

**Short-term plasticity-based working memory spiking model is resilient to synaptic heterogeneity**Gianmarco Tiddia<sup>1,\*</sup>, Luca Sergi<sup>1,2,\*</sup>, Simone Rubiu<sup>2</sup>, Antonio Incollu<sup>1,2</sup> and Bruno Golosio<sup>1,2</sup><sup>1</sup>*Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Cagliari, Cagliari, Italy*<sup>2</sup>*Department of Physics, University of Cagliari, Cagliari, Italy*

(Received 5 May 2025; accepted 19 December 2025; published 22 July 2026)

The presence of activity-silent working memory has been increasingly supported by experimental observation, and short-term synaptic facilitation is suggested to underlie this phenomenon. Originally proposed in the early 2000s, the synaptic theory of working memory was supported by a relatively simple, yet powerful, computational model. Recent advancements in simulation technology eased the introduction of heterogeneity in spiking models, which is a crucial factor in enhancing their dynamical richness and biological plausibility. However, introducing such heterogeneity often disrupts previously observed network dynamics. This work investigates the impact of synaptic heterogeneity in a spiking model of working memory sustained by short-term synaptic facilitation, observing that the model is resilient to additional variability.

DOI: [10.1103/zcmm-3cgh](https://doi.org/10.1103/zcmm-3cgh)**I. INTRODUCTION**

Working Memory (WM) is a cognitive process that allows temporary storage and manipulation of information. Through working memory, the representation of a sensory stimulus can be maintained even after the end of the stimulus itself [1–4]. This mechanism is crucial for many cognitive tasks, such as reasoning, speech, and visual processing.

It is believed that one of the neural bases of working memory is the persistent activity of cortical neurons, which is commonly observed in the prefrontal cortex (PFC) [5–7]. This mechanism is related to prior long-term synaptic modifications that lead to the creation of cell assemblies capable of showing enhanced activity as a response to a specific input [8–10]. Consequently, the information is maintained in the network in a pure spiking form. However, this theory has been challenged by experimental results suggesting that memory can also be maintained in a silent form, i.e., without enhanced spiking activity. It has been shown that memory can also be retrieved when the enhanced activity is interrupted for a short time, and also that neuron populations can exhibit periodic bursts instead of persistent activity during working memory tasks [11–13]. In this framework, short-term synaptic plasticity (STP), and more specifically short-term facilitation, is claimed to be crucial for information maintenance in a nonspiking form [14–17].

The neural basis of working memory is still under debate, but the most accredited hypothesis suggests that persistent activity and activity-silent mechanisms coexist and are responsible for the upkeep of differently prioritized items

[18–21]. Indeed, the introduction of short-term plasticity was beneficial in working memory models relying on persistent activity [22].

One of the models that suggested a prominent role of short-term facilitation in working memory is [15], which shows that a spiking network with synaptic facilitation can exhibit a bistable regime in which it can autonomously retain memories with periodic spiking activity without a significant firing rate increase. According to this model, memories are stored in a synaptic fashion, with spiking activity functional for synaptic facilitation upkeep. Various versions of this model, incorporating different neuron and synaptic current models as well as distinct simulation approaches, have been used to estimate memory capacity [23–25].

However, both the original model and its variants rely on a simplification that weakens their biological realism: parameter homogeneity. In contrast, the brain is an incredibly heterogeneous structure at all scales, and the PFC is no exception [26]. It remains uncertain whether heterogeneity has a role in cognitive processes or is just a consequence of noisy developmental processes. However, several works suggest that heterogeneity positively impacts the network performance of a learning task, but it may also lead to constraints [27,28].

The impact of heterogeneity in a working memory model has been studied in [29], which explored the constraints on the persistent activity of a network with heterogeneity in neuron and synapse parameters. The model investigated in this work relies on attractor networks exploiting persistent activity to store the information in the network, and the study showed that heterogeneity prevented the network from showing working-memory-related behavior, which can be restored by adding more mechanisms at the synaptic level. A similar result has also been discussed in [30].

In this work we investigate the impact of heterogeneity on the spiking model of [15], which relies on the STP mechanism to show the WM behavior. More specifically, we build on the variant of this model described in [24]. In contrast to the original model [15], where incoming spikes directly influence the

\*These authors contributed equally to this work.

†Contact author: gianmarco.tiddia@dsf.unica.it

membrane potential, this variant introduces an independent state variable to represent the post-synaptic current, which is modulated by incoming spikes and decays exponentially over time. To improve the biological plausibility of the model and evaluate its robustness, we introduce heterogeneity in the synaptic parameters of excitatory connections, given that in such a WM model, information is primarily maintained through facilitation of excitatory recurrent synapses.

## II. RESULTS

The model of [15] can store information by exploiting the STP mechanism, specifically short-term facilitation. The model is divided into several selective excitatory populations, each of which encodes a specific input, and an inhibitory population connected to the excitatory neurons in a nonspecific fashion. When an input targets a selective population of the network, it increases the population spiking activity, and the recurrent connections of the neurons belonging to this population temporarily increase their synaptic efficacy because of synaptic facilitation. As a result, random fluctuations of neuron activity in this stage can induce an autonomous Winner-take-all (WTA) mechanism, which, through global inhibition, suppresses the activity of all the excitatory populations but the one targeted by the input, and leads to an increase in the activity of the latter for a short time. This process can be reactivated periodically, and the activation period mainly depends on the synaptic depression time constant  $\tau_d$ . More details on the model architecture can be found in the Methods section.

Given the primary role of STP in the model, we want to investigate the impact of synaptic heterogeneity on the network dynamics. Indeed, heterogeneity reflects the natural diversity in biological systems, where synaptic efficacy, response, and neurotransmitter types vary even across synapses within the same network. To do so, we first test the network resilience to the heterogeneity of single parameters (i.e., synaptic efficacy, the STP time constants, and the fraction of facilitated synapses). Subsequently, we employ the most realistic degree of heterogeneity for all the examined parameters to verify the model's robustness.

The simulations are performed using the NEST [31] implementation of the model described in [24] with a few modifications that ease reproducibility and improve the realism of the external input of the model. For more details, see the Methods section.

The next section describes the simulation and test protocol, showing also the spiking activity of a simulation in which no heterogeneities to parameters such as synaptic efficacies and time constants are applied. This framework is reported as a reference to be compared with the results obtained when applying the rest of the heterogeneities to the synapse parameters.

### A. Simulation and test protocols

The simulation protocol, used for all the simulations reported in this work, can be described as follows: in the first 3000 ms all the populations receive the same background input and the spiking activity of the network assumes an

asynchronous irregular firing pattern, with an average firing rate of the excitatory neurons on the order of 1 Hz.

During the following 350 ms, one of the selective populations receives an additional signal representing the external stimulus (or cue), with the other populations still being targeted by the background input. We refer to this part of the simulation as *item loading*. Thereafter, all the excitatory neurons of the network remain targeted only by the background input for the following 1850 ms. At this time, the selective population targeted by the cue shows different behavior from the other populations as a result of the previous item loading. Finally, the background input intensity is decreased to interrupt the memory-specific response of the targeted selective population.

As discussed in [15], depending on the intensity of the background signal, the spiking activity of the targeted population may fall into three possible states. In this work, we will focus on the state in which the targeted population, after the item loading, enters a persistent state in which the neurons autonomously emit alternated synchronous spiking activity, a phenomenon called here *population spike*. During a population spike, all the neurons of the selective population emit a spike within 20 ms. This state of activity lasts until the background input is decreased. Indeed, this behavior is the most interesting from a neuroscientific point of view since it enables multiple-item maintenance with firing rate differences from standard activity in agreement with experimental observation [32].

Without heterogeneity in dynamic parameters (i.e., synaptic weights  $J_p$ ,  $J_b$  and STP time constants  $\tau_d$ ,  $\tau_f$ ), the targeted selective population generally emits five or six population spikes in the time between item loading and the decrease of the background input. We refer to this case, with homogeneous dynamic parameters, as the *reference case*. Indeed, heterogeneity of this version of the model is limited only to the values of the STP variables  $x$  and  $u$  at the beginning of the simulation. This heterogeneity prevents activity transients in the first timesteps of the simulation, as discussed in the Methods section. To perform the test protocol when heterogeneity is enabled, we take the homogeneous behavior as a reference and consider a simulation to have correctly stored the memory if it successfully reproduces five or more population spikes during the time that follows item loading. Population spikes are detected by measuring the firing rate of each population within a time interval of 25 ms. This simple criterion excludes all cases in which the network falls into a high firing rate regime with asynchronous irregular spiking activity (i.e., one of the possible states identified in [15]), or other anomalous behavior as shown in Fig. 1, which summarizes the test protocol and shows the raster plot and the firing rate of a subset of the network during the simulation.

To test the stability of our results, we performed 80 simulations using different seeds for random number generation, and computed the probability of correct information storage,  $P_{IS}$ , as the fraction of simulations that successfully reproduce five or more population spikes in the selected time window after item loading. The number of simulations has been chosen to have a statistically significant estimation for  $P_{IS}$ , as reported in Appendix A. Additionally, we added to the plot a shaded area corresponding to the SDOM computed over 80 simulations.

