Internal noise promotes heterogeneous synaptic dynamics important for working memory in recurrent neural networks

Nuttida Rungratsameetaweemana*
The Salk Institute for Biological Studies
La Jolla, CA 92037
nrungrat@salk.edu

Robert Kim*
The Salk Institute for Biological Studies
La Jolla, CA 92037
robert.f.kim@gmail.com

Terrence J. Sejnowski
The Salk Institute for Biological Studies
La Jolla, CA 92037
terry@salk.edu

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 sub-population of inhibitory units.

1 Introduction

Even though recent advances in deep learning and artificial intelligence (AI) have greatly increased the functionality and capability of artificial neural networks 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 (Equation (1)) have been widely used to reproduce previously observed experimental findings and to explore neural dynamics associated with cognitive functions including working memory (ability to maintain information for a brief period) [Mante et al., 2013, Song et al., 2016a, Miconi, 2017, Yang et al., 2019].

$$\tau \frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} \phi(x_j) w_{ij} + u \tag{1}$$

The above equation describes a network of recurrently connected N units where x_i represents the synaptic current variable for unit i, $\phi(\cdot)$ is a nonlinear activation function used to estimate the firing

Preprint. Under review.

rate from the synaptic variable, w_{ij} is the synaptic strength from unit j to unit i, and u is the task-specific input data given to the network. τ refers to the synaptic decay time constant, which is often fixed to a large, constant value (usually 100 ms) for all the units.

As the units in these network models communicate via differentiable and continuous signals, gradient-descent methods, such as backpropagation through time (BPTT; [Werbos, 1990]), can be used to train the models to perform tasks commonly used in neuroscience. Previous studies have trained the RNN model (Equation (1)) to perform cognitive tasks that involve a wide range of cognitive computations ranging from simple decision making to context-dependent evidence accumulation ([Yang et al., 2019, Mante et al., 2013]). Despite the success of the model on these various tasks, training the RNN via backpropagation to carry out working memory tasks remains a challenge. For example, the model can learn a simple temporal exclusive-OR (XOR) task only when the delay interval between the two input stimuli is sufficiently short (Figure 1, [Jarne and Laje, 2019]). The training time increases exponentially as a function of the delay duration (Figure 1b). In addition, the training does not converge if the synaptic decay time constant (τ) is fixed to a small constant (i.e., faster decay rate).

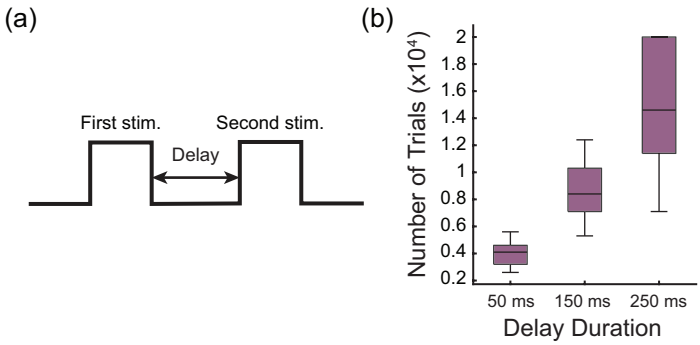


Figure 1: (a) A schematic diagram of a temporal XOR task with two sequential stimuli separated by a delay interval. (b). The number of trials/epochs needed to train continuous-variable RNNs increases exponentially as the delay interval increases. For each delay duration condition, we trained 50 firing-rate RNNs to perform the XOR task shown in (a). The maximum number of trials/epochs was set to 20,000 trials for computational efficiency. Boxplot: central lines, median; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers are not plotted.

In this work, we present a simple framework to produce robust RNNs capable of learning temporal XOR tasks with heterogeneous synaptic time constants. We achieve this by introducing random, internal noise to the RNN model during training. By varying the structure of the random noise, we identified the optimal noise structure that allows the model to learn even a difficult variation of the temporal XOR task. We also demonstrate that adding the intrinsic noise encourages the network to increase the synaptic time constants of the inhibitory units. These findings are aligned with recent experimental and theoretical studies that place specific subtypes of inhibitory neurons at the center of working memory computations [Krabbe et al., 2019, Cummings and Clem, 2019, Mongillo et al., 2018, Xu et al., 2019, Kim and Sejnowski, 2021]. Therefore, our study illustrates how seemingly random noise in the cortex could lead to diverse synaptic dynamics critical for complex cognitive functioning.

2 Related Work

Injecting noise into artificial neural networks. Recent studies have shown that injecting Gaussian noise into multi-layer perceptron (MLP) and vanilla recurrent neural networks during training can lead to improved stability and robustness [Dieng et al., 2018, Camuto et al., 2020, Lim et al., 2021]. For example, Lim et al. [2021] studied the effects of injecting noise into the hidden states of vanilla RNNs and showed that the noise provided stochastic stabilization via implicit regularization [Blanc et al., 2020]. To our knowledge, no studies have investigated the effects of noise injections into RNNs of firing-rate model neurons (Equation (1)).

Cortical noise. It is well documented that there exists a high level of spontaneous ongoing activities in the cortex that are seemingly orthogonal to task-dependent neural codes and activities. However, this 'cortical noise' has been recently shown to contain information about the environmental context and is directly related to downstream behavioral outcomes [Musall et al., 2019, Stringer et al., 2019, 2021]. The relevance and importance of cortical noise to cognitive functions have been elucidated through recent studies showing that random structures of the olfactory system of *Drosophila* optimized the diversity of odor representations in the neural circuits [Caron et al., 2013].

