimulus, respectively.

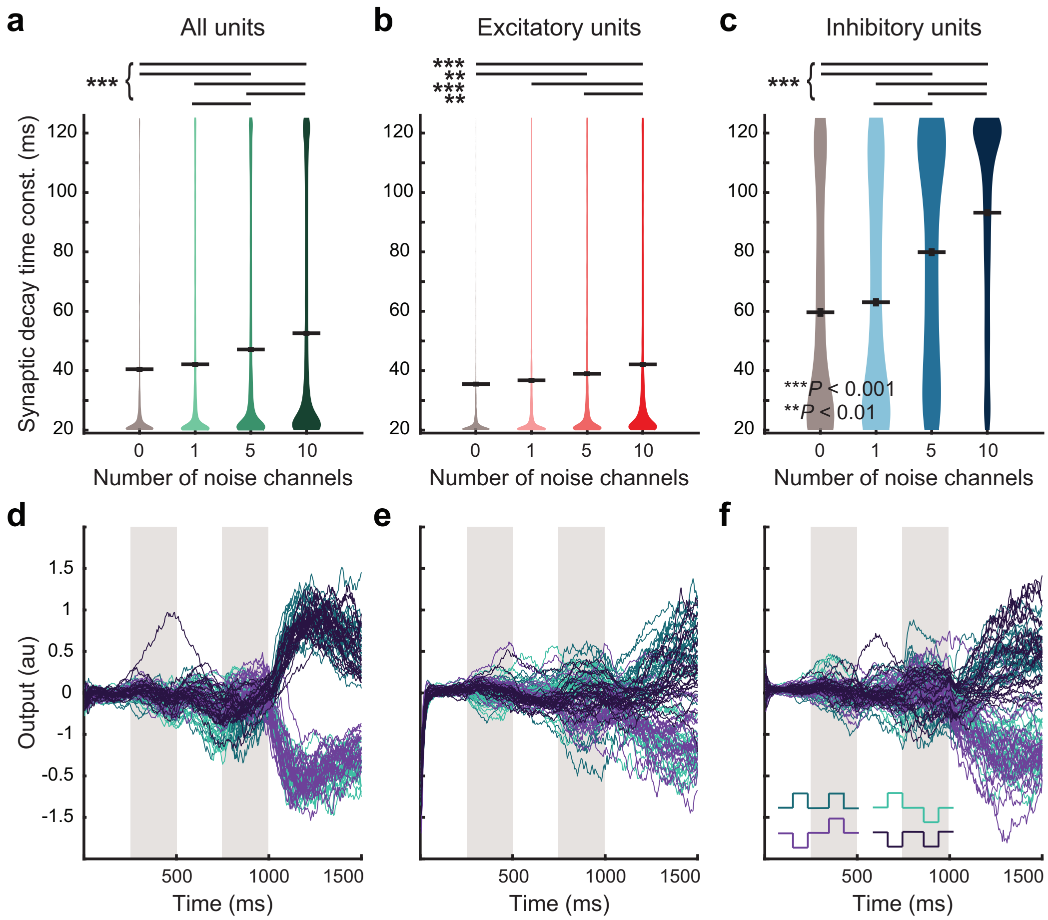


**Fig. 3 | Influence of noise on cell-type specific temporal dynamics.** Comparison of synaptic decay time constants of RNN models trained on the DMS task with varying amount of noise. **a**, For each noise condition, synaptic decay time constants of successfully trained models are reported for all units (n = 33, 40, 46, 50 for the noise conditions of 0, 1, 5, and 10 channels, respectively). Overall, injection of random noise during training increased synaptic decay time constants averaged across all units in the networks ($P$s $<$ 0.001, $H = 113.8$; Kruskal-Wallis test with Dunn's post hoc test). **b**, Comparison of synaptic decay time constants for excitatory units of the trained RNN models ($P$s $< 0.01$, $H = 52.5$; Kruskal-Wallis test with Dunn's post hoc test). **c**, Comparison of synaptic decay time constants for inhibitory units of the trained RNN models ($P$s $< 0.001$, $H = 120.3$; Kruskal-Wallis test with Dunn's post hoc test). Gray horizontal lines, mean. **d**, Network output of a sample RNN model successfully trained without noise to perform the DMS task. The model can differentiate among the four possible stimulus conditions and generate appropriate responses based on the maintained memory (+1 when $s = +1/+1$ (dark green) or $-1/-1$ (dark purple) and -1 when $s = +1/-1$ (light green) or $-1/+1$ (light purple)). **e**, Network output of a RNN model trained without noise where synaptic decay time constants of all units were set to 125 ms (maximal $\tau$; see *Methods*). The model failed to maintain memory and generate correct responses. **f**, Network output of a RNN model trained without noise where synaptic decay time constants of inhibitory units were fixed at 125 ms. The overall performance is higher than that of (**b**), further confirming the differential effect of noise on inhibitory circuits.