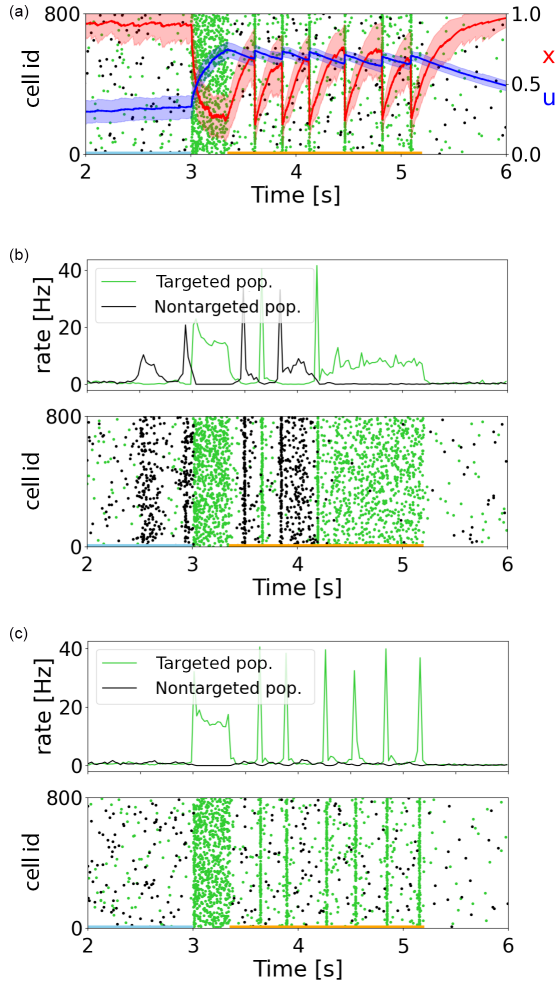


FIG. 1. (a) Simulation protocol. Raster plots of two of the selective neuron populations of the model, one targeted by an external stimulus (green) and a nontargeted (black). Every dot represents a spike emitted by a neuron. Averaged short-term plasticity (STP) variables  $x$  (red line) and  $u$  (blue line) for the synaptic connections within the targeted population are also shown. The shading represents the standard deviation of the STP variables. The first part of the simulation (sky blue horizontal line) represents standard activity, followed by the loading of an item to a selective population. The item loading is followed by a time (yellow horizontal line) when the selective population shows a memory-specific response. In the last part of the simulation, the background input is diminished. (b), (c) Test protocol. During the test protocol, every selective population's firing rate is measured to detect the population spikes. Panel (b) displays a transition to an asynchronous state and the emission of population spikes (PS) from the nontargeted population, and is therefore classified as anomalous behavior. Panel (c) exhibits a correct memory-specific response, as in the first panel.

In the reference case, we estimated this fraction to be 1.0. This value serves as a reference for comparing the  $P_{IS}$  values obtained in simulations with parameter heterogeneity, allowing us to assess whether heterogeneity negatively affects information storage.

Working Memory activity is associated with cortical oscillations in the theta range (4–8 Hz) [26,33–35]. In our model the synchronized activity corresponds to population spikes

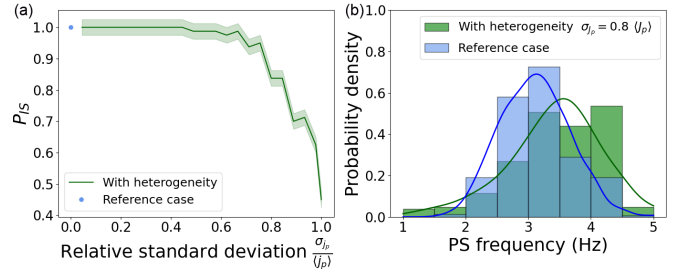


FIG. 2. (a) Probability of correct information storage  $P_{IS}$  as a function of the standard deviation chosen for potentiated synaptic efficacies ( $\langle J_p \rangle = 0.45$  mV). Shaded area indicates the error estimate. (b) Distribution of population spike (PS) frequencies for the case with heterogeneity only in the efficacy of excitatory-excitatory synapses (green), with a standard deviation of  $\sigma_{J_p} = 0.8 \langle J_p \rangle$ , and for the reference case (blue). The solid green and blue lines represent the kernel density estimates (KDEs) for the cases with heterogeneity and for the reference case, respectively.

(PSs), whose frequency mainly depends on the depression time constant  $\tau_d$ , which agrees with the theta frequency range. Thus, studying the PS frequency distribution at different levels of heterogeneity can provide insights into the relationship between synaptic variability and the corresponding theta wave distribution. For this reason, we also compute the population spike frequencies collected over 80 simulations using different seeds for random number generation and compare the distribution of the frequencies measured with the one obtained with a set of an equal number of simulations with homogeneous parameters.

## B. Heterogeneity in synaptic efficacy

As in the models of [15,24], we assume that excitatory-to-excitatory synapses connecting neuron groups selective for the same stimuli have been strengthened through a prior Hebbian learning process. Both experimental measurements and computational approaches suggest that synaptic efficacies follow a positively skewed distribution, which can be described using a lognormal function [36–40], which emerges as a consequence of learning. Thus, we modeled the synaptic efficacies using lognormal distributions, with the mean values of baseline and potentiated connections (i.e.,  $J_b$  and  $J_p$ , respectively) serving as their averages.

To determine the maximum standard deviation our network can sustain before showing deviation from the expected memory-specific response, we study the probability of correct information storage  $P_{IS}$  as a function of the standard deviation of the distribution. We noticed that heterogeneity in the synaptic efficacies of the baseline connections did not affect our network's ability to recall memories (see Fig. 8 in Appendix B). Thus, we investigated here the effects of introducing variability in the potentiated synaptic weights. Figure 2 shows the probability of correct information storage as a function of the standard deviation of the distribution of the potentiated synaptic efficacies and the effect of this heterogeneity on the frequency of population spikes.

We also verified that the probability of correct information storage in a network with heterogeneity in both potentiated

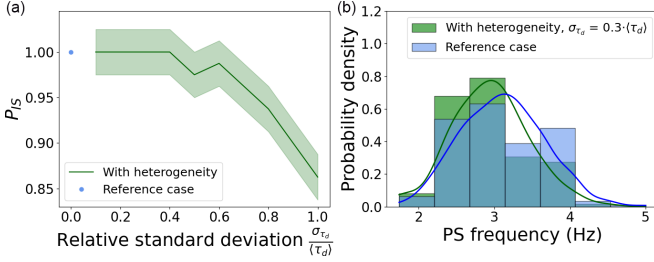


FIG. 3. (a) Probability of correct information storage  $P_{IS}$  as a function of the standard deviation of the normal distribution of the depression time constants  $\tau_d$  ( $\langle \tau_d \rangle = 200$  ms). Shaded area indicates the error estimate. (b) Distribution of population spike frequencies for the case with heterogeneity only in the depression time constant (green), with a standard deviation of  $\sigma_{\tau_d} = 0.3 \langle \tau_d \rangle$ , and for the reference case (blue). The solid green and blue lines represent the kernel density estimates (KDEs) for the cases with and for the reference case, respectively.

and non-potentiated weights is essentially identical to that of the network with heterogeneity only in potentiated weights. This result further highlights the crucial role of  $J_p$  compared to  $J_b$  in the network dynamics of the model.

### C. Heterogeneity on STP time constants

Heterogeneity in facilitation and depression time constants  $\tau_f$  and  $\tau_d$  reflects synapses' physiological characteristics. For instance, different number of calcium-permeable channels or the presence of neuromodulators play a key role in maintaining attention during the delay period of delay-response tasks typically employed to investigate WM behavior [41]. The time constants will follow normal distributions with the average being the same as their value in the homogeneous version of the model and a variable standard deviation.

By introducing variability only in the variable  $\tau_f$ , the model does not show any changes and correctly recalls the item loaded (see Fig. 11 in Appendix C). On the other hand, heterogeneity on depression time constant  $\tau_d$  leads to a substantial change in the network: for values of  $\sigma_{\tau_d} \simeq 100$  ms, the network has higher probability to move into a high firing rate state with asynchronous irregular spiking activity (i.e., no population spikes are emitted). As a result, the probability of correct information storage goes below 0.8. Figure 3 shows the probability of correct information storage as a function of the standard deviation of the depression time constant and the effect of this heterogeneity on the population spike frequency.

### D. Fraction of predominantly facilitated synapses

In the original model, every synapse connecting two excitatory neurons is considered facilitated (i.e.,  $\tau_f \gg \tau_d$ ). However, experimental observations show that only a fraction of the synapses are predominantly facilitated [26]. Thus, incorporating a fraction of nonfacilitated synapses improves the realism of the model and can shed light on the role of synaptic facilitation in working memory networks. We set depression and facilitation time constants of the nonfacilitated synapses driven from a normal distribution with average  $\tau_d = \tau_f = 325$  ms and standard deviation of 50 ms, in agreement

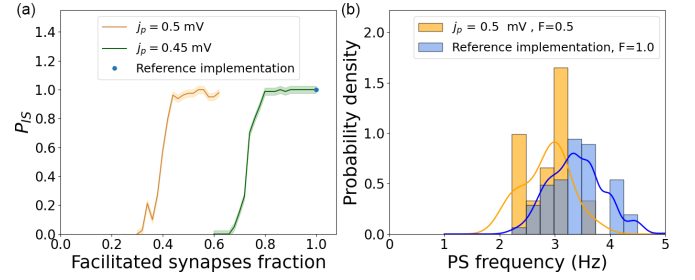


FIG. 4. (a) Probability of correct information storage as a function of the fraction of facilitated synapses with  $J_p = 0.45$  mV (green line) and  $J_p = 0.50$  mV (gold line). The first value is the one employed in [15]. Shaded area indicates the error estimate. (b) Distribution of population spike frequencies for two different fractions of facilitated synapses,  $F$ , without heterogeneity in synaptic properties: a fraction of 0.5 with  $J_p = 0.50$  mV (gold) and a fraction of 1.0 with  $J_p = 0.45$  mV (blue). The solid gold and blue lines represent the kernel density estimates (KDEs) for the corresponding two fractions of facilitated synapses.

with what experimentally observed in [26] for synapses of the prefrontal cortex without dominant facilitation or depression.