3 Methods

3.1 Continuous-rate recurrent neural networks

The recurrent neural network (RNN) model we used in our study is similar to Equation (1). All the units in a network are governed by the following equations:

$$\tau_i \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} w_{ij} r_j(t) + \boldsymbol{w}_{noise}\boldsymbol{\psi}(t) + \boldsymbol{w}_{in}\boldsymbol{u}(t) + \boldsymbol{\xi}(t) \tag{2}$$

$$r_i(t) = \sigma(x_i(t)) = \frac{1}{1+\exp(-x_i(t))} \tag{3}$$

$$o(t) = \boldsymbol{w}_{out}\boldsymbol{r}(t) + b \tag{4}$$

where 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 Equation (3)). Each model contains 200 units (80% excitatory and 20% inhibitory units; i.e., E-I ratio of 80/20). The model receives time-varying input signals ($\boldsymbol{u} \in \mathbb{R}^{1\times T}$) via the input weight matrix, $\boldsymbol{w}_{in} \in \mathbb{R}^{N\times 1}$. The network also receives "intrinsic" noisy signals 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 with zero mean and variance of 0.01. The output (o) of the network was computed as an 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 using the step size (Δt) of 5 ms:

$$\boldsymbol{x}_t = \left(1 - \frac{\Delta t}{\boldsymbol{\tau}}\right)\boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau}}(\boldsymbol{w}\boldsymbol{r}_{t-1} + \boldsymbol{w}_{noise}\boldsymbol{\psi}_{t-1} + \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) + \boldsymbol{\xi}_{t-1} \tag{5}$$

where $\boldsymbol{x}_t = \boldsymbol{x}(t)$. The network was trained using backpropagation through time (BPTT). The parameters of the model included $\boldsymbol{w}$, $\boldsymbol{w}_{noise}$, $\boldsymbol{\tau}$, $\boldsymbol{w}_{out}$, and b. To make the model more biologically plausible, we incorporated Dale's principle (separate populations for excitatory and inhibitory neurons) using the method previously described in [Song et al., 2016b].

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 the method described in Kim et al. [2019]. The parameter was trained to range from 20 ms to 125 ms to model heterogeneous synaptic dynamics of different receptors in the cortex [Duarte et al., 2017, Zador and Dobrunz, 1997]. We initialized the synaptic decay time constant parameter ($\boldsymbol{\tau}$) using

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

where $\sigma(\cdot)$ is the sigmoid function and $\mathcal{N}(0,1)$ refers to the standard normal Gaussian 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 the supplementary materials.

The schematic diagram of the model is shown in Figure 2. All the models were implemented with TensorFlow 1.10.0 and trained on NVIDIA GPUs (Quadro P4000 and Quadro RTX 4000).

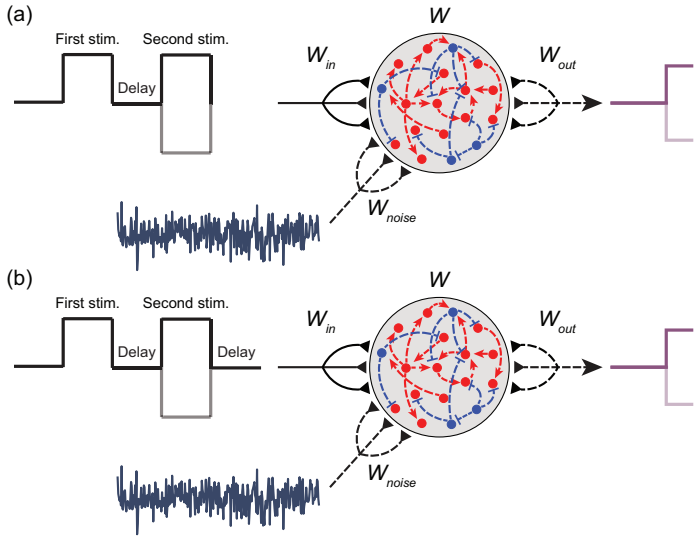


Figure 2: Schematic diagrams illustrating the paradigms used to train our RNN model on two variants of the temporal exclusive OR (XOR) tasks in which either one (a) or two delays (b) were present. In both cases, we introduced and systematically varied the amount of internal noise in the RNN network to study the effects of internal 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.

3.2 Temporal exclusive OR (XOR) task

Two temporal exclusive OR (XOR) tasks were used to train our RNN model and assess how the internal noise influenced the robustness of memory maintenance. Both tasks involved two sequential stimuli (each stimulus lasting for 250 ms) separated by a delay interval of 250 ms. During the stimulus window, the input signal ($\boldsymbol{u}$) was set to either -1 or +1 (Figure 2). If the signs of the two sequential stimuli match, the model was trained to produce an output signal approaching +1. When the signs were opposite, the model had to produce a signal approaching -1. For the first task, the model had to respond immediately after the second stimulus (Figure 2a). A second delay period of 250 ms was added after the second stimulus for the second task (Figure 2b). Due to two delay periods, the second temporal XOR task is considered a more challenging working memory task than the first task.

3.3 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 $> 95\%$

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 internal noise) as the baseline model.

3.4 Testing protocol

To evaluate for robustness and stability of the trained RNNs, we devised a series of testing conditions where different aspects of the one-delay XOR task (Figure 2a) were systematically manipulated. During testing, internal and/or external noise signals were introduced to the networks. For the internal noise condition, we generated a random matrix from the standard normal Gaussian distribution (similar to $\boldsymbol{w}_{noise}$) to feed ten independent white-noise signals during testing. For the networks with

$C = 0$, the noise term similar to $\boldsymbol{w}_{noise}\psi$ in Equation (5) was added during this testing environment. For the models with $C \in \{1, 5, 10\}$, $\boldsymbol{w}_{noise}$ and ψ were replaced with random Gaussian matrices to feed ten independent noise signals.