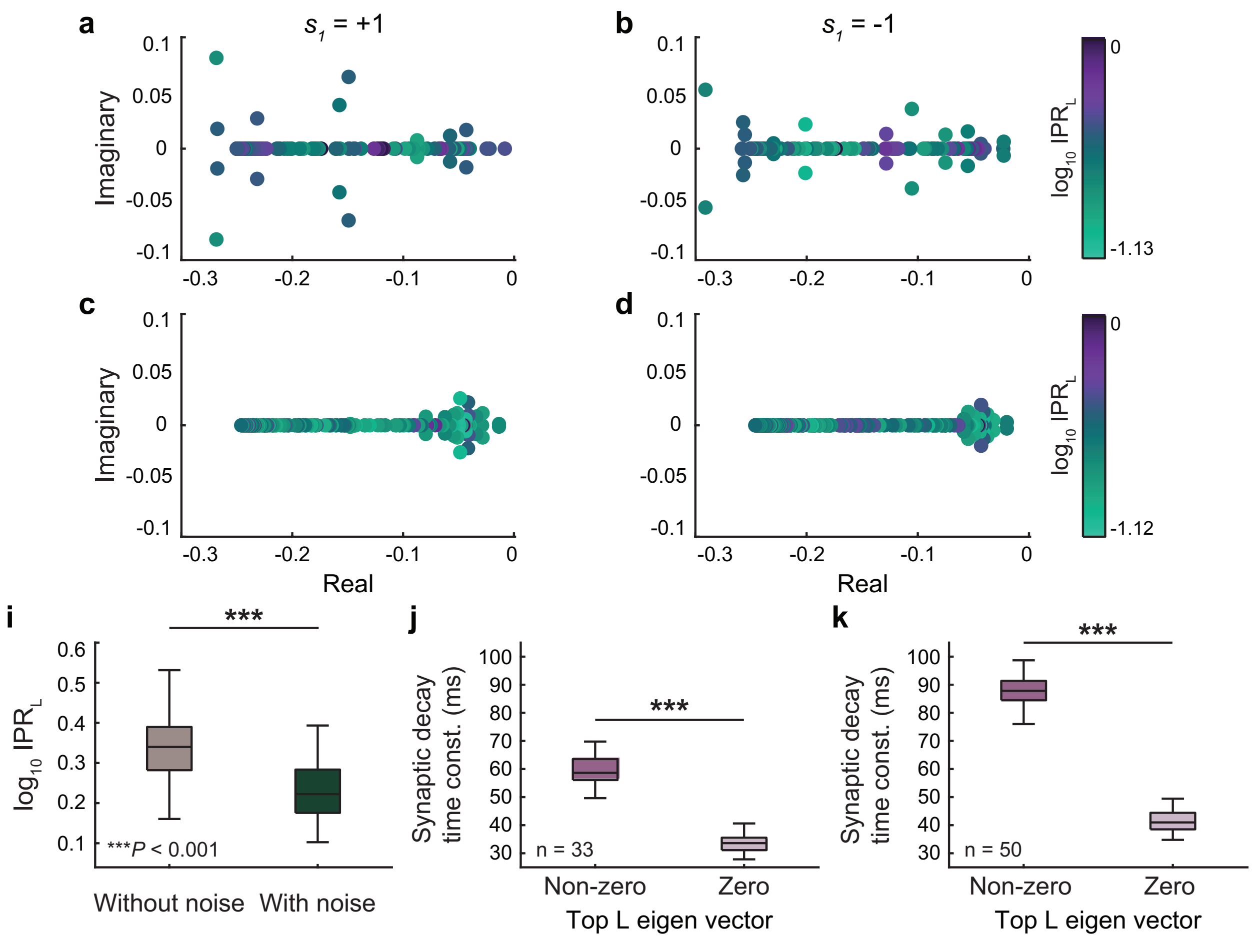


**Fig. 4 | Noise-induced network spectral properties**. Spectra of the Jacobian ($J$) extracted from the network activity during the delay window. **a** and **b**, Spectra of a sample RNN model trained without noise (same RNN as Figure 2**b**) during the delay period following the first stimulus presentation ($s_1 \in \{+1, -1\}$). **c** and **d**, spectra of a sample RNN model trained with noise ($C = 10$; same network as Figure 2**c**) during the delay period following the first stimulus presentation ($s_1 \in \{+1, -1\}$). For both noise 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 non-zero 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**). **i**, Average IPR values from the RNN trained without noise were significantly higher (i.e., more localized) than those from the model trained with noise. **j**, Average synaptic decay time constants of the dominant (non-zero elements in the top ten eigenvectors) and non-dominant (zero elements in the top ten eigenvectors) units from all the RNNs trained without noise. **k**, Average synaptic decay time constants of the dominant and non-dominant units from all the RNNs trained with noise. Boxplot: central lines, median; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers are not plotted. $P < 0.001$, two-sided Wilcoxon rank-sum test.