Figure 4 shows the probability of correct information storage as a function of the fraction of facilitated connections for two values of potentiated synaptic efficacy and the effect of the presence of nonfacilitated connections on population spike frequency. Since the number of population spikes (PSs) decreases as the fraction of facilitated synapses is reduced, we adjusted the criterion for determining whether a simulation successfully stores the memory by lowering the required number of PS from five to two or more.

The model with the value of the potentiated synaptic efficacy ( $J_p$ ) used in [15] maintains the information correctly only when the fraction of facilitated synapses is above 80% (green line in Fig. 4). However, only an increase of  $\sim 11\%$  in  $J_p$  would enable probabilities of correct recall above 0.8 even when only half of the synapses are predominantly facilitated (gold line of Fig. 4). This result reflects the importance of the value of  $J_p$  for the simulation. Nonetheless, the model with the original potentiated synaptic efficacy is enough to correctly retrieve the information in realistic fractions of facilitated synapses, which in the case of organized networks is close to 90% [26]. For this fraction, the probability of correct information storage is around 0.8.

### E. All synaptic heterogeneities activated

In this section all the synaptic heterogeneities described in the previous sections are included, with the degree of heterogeneity driven from experimental results.

Regarding facilitation and depression time constants, we chose values of standard deviation in line with experimental findings from [26]. For predominantly facilitated synapses, we set the standard deviation to 10% of the respective mean values of  $\tau_d$  and  $\tau_f$ . For the nonfacilitated case, we set the standard deviation to 50 ms, approximately 15% of the average values for the time constants. Additionally, the fraction of facilitated synapses is set to 0.9 to reflect that most, but not all, synapses are facilitated.

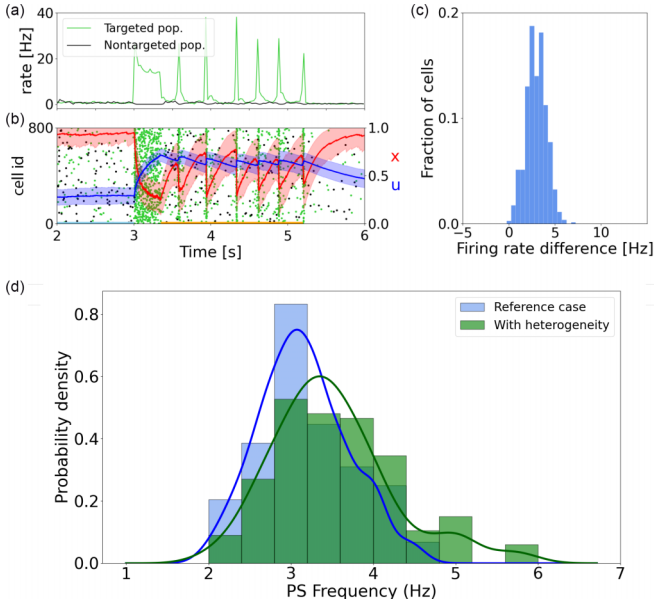


FIG. 5. Simulation with all the heterogeneities enabled. (a) Average firing rate for the targeted selective population and for one of the nontargeted selective populations. (b) Raster plot of a subset of the two selective populations. (c) Firing rate difference for each neuron of the targeted selective population between the delay period (yellow horizontal line) and activity before the item loading (sky blue horizontal line). (d) Distribution of population spike (PS) frequencies for  $F = 0.9$  with (green) and for the reference case (blue), along with their respective kernel density estimates (KDEs).

Regarding baseline and potentiated synaptic efficacies, direct experimental measurements of connection weights within specific neuronal ensembles are unavailable. However, experimental measurements on connections with nonspecific connectivity suggest that the standard deviation is approximately equal to the mean [40]. To account for this, we assume a lognormal distribution with a standard deviation equal to 80% of the mean, considering that variability tends to be lower within specialized neuronal groups.

We performed 80 simulations under these conditions using different seeds for random number generation to assess model robustness and stability. With all heterogeneities activated, only 18 over 80 seeds do not show the correct memory-specific response, resulting in a  $(78 \pm 3)\%$  probability of correct information storage, demonstrating that the network can support this biologically realistic variability in the vast majority of the simulations. Figure 5 shows a typical network behavior with all heterogeneities included, which remains qualitatively identical to the reference case. To further demonstrate that the model is resilient not only to the variability introduced at a specific and fine-tuned stable point corresponding to fixed mean values in the parameter distributions, we provided an additional source of heterogeneity by perturbing those means. We simulated the network by randomly extracting the mean values of the parameter distributions within  $\pm 5\%$  of the original values of the dynamic parameters  $J_p$ ,  $J_b$ ,  $\tau_f$ , and  $\tau_d$ . The results show that the network maintains statistically comparable  $P_{IS}$  values despite this variability, highlighting its robustness not only to hetero-

geneity within the distributions but also to variability in their mean values.

### III. DISCUSSION

In this study we investigate how heterogeneity in the synaptic properties influences the stability of a working memory model driven by short-term plasticity (STP). Our results show that incorporating biologically realistic synaptic heterogeneities into the STP parameters does not disrupt memory-related activity, while in previous works these heterogeneities were identified as a key factor preventing persistent activity from emerging in models relying on sustained firing [29,30]. The following sections discuss the impact of individual synaptic properties on network stability and dynamics, and ultimately, the combined effect of simultaneous heterogeneity across all previously considered parameters.

#### A. Heterogeneity in synaptic efficacy

In biological neuronal networks, synaptic efficacies typically have a skewed distribution with a positive skew. Several works describe this distribution using a lognormal function [36–39], with experimental measurements showing that the standard deviation of the distribution is close to its average [40].

Varying the standard deviations of the baseline synaptic efficacy  $J_b$  does not affect network stability, even when unrealistic values are used. Indeed, with our simulations, we noticed that deviations up to twice the mean do not significantly impact network stability (data not shown). This happens because the baseline synaptic weight is significantly smaller than the potentiated weight  $J_p$ , making nonpotentiated synaptic connections less likely to induce relevant fluctuations in spiking activity able to elicit population spikes.

Conversely, heterogeneity in potentiated synapses affects network stability: when the lognormal distribution reaches a standard deviation close to the mean, the probability of correct memory storage drops to approximately 0.5 (Fig. 2). This is caused by the higher probability of entering a high-rate state with asynchronous irregular spiking activity, as shown in [15] when higher background input is provided to the network (see Fig. 9 in Appendix B 2). A possible explanation for this strong impact on network stability is that, when the standard deviation is sufficiently high, positive feedback becomes more likely, driving high-rate states in the targeted population and triggering fewer population spikes even in nontargeted populations. In this regard, when the standard deviation of the weight distribution exceeds the mean, even nontargeted populations begin to enter a high-rate state and generate multiple population spikes.

Moreover, the choice of a standard deviation close to the mean for the lognormal distribution of the potentiated weights could be unrealistic for connections belonging to a specific neuronal ensemble. As discussed in [40], the distributions of synaptic efficacies follow a lognormal distribution, but neurons connected together have correlated synaptic efficacies, suggesting that the synaptic strengths of potentiated connections are more clustered than weaker ones.

Additionally, heterogeneity on potentiated connections affects the population spike frequency distribution: the log-normal distribution of  $J_p$  leads to higher PS frequencies, as shown in Fig. 2). This increase goes towards average rates in line with activity oscillations in the range of theta waves, as experimentally observed in [33,35,42].

### B. Heterogeneity on STP time constants

Heterogeneity in the facilitation time constant  $\tau_f$  does not affect network stability and memory maintenance. On the other hand, heterogeneity in the depression time constant  $\tau_d$  on network stability and population spike dynamics is not negligible. This is reasonable, since the latter has a lower time constant than the former, and modulates the dynamics of the resources (i.e., neurotransmitters) available to be released for transmitting the spike, with this amount decreasing considerably during the emission of a spike. The probability of correct information storage,  $P_{IS}$ , slightly decreases when the standard deviation of the distribution equals the mean value of  $\tau_d$  (Fig. 3). In this condition it is more likely to have a transition to a high-rate state with asynchronous irregular spiking activity. Such behavior is also observed when heterogeneity is enabled on potentiated synaptic efficacies. However, unlike that latter case, the normal distribution of  $\tau_d$  does not induce population spikes in other nontargeted populations. Moreover, a standard deviation near to the average of the distribution is unrealistic, as shown by [26], which reports a standard deviation for depression time constant closer to 25% of the mean value. Another effect of the heterogeneity of the depression time constant is that it causes the population spike to have broader and lower peaks in the firing rate, indicating that neurons spike at more different times, making their activity less synchronous compared to other cases (see Fig. 10 in Appendix C).

Regarding the impact of the heterogeneity on the rate of the population spikes, we notice a slight decrease in the average rate when heterogeneity on the depression time constant is added, whereas heterogeneity on the facilitation time constant does not lead to significant differences in this regard.

### C. Fraction of predominantly facilitated synapses

The network can reproduce memory-related activity even when not all excitatory connections are predominantly facilitated, as shown in Fig. 4. Obtaining consistent results with different fractions demonstrates that working memory processes are stable and replicable across a wide range of biological conditions in this short-term plasticity-based model. Figure 4 demonstrates that the network can reproduce synchronized activity even with 50% of facilitated synapses, with a relatively small ( $\sim 11\%$ ) increase in potentiated synaptic efficacy. However, this fraction is more biologically realistic for nonspecialized neuron groups [26]. For specialized neuron groups, a higher fraction, closer to 90%, would be more appropriate [26]. With this value, the network is stable without any adjustment in the potentiated synaptic weights. It can also be observed that the fraction of facilitated synapses largely affects the frequency of population spikes: a larger number of nonfacilitated synapses (and thus a smaller frac-

tion of facilitated synapses) results in a lower population spike frequency (see Fig. 12 in Appendix C). This is because the frequency is mainly governed by the depression time constant  $\tau_d$ , which is approximately the period of PS. Non-facilitated synapses have a longer  $\tau_d = 325$  ms [26], leading to a lower PS frequency. Another effect of varying the fraction of nonfacilitated synapses is a broader variability in PS frequency. This is consistent with experimental findings from [33], which reported a wide range of frequencies in intracortical recordings during the Sternberg memory task.