For the external noise testing environment, white-noise signals (drawn from the standard normal distribution) were added to the sensory signals ($\boldsymbol{u}$) to mimic stimulus-related noise. With these manipulations, we had four different testing conditions: (1) no noise, (2) external noise only, (3) internal noise only, and (4) both internal and external noises. For each of these testing conditions, we also varied the duration of the delay interval to range from 250 ms to 1250 ms (with a 250-ms increment) to assess the stability of memory maintenance.

4 Results

4.1 Internal noise improves training efficiency for working memory tasks

As shown in Figure 3a, the training success rate for the baseline model (i.e., no internal noise; $C = 0$) was 66% (33 out of 50 RNNs) for the temporal XOR task with one delay interval (Figure 2a). As the number of the noise channels (C) increased, the training success rate also increased. When $C = 10$, all 50 RNNs were successfully trained to perform the task (dark green in Figure 3a). For the networks successfully trained, we did not see any significant difference in the number of trials/epochs required among the four different noise conditions ($C \in \{0, 1, 5, 10\}$; Figure 3b).

We observed a similar trend for the temporal XOR task with two delay intervals. As the task is more challenging than the one-delay XOR task, the training success rate was overall much lower than the rate for the first task (Figure 3c). However, adding internal noise channels still significantly improved the training success rate (3 out of 50 RNNs with $C = 0$ vs. 16 out of 50 RNNs with $C = 10$; Figure 3c). As expected, it took longer to train our model for this task than for the previous task (Figure 3d).

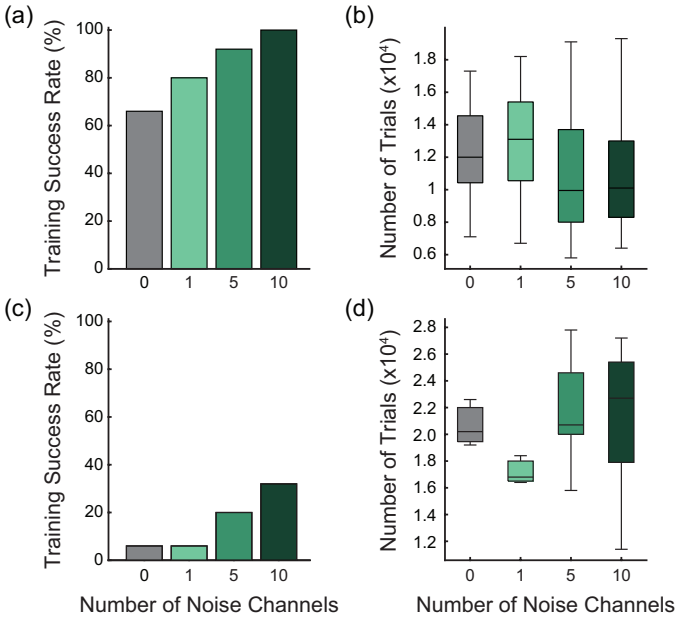


Figure 3: Training performance of the RNN models on the XOR task with one (top) or two delay intervals (bottom). RNN models with varying amount of internal noise (i.e., 0, 1, 5, and 10 noise channels) were trained to perform each variant of this task. Training success rate (a and c) was measured as the number of successfully trained RNNs (out of 50 RNNs). The average number of trials required to reach the training criteria is shown in (b) and (d). Boxplot: central lines, median; bottom and top edges, lower and upper quartiles; whiskers, 1.5 × interquartile range; outliers are not plotted.

4.2 Internal noise induces slow synaptic dynamics in inhibitory units

We studied our RNNs to see how the internal noise led to efficient training. Given the previous studies highlighting the importance of inhibitory connections for information maintenance [Mongillo et al., 2018, Krabbe et al., 2019, Xu et al., 2019, Kim and Sejnowski, 2021], we first compared the inhibitory recurrent connection weights of the RNNs from the four noise conditions. We did not observe any significant difference in the inhibitory weights among the four conditions (see supplementary materials). The excitatory recurrent weights were also comparable among the four models (see supplementary materials).

As we did not see any noticeable changes in the recurrent weight structure induced by the intrinsic 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) for the RNNs trained with internal noise for both tasks (Figure 4a and d). 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 4c and f). In addition, the extent of the shift was correlated with the number of internal noise channels: as C increased, the inhibitory synaptic signals decayed slower. For the excitatory population, we did not observe any marked changes in the decay time constant (Figure 4b and e). These findings are in line with recent modeling studies that emphasized the importance of slow inhibitory dynamics in maintaining information [Kim and Sejnowski, 2021].