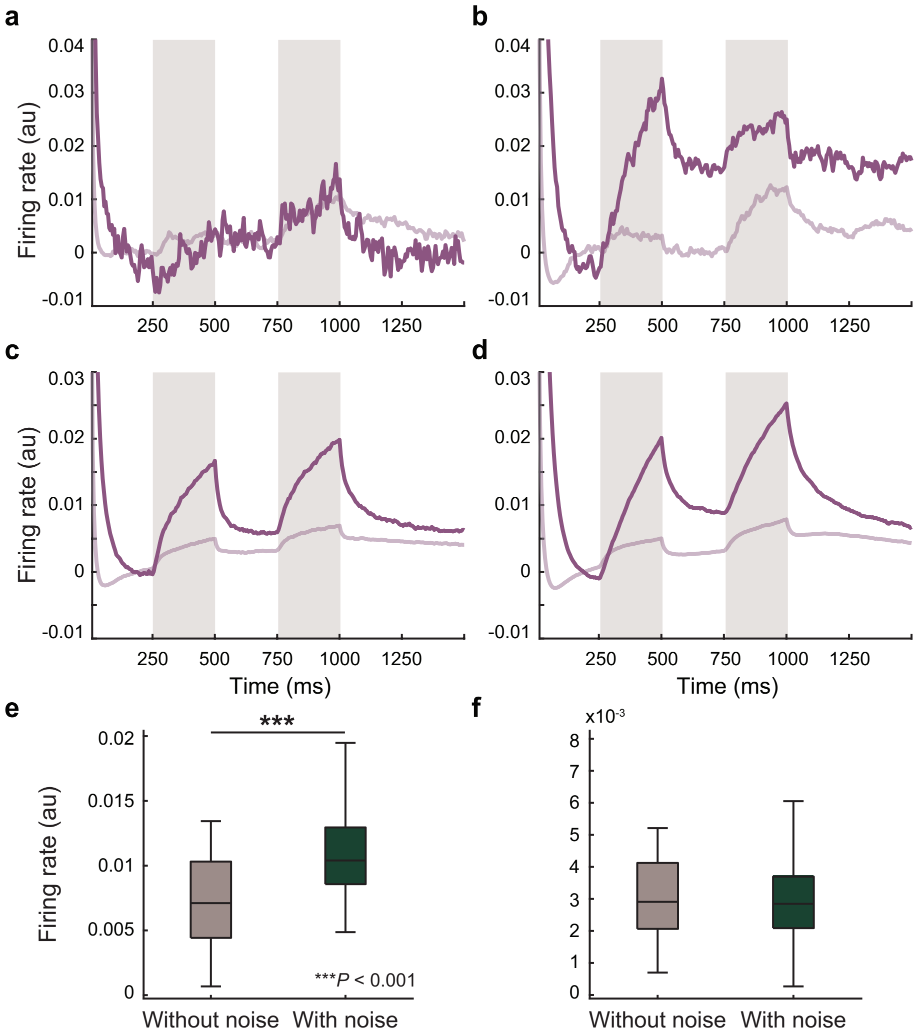


**Fig. 5 | Persistent activity of dominant units from RNNs trained with noise.** **a**, Average firing rate timecourses for the dominant (dark purple) and non-dominant (light purple) units from the example RNN model trained without noise (same RNN as the one used for Figure 2**b**). **b**, Average firing rate timecourses for the dominant (dark purple) and non-dominant (light purple) units from the example RNN model trained with noise (same RNN as the one used for Figure 2**c**). **c**, Similar to **a** but averaged across all RNNs successfully trained ($n$ = 33 RNNs). **d**, Similar to **b** but averaged across all RNNs successfully trained ($n$ = 50 RNNs). **e**, Average firing rate activity during the delay period for the dominant units from RNNs trained without noise (gray) and with noise (dark green). **f**, Average firing rate activity during the delay period for the non-dominant units from RNNs trained without noise (gray) and with noise (dark green). Boxplot: central lines, median; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers are not plotted. $P < 0.001$, two-sided Wilcoxon rank-sum test.