### D. All synaptic heterogeneities activated

When all the synaptic heterogeneities are activated (i.e., synaptic efficacies, STP time constants and fraction of predominantly facilitated synapses), the network can correctly reproduce memory-specific activity with a probability of 78%. This result has been achieved by choosing, for each synaptic parameter, a degree of heterogeneity in line with experimental observation, whereas the tests done on single parameters investigated the maximum degree of heterogeneity the network can have when still keeping the usual behavior. This outlines the robustness of the network and its stability under biologically realistic conditions, without the need for tuning any other network parameter. As can be seen in Fig. 5, a typical simulation with this degree of heterogeneity is qualitatively similar to a simulation used as a reference [see Fig. 1(a)], with firing rate difference between delay period and standard activity lower than 10 Hz, as observed in [11] and the original model [15,24]. The analysis of population spike frequency reveals an increase in the average firing rate, reaching values between 3 and 4 Hz, accompanied by a broader distribution. This trend suggests a more irregularly synchronized activity, a well-documented experimental finding [43], and a known consequence of network heterogeneities in other models [36].

### E. Concluding remarks

Our results show that biologically realistic synaptic heterogeneities do not disrupt working-memory-related activity in STP-based models. Unlike persistent-activity-based models, which often require precise parameter tuning, STP-based models demonstrate robustness against heterogeneity in synaptic efficacies, time constants, and fraction of facilitated synapses.

Moreover, while baseline synaptic efficacy heterogeneity has minimal impact on stability, heterogeneity in potentiated synapses and depression time constants can lead to network instability when exceeding experimentally observed values. However, when all heterogeneities are introduced simultaneously with fluctuations aligned with experimental findings, the network maintains a high probability of correct memory storage with also a slightly higher memory capacity. Interestingly, this result complements recent large-scale studies [44,45], which showed that interarea heterogeneity across cortical regions enhances resistance to distractors in distributed working memory models. In contrast, the intraarea heterogeneity examined here (i.e., variability among neurons within the same cortical population) leads to a slight increase in memory

capacity, suggesting that diversity at different circuit scales may support complementary aspects of information upkeep.

Our model includes both intrinsic and external sources of stochasticity: variability in synaptic parameters and stochastic fluctuations in network inputs caused by Poisson spike train statistics. Both forms of stochasticity can enhance the representation of input signals in a neuronal population, but their combined effect is not simply additive: excessive noise in an already heterogeneous network can degrade information coding [46]. When both sources of variability are present at realistic levels, our network can store items correctly and even achieve a slightly higher memory capacity, as shown in Appendix D. However, if both types of stochasticity are too strong, the  $P_{IS}$  drops significantly. These results suggest that systematically studying how different levels of intrinsic and external heterogeneity interact could provide important insights into the determinants of working memory performance.

Moreover, in this study, we confined heterogeneity to excitatory synaptic parameters. Intrinsic neuronal variability (related to thresholds, membrane time constants and capacitance) and heterogeneity of inhibitory synapses have not been investigated, although both are well known to influence cortical dynamics [47]. While this represents a limitation of the model, focusing on heterogeneities of a few key parameters allows us to isolate their effects and gain a clearer understanding of the underlying mechanisms. The choice to restrict the analysis to excitatory synapses is also motivated by the availability of more comprehensive experimental data, as well as by the fact that pyramidal cells in the PFC generally exhibit short-term facilitation [26,48].

Future work should include a more comprehensive set of biologically realistic heterogeneities, in both neuronal and synaptic parameters, and explore whether these results are generalizable to other network architectures. Furthermore, incorporating more detailed biophysical mechanisms, such as homeostatic plasticity, could provide a deeper understanding of how real neural circuits maintain functional stability while changing synaptic properties.

## IV. MATERIALS AND METHODS

### A. Working memory spiking network model

The model presented in this study has been implemented in the widely known NEST Simulator [31] in [24]. Please refer to [24] and [49] for a detailed description of the NEST implementation. In the following, we provide a general description of the model and its structure.

The network comprises  $N_E$  excitatory and  $N_I$  inhibitory neurons with a 4/1 ratio. We chose the leaky integrate-and-fire neuron model with exponential postsynaptic current to describe the dynamics of each neuron. The excitatory neurons are divided into  $p$  selective populations containing a fraction  $f$  of the excitatory neurons of the network and a nonselective population. Thus, each selective population has  $fN_E$  neurons and the nonselective population counts  $(1 - fp)N_E$  neurons. All the connections between excitatory neurons are modulated by short-term plasticity. See Appendixes F and G for more details on the STP synapses dynamics and neuron model. Synapses connecting excitatory neurons belonging to the

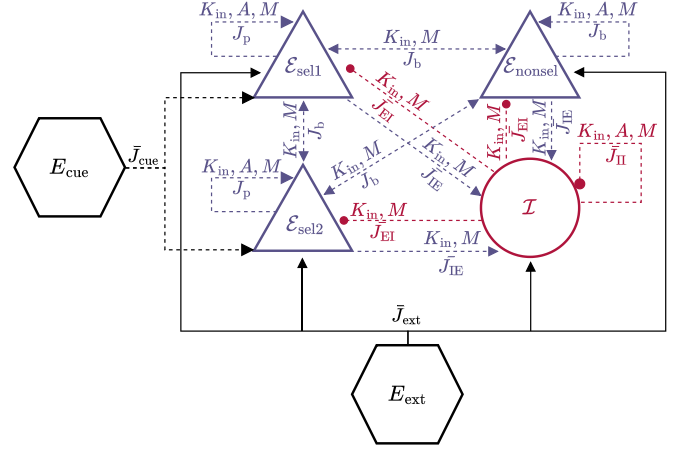


FIG. 6. Representation of the network following the graphical notation of [50]. For simplicity, only two selective populations  $\mathcal{E}_{sel}$  are shown. All the connections between neurons of the network follow the fixed-indegree connection rule and the synaptic delays are uniformly distributed. Inhibitory population and its outgoing connections are represented in red, and excitatory neuron populations and their connections are represented in blue. The contribution from external stimuli is depicted in black.

same selective population have a potentiated synaptic efficacy  $J_p$ . In contrast, all the other excitatory-to-excitatory connections show a baseline value  $J_b < J_p$ . The inhibitory population is connected to the excitatory neurons in a nonspecific fashion. This enables, through global inhibition, the mechanism of winner-take-all (WTA) for the competition between selective populations following the arrival of an external stimulus.

Figure 6 depicts a diagram of the model, using the graphical notation proposed in [50]. Appendix E summarizes the key synaptic parameters of the reference model. A comprehensive description of the network structure and the rest of the model parameters is also available in [49] using a tabular form that follows the guidelines of [51].

### B. Features added to the spiking model

We implemented two new features in the spiking model with respect to the previous version of [24]. The first one is an improvement in the reproducibility of the model, whereas the second one is an improvement in the realism of the external input provided to the network.

#### 1. Reproducibility

The previous network model was equipped with a custom synapse derived from an existing NEST synapse model. It was written in C++ and implemented in a specific version of the NEST Simulator (see [24] for the implementation details), and used the STP model employed in [15]. However, to improve the reproducibility of the model, we implement the synapse model with the domain-specific language NESTML [52] (version 8.0.0 [53]). In this way, the model can be easily executed with every NEST 3 installation. We verified that using the same seed for random number generation and the same version of NEST employed in [24], the C++ implementation and the NESTML implementation of the spiking

model produce the same results. This is also true for different NEST 3 versions. The link to access the code used to simulate the model is provided in the Data Availability Statement.

## 2. External input

In [15] the neurons were stimulated using a Gaussian white noise current to represent the stimuli from neurons external to the simulated model. This approximates the asynchronous-irregular firing patterns observed in biological neurons. However, to better represent this firing pattern, a more realistic option would be to provide the neurons with Poisson spike trains with a given frequency and amplitude. We decided to provide the model with both options, to be chosen by the user, with the Poisson spike train being the default option. In Appendix H we derive the frequency and the amplitude of the Poisson spike train needed to elicit the same depolarization as the Gaussian white noise current employed in the first version of the model.

In the NEST implementation of the model, the Gaussian white noise current is created using the `noise_generator` device (as described in [24]), whereas here the Poisson spike train is generated using the `poisson_generator` device.

## 3. Heterogeneity in the STP variables

In the implementation of [24], the initial values for the STP variables  $x$  and  $u$  were constant for all the synapses. However, the nonhomogeneity of the initial conditions mimics previous neural activity and mitigates transients at the beginning of the simulation. Here  $u$  follows a normal distribution, whereas  $x$  is uniformly distributed between 0 and 1.

## ACKNOWLEDGMENTS

This study was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreements No. 945539 (Human Brain Project SGA3) and No. 785907 (Human Brain Project SGA2), and by the Italian Ministry of University and Research (MUR) Piano Nazionale di Ripresa e Resilienza (PNRR), project e.INS Ecosystem of Innovation for Next Generation Sardinia-spoke 10-CUP F53C22000430001-MUR code ECS00000038, and the BRAINSTAIN project, funded by INFN-CSN5. We acknowledge the use of Fenix Infrastructure resources, which are partially funded from the European Union's Horizon 2020 research and innovation program through the ICEI project under the Grant Agreement No. 800858.