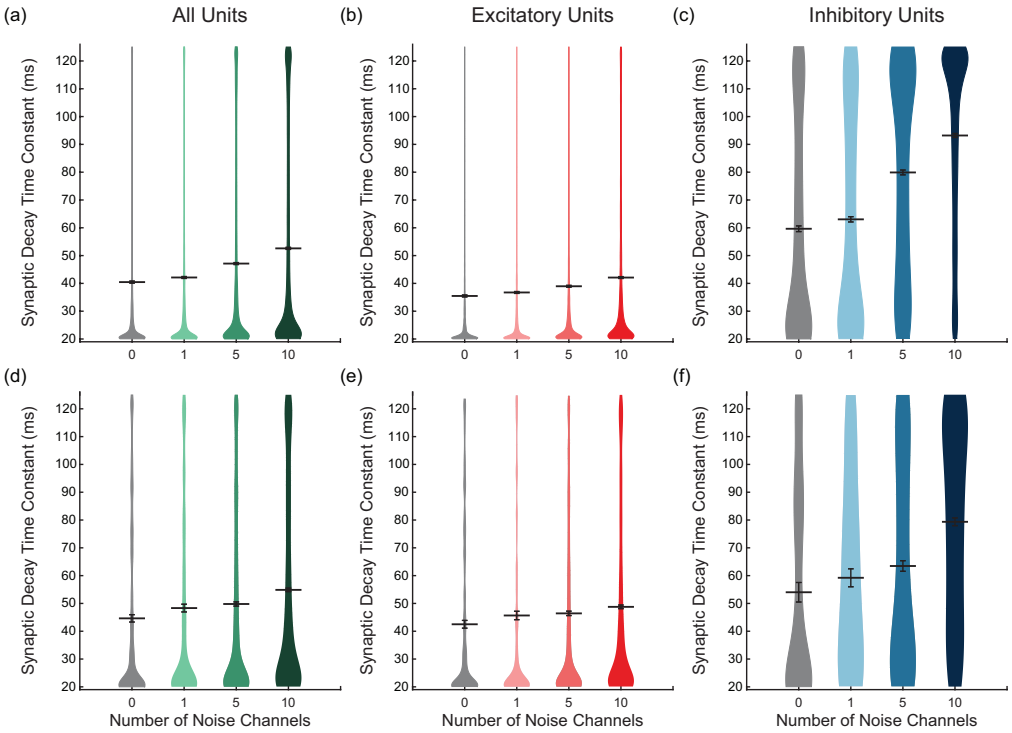


Figure 4: Comparison of synaptic decay time constants of RNN models trained on the XOR task with one (top) or two delay intervals (bottom). For each of these tasks and for each noise condition, synaptic decay time constants of successfully trained models are reported for all units (a and d), and separately for excitatory (b and e) and inhibitory units (c and f). Gray horizontal lines, mean.

4.3 RNNs trained with internal noise are robust to intrinsic and extrinsic perturbations

Our next analyses focused on unraveling the role of slow inhibitory synaptic decay dynamics in robustness and generalizability of the model. We hypothesized that injected internal noise during training contextualized sensory encoding and resulted in a more robust representation of the input stimuli. In other words, we predicted that the stability and robustness of the trained RNN models

would vary depending on the noise condition they were trained in. Specifically, the sensory input representation in the noise-injected models should be more robust, allowing these models to generalize in a new environment. To test this hypothesis, we evaluated the successfully trained RNNs on four different testing conditions and different delay intervals as described in Section 3.4.

Table 1: Overall testing performance. Average accuracy of the trained RNN models on the one-delay XOR task in four testing conditions (no noise, external noise only, internal noise only, both external and internal noise). We also varied the delay duration for each condition. Highest performance accuracy within each testing condition is boldfaced.

Delay	Testing conditions	C (# of noise channels)			
		0	1	5	10
250 ms	No noise	**96.52±2.94**	96.32±2.72	95.93±3.46	95.72±3.82
	Ext. noise only	92.09±3.74	92.10±4.24	**92.54±3.23**	92.28±3.96
	Int. noise only	86.27±5.16	84.28±7.21	88.48±5.65	**90.68±4.98**
	Ext. and int. noise	83.76±6.06	83.10±7.09	84.98±6.19	**88.02±3.87**
500 ms	No noise	94.58±4.51	93.57±3.71	94.85±3.51	**94.96±5.41**
	Ext. noise only	88.58±6.45	87.15±5.54	88.78±5.09	**89.26±5.55**
	Int. noise only	80.15±7.62	79.63±9.26	84.59±7.74	**87.54±4.73**
	Ext. and int. noise	78.27±7.10	78.10±9.18	82.09±7.26	**83.70±5.85**
750 ms	No noise	90.24±8.79	88.93±10.45	92.04±6.73	**92.66±6.89**
	Ext. noise only	83.21±10.96	82.33±9.68	84.65±8.21	**85.82±6.83**
	Int. noise only	76.36±9.82	74.63±9.62	80.65±8.99	**84.24±7.15**
	Ext. and int. noise	73.33±7.91	72.03±9.39	76.98±8.97	**79.22±6.60**
1000 ms	No noise	87.55±13.33	86.20±12.08	87.26±10.38	**89.74±10.85**
	Ext. noise only	78.91±12.64	77.50±11.83	80.41±10.85	**81.82±9.43**
	Int. noise only	70.06±9.77	69.92±10.88	76.48±10.37	**80.62±8.04**
	Ext. and int. noise	71.09±6.82	69.63±9.74	73.57±8.86	**75.84±7.87**
1250 ms	No noise	83.88±14.24	82.50±12.70	84.57±14.21	**87.22±12.73**
	Ext. noise only	75.39±13.71	73.88±13.16	77.09±12.33	**78.28±12.70**
	Int. noise only	67.91±9.03	68.60±10.66	73.39±11.74	**77.70±9.68**
	Ext. and int. noise	65.82±8.41	65.35±8.95	69.83±10.72	**71.98±10.30**

First, we found that the overall performance averaged across all the models was highest in the testing condition with no noise and lowest in the scenario where both external and internal noises were present (Table 1). Furthermore, the average performance was higher when only external noise was present than when only internal noise was present. In addition, the average performance across all the models declined as the duration of the delay interval increased. Importantly, we found that the models trained with the internal noise ($C \in \{1, 5, 10\}$) generally outperformed the baseline RNNs trained without any noise ($C = 0$), particularly under the most challenging testing conditions (i.e., extended delay intervals with both extrinsic and intrinsic noises). As shown in Table 1, the models trained with the noise condition of $C = 10$ had the highest task performance in every testing condition when the delay interval was greater than 250 ms. These results demonstrate that the internal noise introduced during training for the temporal XOR task increased the model stability and led to robust task representations that could be exploited in different environments. These learned dynamics are particularly useful for optimizing performance in a new task setting with higher task demands and sensory complexities.