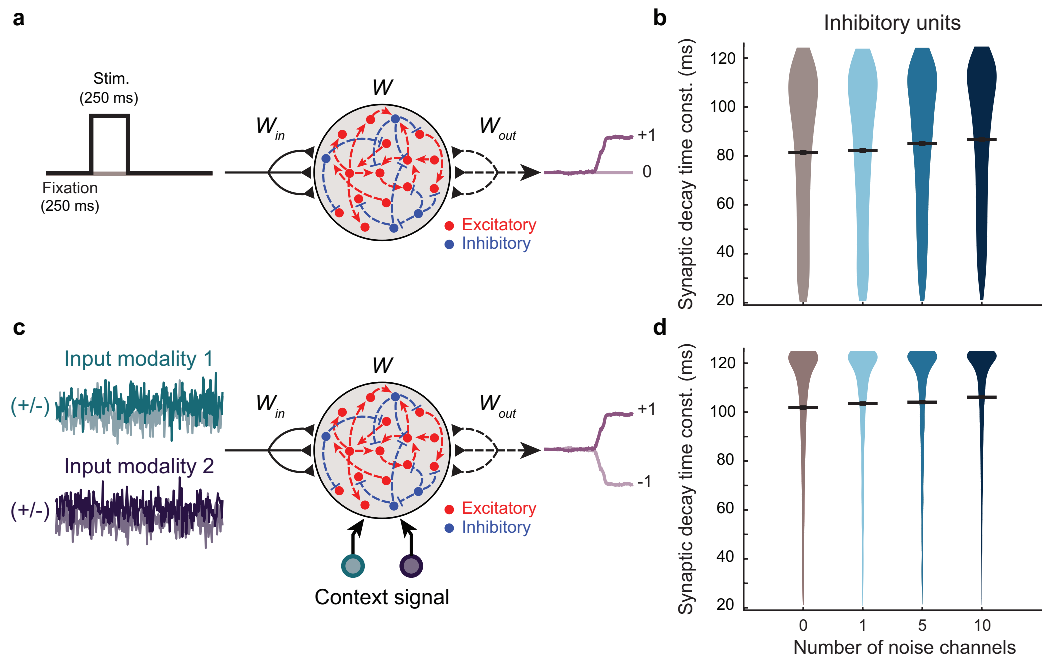


**Fig. 6 | Network functional motifs underlying working memory-independent computation.** Schematics diagrams illustrating working memory-independent tasks and the corresponding network dynamics of the RNN models successfully trained on these tasks. **a**, Two-alternative forced choice (AFC) task, in which the RNN modes were trained to produce an output indicating the presence of a brief input pulse. **b**, For the inhibitory units from the RNNs trained on the AFC task, synaptic decay time constants were similar across all noise conditions. **c**, 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. **d**, Across all the noise conditions, the inhibitory units from the networks trained on the sensory integration task exhibited similar synaptic decay time constants. **e**, For the CTX task, similar PIPR was observed for a sample RNN model trained without and with noise ($C = 10$). **f**, Task performance on the CTX task after perturbation was similar regardless of whether intrinsic noise was introduced during training. Gray horizontal lines, mean.

## References

[1] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019.

[2] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364 (6437):eaav7893, 2019.

[3] Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778, 2021.

[4] Sophie JC Caron, Vanessa Ruta, LF Abbott, and Richard Axel. Random convergence of olfactory inputs in the drosophila mushroom body. *Nature*, 497(7447):113–117, 2013.

[5] Adji B Dieng, Jaan Altosaar, Rajesh Ranganath, and David M Blei. Noise-based regularizers for recurrent neural networks. 2018.

[6] Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J Roberts, and Chris C Holmes. Explicit regularisation in gaussian noise injections. *Advances in Neural Information Processing Systems*, 33: 16603–16614, 2020.

[7] Soon Hoe Lim, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Noisy recurrent neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[9] Sabine Krabbe, Enrica Paradiso, Simon d' Aquin, Yael Bitterman, Julien Courtin, Chun Xu, Keisuke Yonehara, Milica Markovic, Christian Müller, Tobias Eichlisberger, and et al. Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nature Neuroscience*, 22(11):1834–1843, Oct 2019.