Conceptualization, Investigation, Methodology, Validation, Visualization: B.G., G.T., L.S., S.R., A.I. Data curation, Formal analysis, Software: G.T., L.S., S.R., A.I. Supervision: B.G., G.T. Funding acquisition, Project administration, Resources: B.G. Writing—original draft: B.G., G.T., L.S., S.R., A.I. Writing—review and editing: B.G., G.T., L.S., S.R., A.I.

## DATA AVAILABILITY

The data that support the findings of this article are openly available [54].

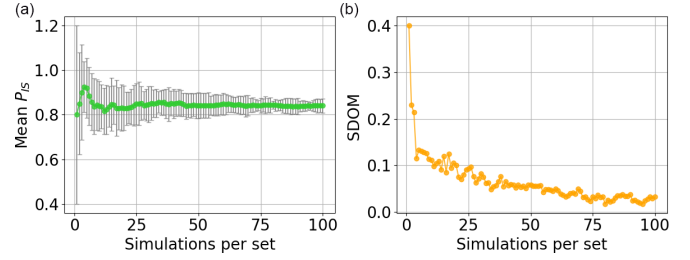


FIG. 7. (a) Average  $P_{IS}$  over 10 sets as a function of the set size, ranging from 1 to 100 simulations. (b) Standard deviation of the mean (SDOM) as a function of the set size, ranging from 1 to 100 simulations.

## APPENDIX A: EVALUATION OF A STATISTICALLY SIGNIFICANT NUMBER OF SIMULATIONS

To assess the number of simulations needed for estimating with statistical significance the probability of correct information storage, we study the mean of  $P_{IS}$  for 10 independent blocks of simulations and depict the average of those blocks as a function of the block size  $N$ . The results are depicted in Fig. 7. Error bars represent the standard deviation across the 10 block means, i.e., the empirical variability of the estimated mean when repeating the experiment with independent seeds. Both the mean and its interblock standard deviation converge by  $N \approx 80$ , indicating that averaging over 80 simulations provides a stable and representative estimate of  $P_{IS}$ . Thus, additional simulations do not further improve the accuracy or reduce the uncertainty of the estimate, meaning that this number of simulations is sufficient to ensure statistical reliability while maintaining computational efficiency.

## APPENDIX B: HETEROGENEITY IN SYNAPTIC EFFICACY

### 1. Heterogeneity in the synaptic efficacy of baseline connections

In the main body of the work, we discuss the effect of heterogeneity on potentiated synaptic efficacies, saying that in the case of baseline efficacies, the effect of heterogeneity is negligible. Indeed, Fig. 8 shows it, by performing the test protocol also used for potentiated connections. We can see that the probability of correct information storage,  $P_{IS}$ , remains above 0.9 even for large standard deviations, confirming that the network is robust to heterogeneity in baseline synaptic efficacy. As can be noticed, the impact on population spike frequency is negligible as well, and shows a slight shift towards lower frequencies.

### 2. Heterogeneity in the synaptic efficacy of potentiated connections

This section presents additional results obtained introducing heterogeneity in the synaptic efficacy of excitatory-to-excitatory neurons of the potentiated selective population  $J_p$ .

Figure 9 shows raster plot, average firing and firing rate difference between the delay period and standard activity for two scenarios that can arise when introducing variability on

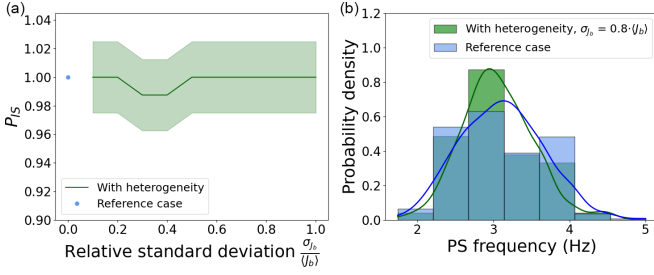


FIG. 8. (a) Probability of correct information storage  $P_{IS}$  as a function of the standard deviation chosen for baseline synaptic efficacies. Shaded area indicates the error estimate. (b) Distribution of population spike (PS) frequencies for the case with heterogeneity only in the efficacy of excitatory-excitatory synapses of different populations (green), with a standard deviation of  $J_b = 0.8 \langle J_b \rangle$ , and without heterogeneity (blue). The solid green and blue lines represent the kernel density estimates (KDEs) for the cases with and without heterogeneity, respectively.

$J_p$ . In the first three panels [Figs. 9(a)–9(c)], the network exhibits a response similar to the case without heterogeneity. This behavior is mainly observed when the standard deviation is approximately below 80% of the mean. In contrast, the last three panels [Figs. 9(d)–9(f)] illustrate a distinct behavior that emerges as the standard deviation approaches the mean  $\langle J_p \rangle = 0.45$  mV.

As shown in panel Fig. 9(e), when the standard deviation is approximately 90% of the mean, only a few population spikes occur before the network transitions into a high-rate state, resembling the state observed in the original model with a higher background input. Under this condition of high heterogeneity, positive feedback is more likely to occur, triggering population spikes even in nontargeted populations. This is due to the presence of synapses with very high efficacy, which eases the process.

Another effect of high heterogeneity is a broader distribution of firing rate differences, as Fig. 9(f) shows. The greater variability in potentiated efficacy drives this effect.

## APPENDIX C: HETEROGENEITY IN STP TIME CONSTANTS

### 1. Heterogeneity in depression time constant

This section presents additional results obtained introducing heterogeneity in depression time constant  $\tau_d$ .

Figure 10 shows the raster plot, average firing rate, and rate difference between the delay period and the spontaneous state for simulations with different standard deviations for depression time constant  $\sigma_{\tau_d}$ . The first standard deviation,  $\sigma_{\tau_d} = 55$  ms, was selected due to its similarity to values reported by [26]. The second one,  $\sigma_{\tau_d} = 110$  ms, was chosen because it shows the threshold beyond which the probability of correct information storage decreases substantially.

The effect of a higher standard deviation, as shown in Fig. 10, is an increase in the number of population spikes, which exhibit a lower average firing rate but a longer duration.

### 2. Heterogeneity in facilitation time constant

As anticipated in the results section, despite the high variability in the facilitation time constant  $\tau_f$ , the network is still capable of generating high synchronized activity, effectively storing the information. Here, we present the study of the probability of a correct memory response as a function of the standard deviation  $\sigma_{\tau_f}$ . The probability of a correct memory response remains stable, even with an unrealistically large standard deviation equal to the mean value, as shown in Fig. 11(a). The right panel of the figure shows that the population spike frequency has a negligible change due to this kind of heterogeneity.

### 3. Fraction of predominantly facilitated synapses

This section shows additional data from simulations in which the fraction of predominantly facilitated synapses is around 50%. As discussed in the main body of the work, this fraction can be achieved in the model with a modest increase in the value of potentiated synaptic efficacy, and the result of the test protocol shows a significant drop in the population spike rate. Figure 12 shows a simulation with this fraction of facilitated synapses, with a significant decrease in the number of PS emitted.

## APPENDIX D: MEMORY CAPACITY OF THE NETWORK

In [24] some of us showed that the maximum memory capacity, when using the original dynamic parameters of Mongillo [15], was limited to three items. We further examined this property by analyzing the number of population spikes in both target and nontarget populations when four items were simultaneously loaded into the network.

In the reference case, the first population that received an item rarely maintained its synchronized activity once additional items were loaded into other populations. In contrast, when heterogeneity was introduced, the first population was more likely to preserve its oscillatory and synchronized activity, indicating that its dynamics were not disrupted by the encoding of subsequent items. This effect can be explained by the variability in the facilitation time constant  $\tau_f$ , which introduces diversity in synaptic recovery times. Such heterogeneity reduces interference between populations and helps stabilize reverberatory activity, allowing earlier-activated populations to maintain their memory-related activity despite subsequent item encoding.

Specifically, the fraction of simulations in which all four populations exhibited at least three population spikes was zero in the reference case and 0.2 when synaptic heterogeneity was present. Figure 13 presents a raster plot illustrating a correct four-item recall in the presence of heterogeneity in all dynamics parameters.

This shows that intrinsic heterogeneity in synaptic parameters enhances the storage of items within the network.

## APPENDIX E: REFERENCE MODEL PARAMETERS

Table I describes the key parameters of the network model for the reference implementation. For the sake of simplicity, not all the model parameters are described. For a