4.4 Robustness and increased efficiency due to internal noise are specific to working memory tasks

Finally, we asked if the modulatory effects of the intrinsic noise during training were specific to working memory computations. To answer this question, we devised two cognitive tasks that do not require maintenance of sensory information over time. In the sensory detection (Go-NoGo) task (Figure 5a), the RNN model had to generate an output signal that indicated whether a target sensory signal was present. The context-dependent sensory integration task is a more challenging variant

of the sensory detection 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 [Mante et al., 2013]. 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 without any intrinsic noise, and that adding noise during training did not further improve training efficiency for either task. To investigate if intrinsic noise modulated the model 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 among the models that trained with no noise and those trained with internal noise. These findings suggest that the slow synaptic decay dynamics induced by internal noise is specific to working memory function where robust information maintenance is needed to ensure successful performance.

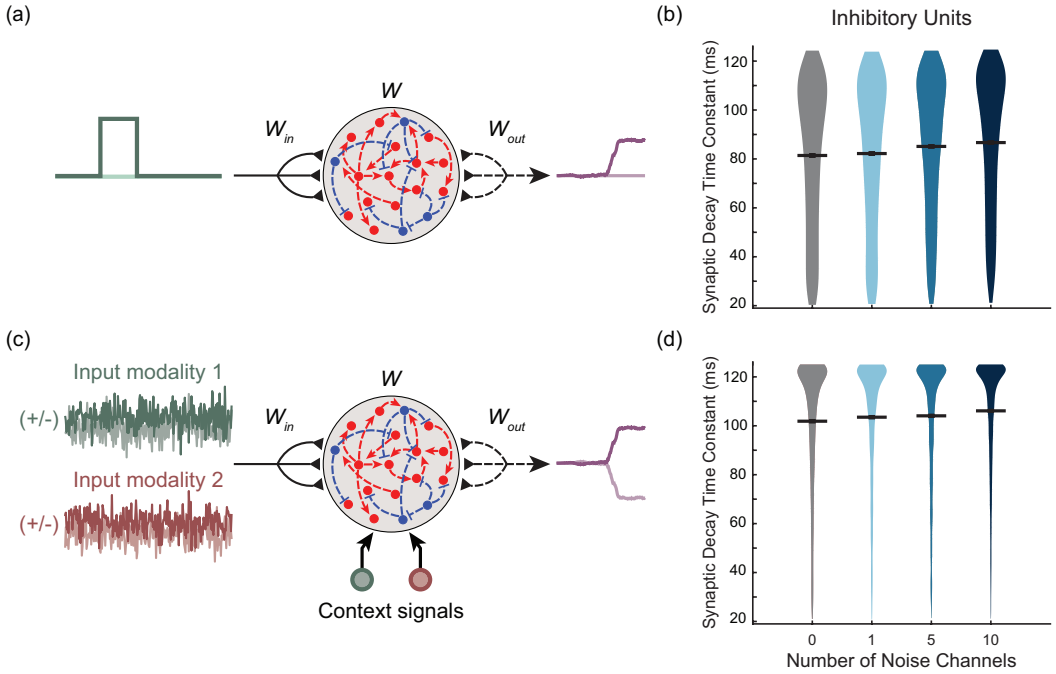


Figure 5: Schematics diagrams illustrating non-working memory tasks and the corresponding synaptic decay time constants for the inhibitory units of the RNN models successfully trained on these tasks. (a) Sensory detection (Go-NoGo) task, in which the RNN model was trained to produce an output indicating the presence of a brief input pulse. (b) For the inhibitory units from the RNNs trained on the Go-NoGo task, synaptic decay time constants were similar across all noise conditions. (c) Context-dependent sensory integration task, where the RNN model was 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. Gray horizontal lines, mean.

5 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 robust to perturbations and able to generalize to sustain stimulus-related information during a delay interval longer than the delay period they were trained. Further analyses revealed that the internal noise resulted in slow synaptic decay dynamics in inhibitory units. The changes were strongly correlated to the number of noise channels.

In addition, these changes were specific to the models trained to perform working memory tasks, suggesting that noise-induced stability was crucial for 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 [Zhang et al., 2021, McDonnell and Ward, 2011]. 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 [Caron et al., 2013, Pashkovski et al., 2020, Litwin-Kumar et al., 2017, Zhang et al., 2019]. 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 enhance 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.

6 Limitations

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 [Sussillo and Abbott, 2009, Nicola and Clopath, 2017]. 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 [Song et al., 2016c].

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

While we used non-working memory tasks (Go-NoGo and context-based sensory integration tasks) to illustrate that slow inhibitory synaptic decay dynamics were specific to working memory, it is possible that the non-working memory tasks were much less challenging than the temporal XOR tasks. Future directions include devising a non-working memory task that is as difficult as the working memory tasks. By doing so, we would be able to gain an improved understanding of the role of intrinsic noise in broader cognitive functions.

Broader Impact

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 effect on model dynamics and performance under different testing conditions. We believe that this framework offers a unique opportunity for researchers to gain a better understanding of how internal noise improves the training of machine learning models while also generating testable hypotheses about its function in the brain. 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 epilepsy, Parkinson's disease, schizophrenia, and autism spectrum disorder. We believe that the present study poses no negative ethical or societal issues.

References

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.

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, 2016a.

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

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.

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

C. Jarne and R. Laje. Graceful degradation of recurrent neural networks as a function of network size, memory length, and connectivity damage, 2019.

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.