[10] Kirstie A. Cummings and Roger L. Clem. Prefrontal somatostatin interneurons encode fear memory. *Nature Neuroscience*, 23(1):61–74, 2019.

[11] Gianluigi Mongillo, Simon Rumpel, and Yonatan Loewenstein. Inhibitory connectivity defines the realm of excitatory plasticity. *Nature Neuroscience*, 21(10):1463–1470, Sep 2018.

[12] Haifeng Xu, Ling Liu, Yuanyuan Tian, Jun Wang, Jie Li, Junqiang Zheng, Hongfei Zhao, Miao He, Tian-Le Xu, Shumin Duan, and et al. A disinhibitory microcircuit mediates conditioned social fear in the prefrontal cortex. *Neuron*, 102(3):668–682, 2019.

[13] Robert Kim and Terrence J Sejnowski. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nature Neuroscience*, 24(1):129–139, 2021.

[14] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[15] H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2): e1004792, 2016.

[16] Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6:e20899, 2017.

[17] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.

[18] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[19] Kevin R Moon, David van Dijk, Zheng Wang, William Chen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, and Smita Krishnaswamy. Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *BioRxiv*, 120378, 2017.

[20] Chi Zhang, Danke Zhang, and Armen Stepanyants. Noise in neurons and synapses enables reliable associative memory storage in local cortical circuits. *eNeuro*, 8(1), 2021.

[21] Mark D McDonnell and Lawrence M Ward. The benefits of noise in neural systems: bridging theory and experiment. *Nature Reviews Neuroscience*, 12(7):415–425, 2011.

[22] Stan L. Pashkovski, Giuliano Iurilli, David Brann, Daniel Chicharro, Kristen Drummey, Kevin M. Franks, Stefano Panzeri, and Sandeep Robert Datta. Structure and flexibility in cortical representations of odour space. *Nature*, 583(7815):253–258, 2020.