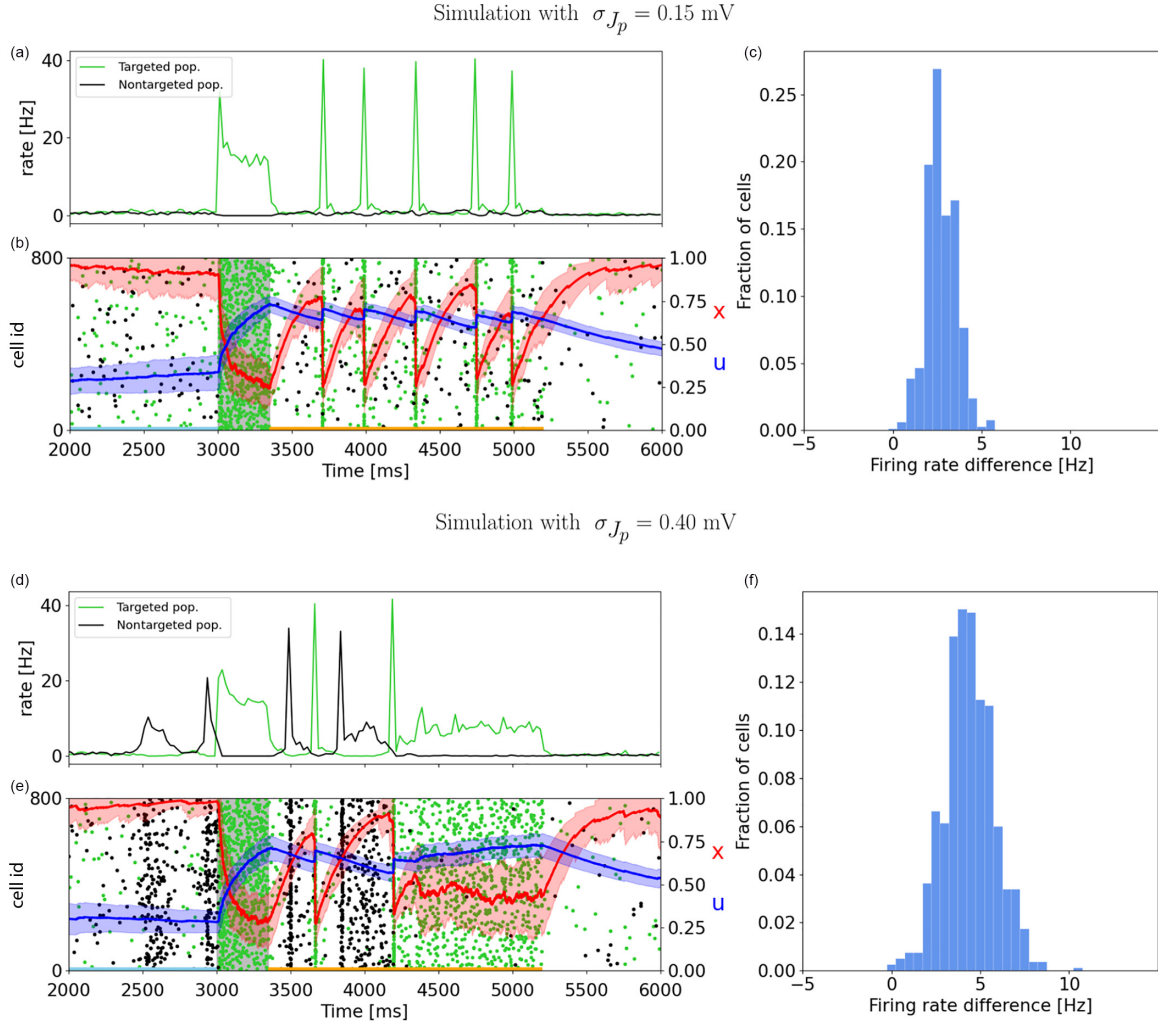


FIG. 9. Simulations with different degrees of heterogeneity of potentiated synaptic efficacy. (a), (d) Average firing rate for the targeted selective population and for one of the nontargeted selective populations. Averaged short-term plasticity variables  $x$  (red line) and  $u$  (blue line) for the synaptic connections within the targeted population are also shown. Bands in the same color code represent the standard deviation of the values of  $x$  and  $u$ . (b), (e) Raster plots of a subset of the two selective populations. (c), (f) Firing rate difference for each neuron between delay period and standard activity.

comprehensive description of, respectively, model architecture and parameters without parameter heterogeneity, please refer to Tables 7.1 and 7.2 of [49]. Indeed, the only parameter difference in this table with respect to [49] is that the initial values of the STP variables  $x$  and  $u$ , as stated in Sec. IV, are distributed. In the original version, the initial values for these variables were  $x = 1.0$  and  $u = U$ .

#### APPENDIX F: SHORT-TERM SYNAPTIC PLASTICITY

Short-term plasticity (STP) is a synaptic mechanism that reversibly changes the synaptic efficacy of a synapse on a timescale from hundreds to thousands of milliseconds. It results from the dynamics of neurotransmitters modulated by the calcium concentration in the presynaptic terminal. Indeed, the emission of a spike leads to an increase in presynaptic calcium concentration, which triggers the release of neurotransmitters. When the neuron is silent, the presynaptic calcium concentration decreases and returns to its baseline concentration, whereas the number of

neurotransmitters available for release in the presynaptic terminal returns to its initial value. These two processes have different characteristic times, and can be described using a phenomenological model [15,55].

Calling  $u$  the normalized presynaptic calcium concentration,  $\tau_f$  the time constant related to the decrease of the concentration when the neuron is silent,  $x$  the normalized amount of neurotransmitters ready to be released and  $\tau_d$  the time constant related to the neurotransmitters recovery after being released, the dynamics of a synapse connecting the presynaptic neuron  $i$  and the postsynaptic neuron  $j$  can be described by the following equations:

$$\begin{aligned} \frac{du_{i,j}}{dt} &= -\frac{u_{i,j}}{\tau_f} + U(1 - u_{i,j}) \sum_s \delta(t - t_s^{(i)}), \\ \frac{dx_{i,j}}{dt} &= \frac{1 - x_{i,j}}{\tau_d} - u_{i,j}x_{i,j} \sum_s \delta(t - t_s^{(i)}), \end{aligned} \quad (\text{F1})$$

where  $\delta(\cdot)$  is the Dirac delta function,  $U$  is the baseline utilization factor, and the sum is over the spike times  $t_s^{(i)}$  of

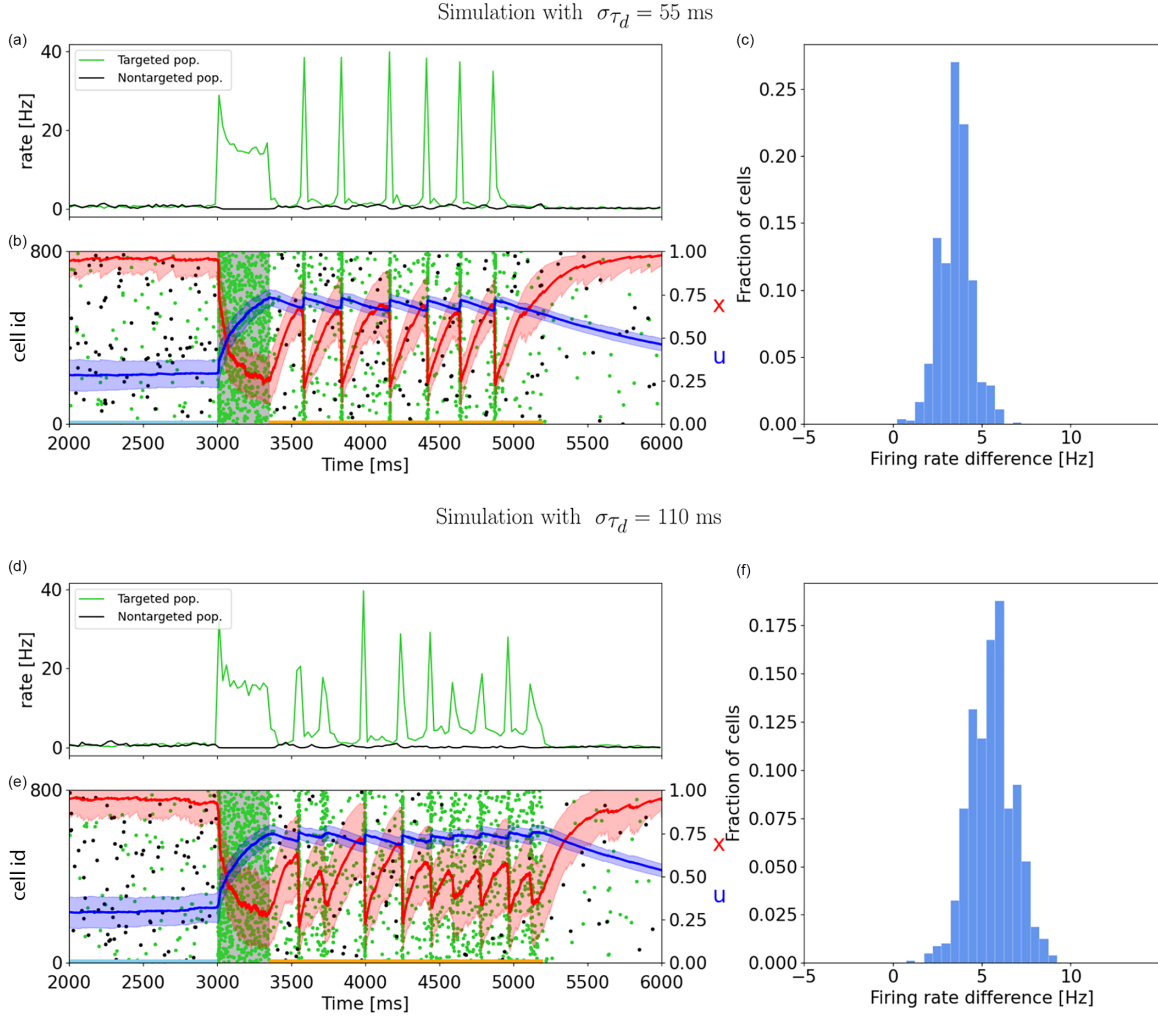


FIG. 10. Simulations with different degrees of heterogeneity of depression time constant. (a), (d) Average firing rate for the targeted selective population and for one of the nontargeted selective populations. (b), (e) Raster plots of a subset of the two selective populations. Averaged short-term plasticity variables  $x$  (red line) and  $u$  (blue line) for the synaptic connections within the targeted population are also shown. Bands in the same color code represent the standard deviation of the values of  $x$  and  $u$ . (c), (f) Firing rate difference for each neuron between delay period and standard activity.

the presynaptic neuron  $i$ . The synaptic modulation takes place during the spike emission, so that

$$J_{i,j}(t) = J_{i,j}^{(\text{abs})} u_{i,j}(t - \hat{\delta}_{i,j}) x_{i,j}(t - \hat{\delta}_{i,j}), \quad (\text{F2})$$

where  $J_{i,j}^{(\text{abs})}$  is the absolute synaptic efficacy for the synapse connecting neurons  $i$  to neuron  $j$  and  $\hat{\delta}_{i,j}$  is the synaptic delay. Thus, when a spike is fired, the product  $Jux$  describes the synaptic efficacy.