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

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

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.

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.

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

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.

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

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.

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.

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.

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

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.

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, 2016b.

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.

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.

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

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.

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.

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.

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

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.

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

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

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 2016c.

Sahil Loomba, Jakob Straehle, Vijayan Gangadharan, Natalie Heike, Abdelrahman Khalifa, Alessandro Motta, Niansheng Ju, Meike Sievers, Jens Gempt, Hanno S Meyer, et al. Connectomic comparison of mouse and human cortex. *Science*, 377(6602):eabo0924, 2022.

A Gradient loss with respect to the synaptic decay time constant

In this section, we derive the gradient of the loss function with respect to the synaptic decay time constant parameter ($\boldsymbol{\tau}$). First, we define the loss function ($\mathcal{L}$) as

$$\mathcal{L} = \sum_{t=1}^{T} (z(t) - o(t))^2 \tag{6}$$

$$= \sum_{t=1}^{T} (z(t) - (\boldsymbol{w}_{out}\boldsymbol{r}(t) + b))^2 \tag{7}$$

We have the following gradient of the loss function at time t (i.e., $\mathcal{L}_t$) with respect to $\tau^{(i)}$ (i.e., synaptic decay time constant for unit i):

$$\frac{\partial \mathcal{L}_t}{\partial \tau^{(i)}} = \frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}} \frac{\partial r_t^{(i)}}{\partial x_t^{(i)}} \frac{\partial x_t^{(i)}}{\partial \tau^{(i)}} \tag{8}$$

For $\frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}}$, we can first rewrite the loss function,

$$\begin{aligned} \mathcal{L}_t &= z_t - \left(\sum_{j=1}^{N} w_{out}^{(j)} r_t^{(j)} + b \right)^2 \\ &= z_t - \left(w_{out}^{(i)} r_t^{(i)} + \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{out}^{(j)} r_t^{(j)} + b \right)^2 \end{aligned}$$

Then we have

$$\frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}} = (-2) \cdot w_{out}^{(i)} \cdot \left(w_{out}^{(i)} r_t^{(i)} + \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{out}^{(j)} r_t^{(j)} + b \right) \tag{9}$$

For $\frac{\partial r_t^{(i)}}{\partial x_t^{(i)}}$, we use Equation (3) to get

$$\frac{\partial r_t^{(i)}}{\partial x_t^{(i)}} = \sigma(x_t^{(i)}) \cdot (1 - \sigma(x_t^{(i)})) \tag{10}$$

Lastly, we substitute $\alpha = \frac{\Delta}{\tau^{(i)}}$ into the discretized equation (Equation (5)) for $\frac{\partial x_t^{(i)}}{\partial \tau^{(i)}}$

$$x_t^{(i)} = (1-\alpha) x_{t-1}^{(i)} + \alpha(\boldsymbol{w}\boldsymbol{r}_{t-1} + \boldsymbol{w}_{noise}\boldsymbol{\psi}_{t-1} + \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) + \xi_{t-1}^{(i)}$$

Differentiating the above equation with respect to α leads to

$$\begin{aligned} \frac{\partial x_t^i}{\partial \alpha} &= \frac{\Delta}{(\tau^{(i)})^2} x_{t-1}^{(i)} - \frac{\Delta}{(\tau^{(i)})^2} (\boldsymbol{w}\boldsymbol{r}_{t-1} + \boldsymbol{w}_{noise}\boldsymbol{\psi}_{t-1} + \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) \\ &= \frac{\Delta}{(\tau^{(i)})^2} (x_{t-1}^{(i)} - \boldsymbol{w}\boldsymbol{r}_{t-1} - \boldsymbol{w}_{noise}\boldsymbol{\psi}_{t-1} - \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) \end{aligned}$$

Since $\partial\alpha / \partial\tau^{(i)} = -\Delta/(\tau^{(i)})^2$,

$$\frac{\partial x_t^{(i)}}{\partial \tau^{(i)}} = -\frac{\Delta^2}{(\tau^{(i)})^4} (x_{t-1}^{(i)} - \boldsymbol{w}\boldsymbol{r}_{t-1} - \boldsymbol{w}_{noise}\boldsymbol{\psi}_{t-1} - \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) \tag{11}$$

Multiplying the three equations (Equations (9) to (11)) leads to the gradient of the loss function with respect to the synaptic decay parameter.

B Parameter initialization

In this section, we discuss how we initialized all the parameters in our model.

The recurrent connectivity matrix ($\boldsymbol{w} \in \mathbb{R}^{N \times N}$) was initialized as a sparse random matrix drawn from $\mathcal{N}(0, 1.5/\sqrt{N \cdot P_c})$ with the initial connectivity probability (P_c) of 0.20.

The "noise" input matrix ($\boldsymbol{w}_{noise} \in \mathbb{R}^{N \times C}$) was initialized as a full random matrix drawn from the standard normal distribution (i.e., $\mathcal{N}(0, 1)$). The noise signals with C independent components ($\boldsymbol{\psi} \in \mathbb{R}^{C \times T}$) were initialized in the same manner.