[23] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and LF Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

[24] Xiaoxing Zhang, Wenjun Yan, Wenliang Wang, Hongmei Fan, Ruiqing Hou, Yulei Chen, Zhaoqin Chen, Chaofan Ge, Shumin Duan, Albert Compte, and Chengyu T Li. Active information maintenance in working memory by a sensory cortex. *eLife*, 8:e43191, jun 2019.

[25] David Sussillo and L.F. Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544 – 557, 2009.

[26] Wilten Nicola and Claudia Clopath. Supervised learning in spiking neural networks with force training. *Nature Communications*, 8:2208, Dec 2017.

[27] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Computational Biology*, 12 (2):1–30, 02 2016.

[28] Stewart H Hendry, HD Schwark, EG Jones, and J Yan. Numbers and proportions of gaba-immunoreactive neurons in different areas of monkey cerebral cortex. *Journal of Neuroscience*, 7(5):1503–1519, 1987.

[29] Arish Alreja, Ilya Nemenman, and Christopher J Rozell. Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices. *PLOS Computational Biology*, 18(1):e1009642, 2022.

[30] Chet C Sherwood, Mary Ann Raghanti, Cheryl D Stimpson, Christopher J Bonar, Alexandra A de Sousa, Todd M Preuss, and Patrick R Hof. Scaling of inhibitory interneurons in areas v1 and v2 of anthropoid primates as revealed by calcium-binding protein immunohistochemistry. *Brain, Behavior and Evolution*, 69(3):176–195, 2007.

[31] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Computational Biology*, 12 (2):e1004792, 2016.

[32] Robert Kim, Yinghao Li, and Terrence J Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the national academy of sciences*, 116(45):22811–22820, 2019.

[33] Renato Duarte, Alexander Seeholzer, Karl Zilles, and Abigail Morrison. Synaptic patterning and the timescales of cortical dynamics. *Current Opinion in Neurobiology*, 43:156–165, 2017.

[34] Anthony M Zador and Lynn E Dobrunz. Dynamic synapses in the cortex. *Neuron*, 19(1):1–4, 1997.

[35] Fernando Lucas Metz, Izaak Neri, and Tim Rogers. Spectral theory of sparse non-hermitian random matrices. *Journal of Physics A: Mathematical and Theoretical*, 52(43):434003, oct 2019. doi:10.1088/1751-8121/ab1ce0. URL `https://dx.doi.org/10.1088/1751-8121/ab1ce0`.

[36] Elihu Abrahams. *50 Years of Anderson Localization*. WORLD SCIENTIFIC, 2010. doi:10.1142/7663. URL `https://www.worldscientific.com/doi/abs/10.1142/7663`.

## Acknowledgements

This work was supported by the Swartz Foundation (N.R.), the Kavli Institute for Brain and Mind (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 DEVCOM Army Research Laboratory or the United States Government.

## Author information

**Department of Biomedical Engineering, Columbia University, New York, NY, USA**

Nuttida Rungratsameetaweemana

**Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, USA**

Nuttida Rungratsameetaweemana, Robert Kim & Terrence J. Sejnowski

**Neurology Department, Cedars-Sinai Medical Center, Los Angeles, CA, USA** Robert Kim

**Chula Intelligent and Complex Systems, Department of Physics, Chulalongkorn University, Bangkok, Thailand**

Thiparat Chotibut

**Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA**

Terrence J. Sejnowski

**Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA**

Terrence J. Sejnowski

## Contributions

N.R. and R.K. conceived, designed, and performed the research; N.R., R.K., and T.C. analyzed data; N.R., R.K., T.C., and T.J.S. wrote the manuscript.

## Corresponding author

Corresponding authors: correspondence to Thiparat Chotibut or Terrence J. Sejnowski

## Declaration of interests

The authors declare no competing interests.

## Code availability

The code for training the networks and for the analyses performed in this work will be made available at `https://github.com/NuttidaLab/Noisy_RNN`.

## Data availability

All data used in the present study will be deposited as MATLAB-formatted data in Open Science Framework, `https://osf.io/dqy3g/`.