#### APPENDIX G: NEURON MODEL

The neuron model employed for the simulations is the leaky integrate-and-fire model with exponential postsynaptic currents. The subthreshold dynamics is described by the differential equation

$$\tau_m \frac{dV_j}{dt} = -V_j + R_m (I_j^{\text{exc}} + I_j^{\text{inh}} + I_{\text{ext},j}), \quad (\text{G1})$$

where  $\tau_m$  is the membrane time constant,  $V_j$  is the neuron's membrane potential,  $R_m$  is the membrane resistance,  $I_j^{\text{exc}}$  and

$I_j^{\text{inh}}$  the excitatory and inhibitory synaptic current and  $I_{\text{ext},j}$  represents the external input to the network.

Since the synaptic input from other neurons of the network is designed as exponential postsynaptic currents,  $I_j^{\text{exc}}$  and  $I_j^{\text{inh}}$  can be described by the following equations:

$$\begin{aligned} \tau_{\text{exc}} \frac{dI_j^{\text{exc}}}{dt} &= -I_j^{\text{exc}} + \sum_i \alpha J_{i,j}^{\text{exc}}(t) \sum_s \delta(t - t_s^{(i)} - \hat{\delta}_{i,j}), \\ \tau_{\text{inh}} \frac{dI_j^{\text{inh}}}{dt} &= -I_j^{\text{inh}} + \sum_i \alpha J_{i,j}^{\text{inh}} \sum_s \delta(t - t_s^{(i)} - \hat{\delta}_{i,j}), \end{aligned} \quad (\text{G2})$$

where  $i$  is the index of the presynaptic neurons targeting the neuron  $j$ .  $\tau_{\text{exc}}$  and  $\tau_{\text{inh}}$  represent the time constant of the excitatory and inhibitory synaptic currents, respectively. The term  $\hat{\delta}_{i,j}$  indicates the synaptic delays.  $\alpha$  is a term that, multiplied by the synaptic efficacy  $J$  (expressed in mV), return the synaptic efficacy expressed in pA (see next section). Please note that only the equation describing the behavior of the excitatory current has a time-dependent synaptic efficacy. This

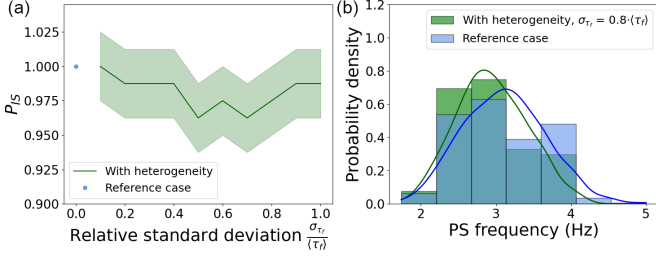


FIG. 11. (a) Probability of correct information storage  $P_{IS}$  as a function of the standard deviation chosen for facilitation time constant. Shaded area indicates the error estimate. (b) Distribution of population spike (PS) frequencies for the case with heterogeneity only in facilitation time constant  $\tau_f$  (green), with a standard deviation  $\sigma_{\tau_f} = 0.8 \langle \tau_f \rangle$ , and without heterogeneity (blue). The solid green and blue lines represent the kernel density estimates (KDE) for the cases with and without heterogeneity, respectively.

is due to the fact that STP is implemented only for excitatory-to-excitatory connections.

The external input is modeled as using Poisson spike train input, with results compatible with the ones achieved in the original model and the implementation of [24] when the input is modeled as a Gaussian white noise current.

### 1. Current able to elicit a PSP of amplitude $J$

In [15] the synaptic efficacy is expressed in mV, and indicates the change of the membrane potential of the postsynaptic neuron due to a spike received from a presynaptic neuron connected through a synapse of strength  $J$ . However, the actual signal delivered through the connection is a current, and in NEST the synaptic strength of a synapse must be expressed in pA. In [24] we delivered an equation that computes a factor  $\alpha$  that multiplied by  $J$  is able to return the synaptic efficacy of the synapse in pA. The following equation is valid for LIF neurons with exponential postsynaptic current, refer to [24] for the full derivation:

$$\alpha = \frac{I_0}{V(t_{\max})} = \left\{ \frac{\tau_m \tau_s}{C_m (\tau_s - \tau_m)} \left[ \left( \frac{\tau_m}{\tau_s} \right)^{-\tau_m/(\tau_m - \tau_s)} - \left( \frac{\tau_m}{\tau_s} \right)^{-\tau_s/(\tau_m - \tau_s)} \right] \right\}^{-1}. \quad (G3)$$

Here  $I_0$  represents the input current, which has a form  $I_0 \approx e^{t/\tau_s}$  and  $V(t_{\max})$  represents the maximum value of the postsynaptic potential due to the injection of the input  $I_0$ , assuming that  $V(t=0) = 0$  and the input targets the neuron at  $t = 0$ .

## APPENDIX H: FINDING THE RATE AND SPIKE AMPLITUDE OF A POISSON SPIKE TRAIN TO ELICIT A GIVEN AVERAGE AND STANDARD DEVIATION OF THE MEMBRANE POTENTIAL

To obtain the analytic expression for the value of current needed to elicit a given postsynaptic potential, the equation describing the LIF neuron model's subthreshold dy-

TABLE I. Main model parameters of the reference model. The term  $\mathcal{U}(n_1, n_2)$  denotes a uniform distribution within the range  $[n_1, n_2]$ , whereas  $\mathcal{D}(\mu, \sigma)$  represents a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

| Network and connectivity |                            |                                                                       |
|--------------------------|----------------------------|-----------------------------------------------------------------------|
| Name                     | Value                      | Description                                                           |
| $f$                      | 0.1                        | Fraction of neurons in a synaptic population                          |
| $p$                      | 5                          | Number of selective populations                                       |
| $c$                      | 0.2                        | Connectivity level                                                    |
| $N_E$                    | 8000                       | Number of excitatory neurons                                          |
| $N_I$                    | 2000                       | Number of inhibitory neurons                                          |
| Synaptic parameters      |                            |                                                                       |
| Name                     | Value                      | Description                                                           |
| $J_b$                    | 0.10 mV                    | Baseline weight (PSP amplitude) for excitatory to excitatory synapses |
| $J_p$                    | 0.45 mV                    | Potentiated weight of excitatory to excitatory synapses               |
| $J_{IE}$                 | 0.135 mV                   | Weight of excitatory to inhibitory synapses                           |
| $J_{EI}$                 | 0.25 mV                    | Weight of inhibitory to excitatory synapses                           |
| $J_{II}$                 | 0.2 mV                     | Weight of inhibitory to inhibitory synapses                           |
| $\hat{\delta}$           | $\mathcal{U}(0.1, 1.0)$ ms | Synaptic delays                                                       |
| $U$                      | 0.19                       | Baseline STP utilization factor                                       |
| $u$                      | $\mathcal{D}(U, 0.05)$     | Initial fraction of resources released for a spike                    |
| $x$                      | $\mathcal{U}(0, 1.0)$      | Initial fraction of resources ready to be released                    |
| $\tau_f$                 | 1500 ms                    | Facilitation time constant                                            |
| $\tau_d$                 | 200 ms                     | Depression time constant                                              |

namics and the postsynaptic current's behavior are needed. In particular, the subthreshold dynamics is described by

$$\tau_m \frac{dV}{dt} = -V + RI, \quad (H1)$$

and the input current of a Poisson spike train targeting a LIF neuron with exponential postsynaptic current is:

$$I(t) = \omega \sum_{i=1}^{N_S} e^{-\frac{t-t_i}{\tau_s}} \Theta(t - t_i). \quad (H2)$$

Here  $\tau_s$  and  $\tau_m$  represent the synaptic and the membrane time constants, respectively, while  $R = \tau_m / C_m$ . In the following, we denote by  $N(t)$  the number of spikes that have reached the neuron at time  $t$ . In this way we define  $N_S = N(T)$ , with  $T$  being the total time of the simulation.

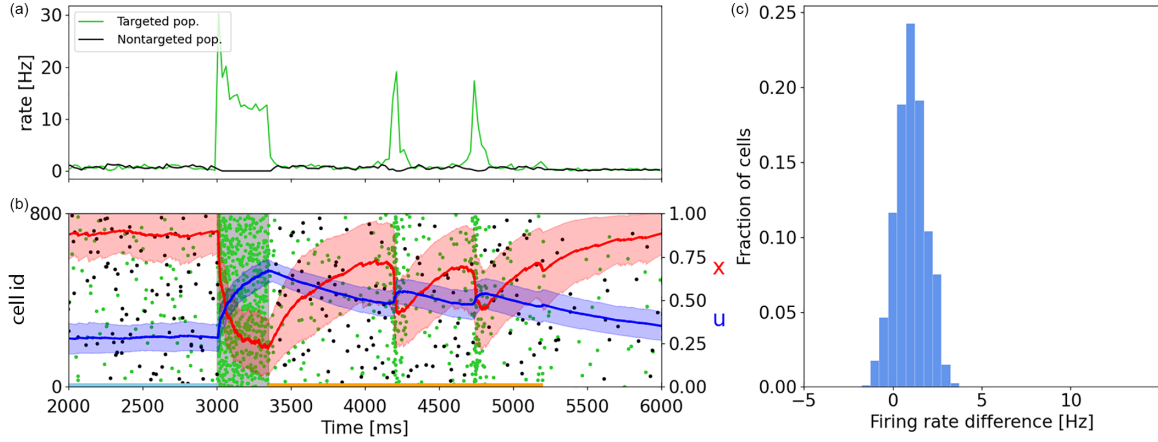


FIG. 12. Simulation with 50% of facilitated synapses. (a) Average firing rate for the targeted selective population and for one of the nontargeted selective populations. (b) Raster plot of a subset of the two selective populations. Averaged short-term plasticity variables  $x$  (red line) and  $u$  (blue line) for the synaptic connections within the targeted population are also shown. Bands in the same color code represent the standard deviation of the values of  $x$  and  $u$ . (c) Firing rate difference between delay period and standard activity.