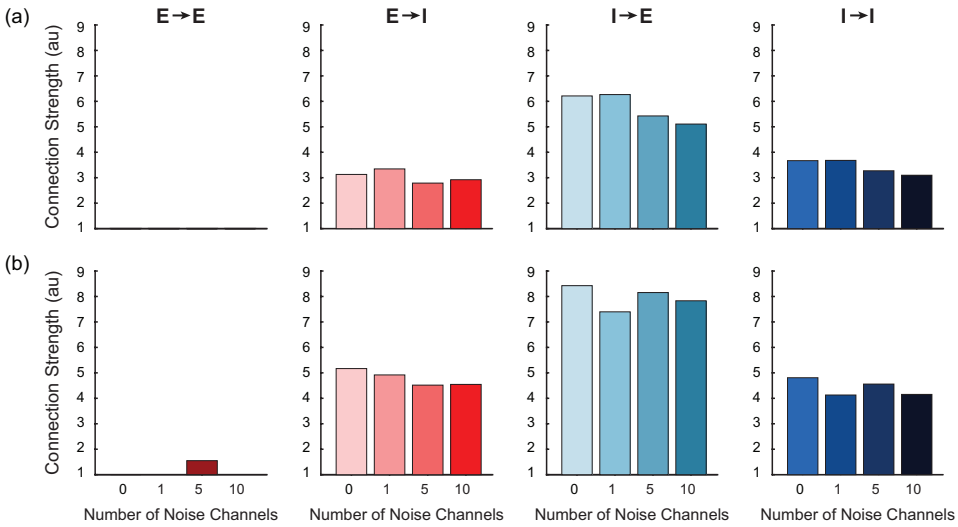
The input weight matrix ($\boldsymbol{w}_{in} \in \mathbb{R}^{N \times 1}$) was initialized as a random matrix drawn from the standard normal distribution.

The external noise ($\boldsymbol{\xi}$) and the readout weights ($\boldsymbol{w}_{out}$) were initialized as random matrices drawn from $\mathcal{N}(0, 0.01)$.

The bias term (b in Equation (4)) was initialized to 0.

C Comparison of the excitatory and inhibitory connection strength

In order to demonstrate that adding internal noise did not lead to significant changes in the recurrent connectivity structures, we averaged the recurrent synaptic weights separated by the four connection types for each noise condition ($C = 0, 1, 5, 10$): excitatory-to-excitatory ($E \to E$), excitatory-to-inhibitory ($E \to I$), inhibitory-to-excitatory ($I \to E$), and inhibitory-to-inhibitory ($I \to I$). We analyzed the models trained to perform the one-delay temporal XOR task (Supplementary Figure 1a) and two-delay XOR task (Supplementary Figure 1b). All the models, regardless of the noise condition, contained sparse $E \to E$ recurrent connections (Supplementary Figure 1a and b). Furthermore, we did not observe any significant changes in the mean synaptic strength across the four noise conditions, suggesting that the robustness conferred by the internal noise was not due to changes in synaptic connection strength.



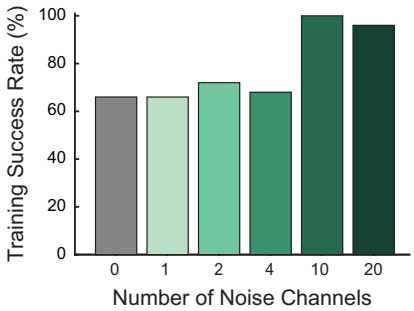
Supplementary Figure 1: Comparison of the mean synaptic connection strength across the four noise conditions ($C = 0, 1, 5, 10$). (a) Mean synaptic strength separated by the four synaptic connection types ($E \to E$, $E \to I$, $I \to E$, $I \to I$) for the RNNs trained to perform the one-delay temporal XOR task. (b) Mean synaptic strength separated by the four synaptic connection types for the RNNs trained to perform the two-delay temporal XOR task.

D Comparison of different internal noise structures

In this section, we show the effects of different internal noise structures on the training success rate and synaptic decay time constants. Particularly, we tried the following configurations:

- 1 channel of noise signals (i.e., $C = 1$) drawn from $\mathcal{N}(0, 400)$
- 2 channels of noise signals, each drawn from $\mathcal{N}(0, 100)$
- 4 channels of noise signals, each drawn from $\mathcal{N}(0, 25)$
- 10 channels of noise signals, each drawn from $\mathcal{N}(0, 4)$
- 20 channels of noise signals, each drawn from $\mathcal{N}(0, 1)$

For each configuration, we trained 50 RNNs to perform the one-delay temporal XOR task described in the main text. As shown in Supplementary Figure 2, the internal noise with high variance ($C = 1, 2, 4$) did not provide any improvement in the training success rate compared to the RNNs trained without any noise ($C = 0$). Adding more than 10 noise channels ($C = 20$) did not result in additional improvement in the success rate (darkest green in Supplementary Figure 2).



Supplementary Figure 2: Comparison of the training success rate between RNNs trained with no noise and RNNs trained with varying amount of internal noise. The grey bar is the baseline model (RNNs trained without any noise). The number and variance of the noise signals were varied (described in the text above).

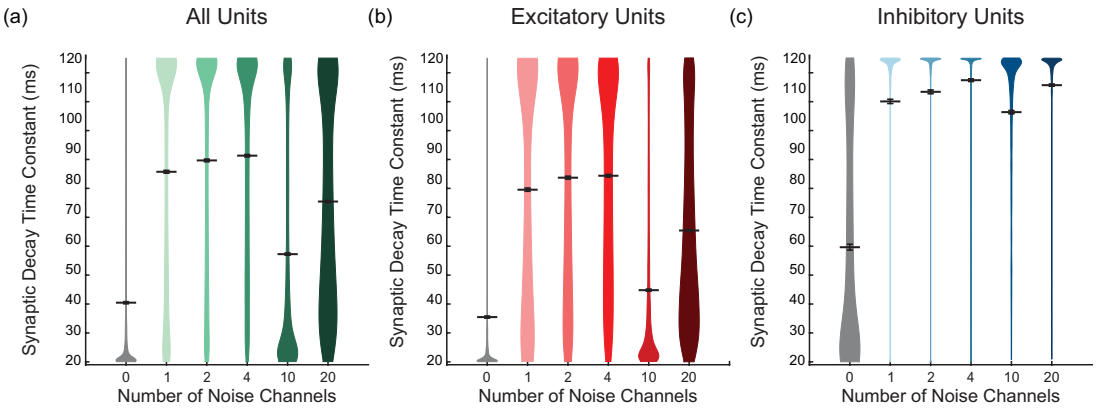
For the high-variance conditions ($C = 1, 2, 4$), the synaptic decay constant values for both excitatory and inhibitory were prolonged (Supplementary Figure 3). Consistent with our main findings (Figure 4), adding any amount of internal noise led to a significant increase in the inhibitory synaptic decay parameter (Supplementary Figure 3c). These results demonstrate both the number and variance of the internal noise signals can significantly impact the optimization of the synaptic decay parameter.

E Comparison of different excitation-inhibition (E-I) ratios

In this section, we show the effects of different E-I ratios on the training performance and synaptic decay time constants. To examine an interaction between E-I ratio and the modulatory effects of internal noise on working memory performance, we considered two additional network configurations: E-I ratio of 60/40 and 50/50. For each of the E-I ration conditions, we trained 20 RNNs on the one-delay temporal XOR task under the four noise conditions ($C \in \{0, 1, 5, 10\}$; see section 3.1 Continuous-rate recurrent neural networks). For both E-I ratio conditions, the training success rate was 100% for the baseline model (i.e., no internal noise; $C = 0$) as well as for the models trained with internal noise ($C \in \{1, 5, 10\}$).

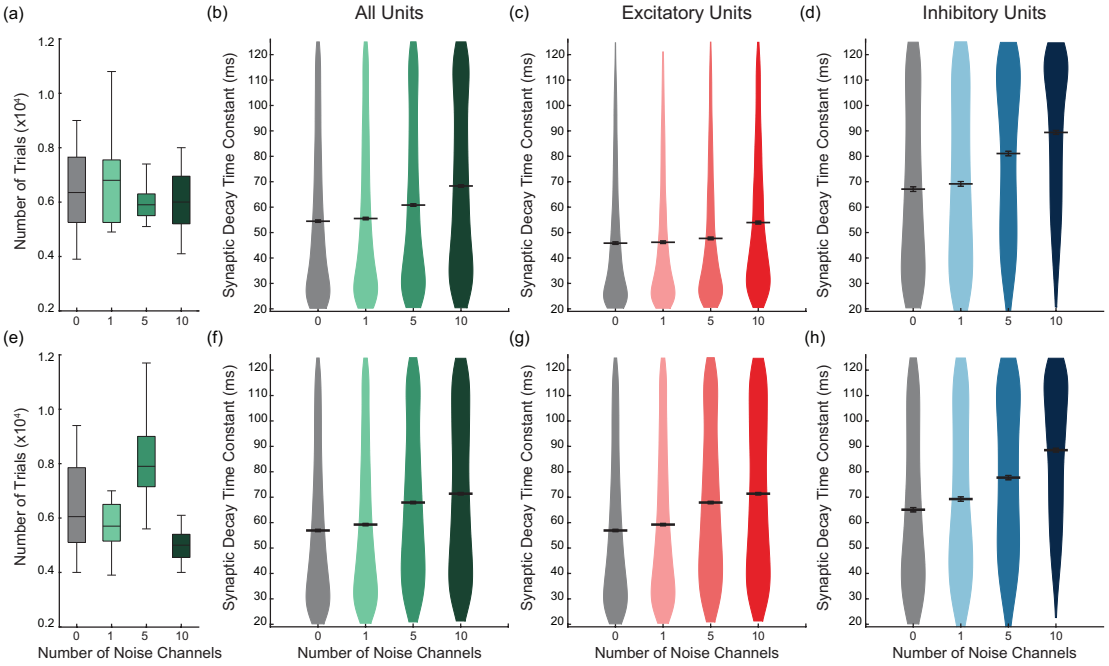
As shown in Supplementary Figure 4a and e, it took significantly fewer trials to train RNNs with E-I of 60/40 and 50/50 to perform one-delay XOR task as compared to models with E-I of 80/20 (Figure 4b). These results point to the importance of inhibitory units in working memory function and are in line with a recent study that links increased inhibitory signaling with improved cognitive abilities [Loomba et al., 2022].

Through analyzing the distribution of the optimized synaptic decay time constants (τ), we found that the synaptic decay constant distribution shifted toward the maximum value (125 ms) for the RNNs



Supplementary Figure 3: Comparison of the synaptic decay time constant values between RNNs trained with no noise and RNNs trained with varying amount of internal noise. The trained synaptic decay values are shown for all (a), excitatory (b), and inhibitory (c) units from the RNNs with different internal noise structures. Gray horizontal lines, mean.

trained with internal noise for both E-I ratio conditions (Figure 4d and h). This pattern of results is consistent with the observed findings of the RNNs with E-I ratio of 80/20 (Figure 5c).



Supplementary Figure 4: Training performance and synaptic decay time constant values of RNNs with E-I ratio of 60/40 (top) or 50/50 (bottom). RNN models of each E-I ratio condition were trained with varying amount of internal noise (i.e., 0, 1, 5, and 10 noise channels) to perform the one-delay temporal XOR task. The average number of trials required to reach the training criteria is shown in (a) and (e). The trained synaptic decay values for each E-I ratio condition are shown for all (b and f), excitatory (c and g), and inhibitory (d and h) units from the RNNs with different internal noise structures. Boxplot: central lines, median; bottom and top edges, lower and upper quartiles; whiskers, 1.5 x interquartile range; outliers are not plotted. Gray horizontal lines, mean.