Within the interval  $(t_j, t_{j+1})$ , we seek a particular solution  $t \in (t_j, t_{j+1})$ :  
of the differential equation of the form

$$V_{m,j}^{(\text{part})}(t) = \sum_{i=1}^j A_{ji} e^{-(t-t_i)/\tau_s}, \quad (\text{H3})$$

$$\sum_{i=1}^j \left[ A_{ji} \left( 1 - \frac{\tau_m}{\tau_s} \right) - \omega R \right] e^{-(t-t_i)/\tau_s} = 0, \quad (\text{H4})$$

where  $A_{ji}$  denotes the amplitude coefficient associated with the  $i$ -th spike contributing to the membrane potential at time  $t_j$ , to be determined in the following steps. Substituting it into the equation, we obtain the following relation for all

where we used the fact that

$$\begin{aligned} \Theta(t - t_i) &= 1 & \text{for } i \leq j, \\ \Theta(t - t_i) &= 0 & \text{for } i > j, \end{aligned}$$

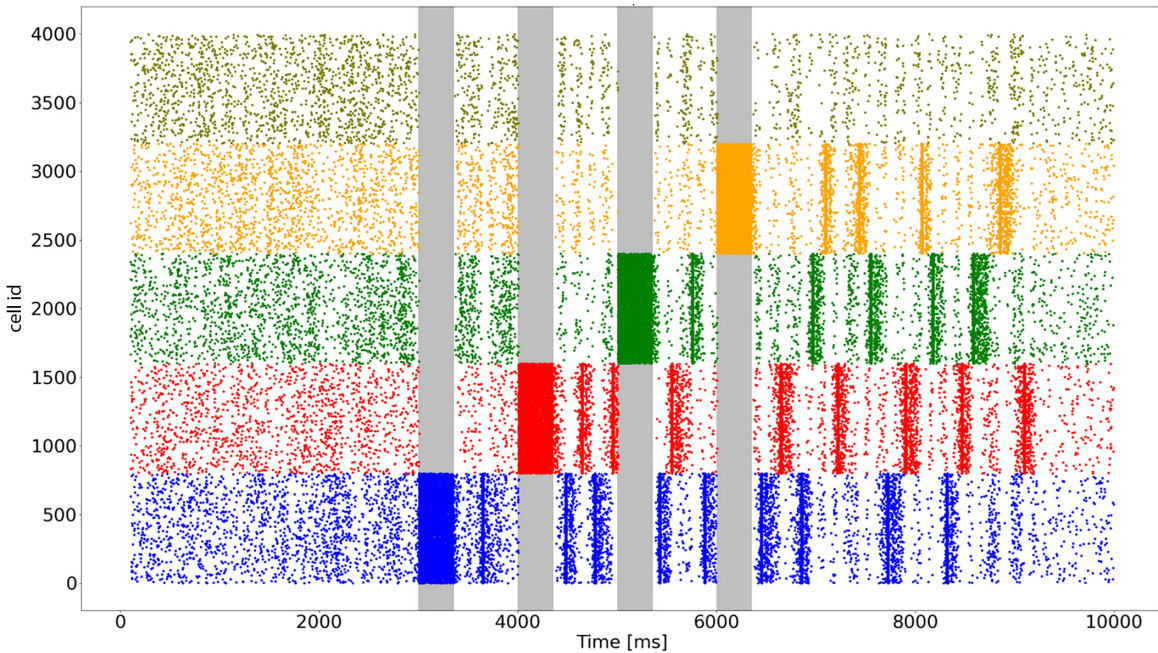


FIG. 13. Raster plot of excitatory neurons belonging to five selective populations during the working memory task. Gray bars indicate the time windows of item presentation (loading phase). Each color corresponds to a different selective population. After the sequential loading of all items, population-specific reverberatory activity emerges, maintaining active representations of the stored items through short-term synaptic facilitation.

where  $t_j < t < t_{j+1}$ . Solving Eq. (H4) for every  $i = 1, \dots, j$  we get

$$A_{ji} = \frac{\omega R \tau_s}{\tau_s - \tau_m}, \quad (\text{H5})$$

from which we find our particular solution as

$$V_{m,j}^{(\text{part})}(t) = \frac{\omega R \tau_s}{\tau_s - \tau_m} \sum_{i=1}^j e^{-(t-t_i)/\tau_s}. \quad (\text{H6})$$

The general solution in the interval  $(t_j, t_{j+1})$  is then of the form

$$V_{m,j}(t) = B_j e^{-t/\tau_m} + V_{m,j}^{(\text{part})}(t), \quad (\text{H7})$$

where  $B_j$  is an integration constant yet to be determined by applying the continuity conditions at the boundaries. Without any loss of generality, we can redefine  $B_j$  as

$$B_j = \sum_{i=1}^j K_i e^{t_i/\tau_m}, \quad (\text{H8})$$

where  $K_i$  are constant coefficients to be determined. Thus, Eq. (H7) can be rewritten as

$$\begin{aligned} V_{m,j}(t) &= \sum_{i=1}^j \left[ K_i e^{-\frac{(t-t_i)}{\tau_m}} + \frac{\omega R \tau_s}{\tau_s - \tau_m} e^{-\frac{(t-t_i)}{\tau_s}} \right] \\ &= \sum_{i=1}^{N_S} \left[ K_i e^{-\frac{(t-t_i)}{\tau_m}} + \frac{\omega R \tau_s}{\tau_s - \tau_m} e^{-\frac{(t-t_i)}{\tau_s}} \right] \Theta(t - t_i), \end{aligned} \quad (\text{H9})$$

where the last step is justified by the fact that  $t_j < t < t_{j+1}$ . Enforcing the continuity condition

$$\lim_{t \rightarrow t_{j+1}} V_{m,j}(t) = \lim_{t \rightarrow t_{j+1}} V_{m,j+1}(t), \quad (\text{H10})$$

which matches the condition

$$\lim_{t \rightarrow t_j} \left[ K_j e^{-(t-t_j)/\tau_m} + \frac{\omega R \tau_s}{\tau_s - \tau_m} e^{-(t-t_j)/\tau_s} \right] = 0, \quad (\text{H11})$$

allows us to determine the value of  $K_i$ :

$$K_i = -\frac{\omega R \tau_s}{\tau_s - \tau_m}. \quad (\text{H12})$$

Given that  $R = \tau_m/C_m$ , we are finally able to write the general solution of Eq. (H1) as

$$V(t) = \frac{\tau_m}{C_m} \frac{\tau_s}{\tau_s - \tau_m} \omega \sum_{i=1}^{N_S} \left( e^{-\frac{t-t_i}{\tau_s}} - e^{-\frac{t-t_i}{\tau_m}} \right) \Theta(t - t_i). \quad (\text{H13})$$

To compute the mean and variance of the membrane potential, we have to define the probability density  $\rho(t_1, \dots, t_{N_S})$  of having a total number of spikes  $N(T) = N_S$  occurring at times  $t_1, \dots, t_{N_S}$ . Each of the events is independent, and the probability of its occurrence is uniformly distributed over the interval  $[0, T]$ . However, we need to consider that the probability of getting  $N_S$  spikes in that interval follows a Poisson distribution with rate  $\lambda$ . With this in mind, we can compute the average of a generic function  $f(t_1, \dots, t_{N_S})$  as

$$\begin{aligned} \langle f \rangle_T &= \sum_{N_S=0}^{\infty} P(N_S) \int_0^T dt_1 \dots \\ &\times \int_0^T dt_{N_S} \rho_1(t_1) \dots \rho_{N_S}(t_{N_S}) f(t_1, \dots, t_{N_S}), \end{aligned} \quad (\text{H14})$$

where

$$\begin{aligned} P(N_S) &= e^{-\lambda T} \frac{(\lambda T)^{N_S}}{N_S!} \quad \text{and} \\ \rho_i(t_i) &= \rho_i(t_i; T) = \frac{1}{T} \quad \forall i = 1, \dots, N_S, \end{aligned} \quad (\text{H15})$$

with  $\rho_i$  being the probability density over  $t_i$ , parameterized by the total time interval  $T$ , and where we used the fact that the  $\rho_i$  are uniform over  $T$ . By taking into account also the time average over the interval  $[0, T]$  for the time-dependent solution (H13), we compute

$$\begin{aligned} \langle V(t) \rangle_T &= \frac{1}{T} \int_0^T dt \sum_{N_S=0}^{\infty} P(N_S) \int_0^T \frac{dt_1}{T} \dots \int_0^T \frac{dt_{N_S}}{T} V(t, t_i) \\ &= \frac{\tau_m}{C_m} \frac{\tau_s}{\tau_s - \tau_m} \omega \sum_{N_S=0}^{\infty} P(N_S) \int_0^T \frac{dt}{T} \int_0^T \frac{dt_1}{T} \dots \int_0^T \frac{dt_{N_S}}{T} \sum_{i=1}^{N_S} \left( e^{-\frac{t-t_i}{\tau_s}} - e^{-\frac{t-t_i}{\tau_m}} \right) \Theta(t - t_i). \end{aligned} \quad (\text{H16})$$

We now note that since every  $t_i$  is integrated, the terms of the sum are all equal, thus we can write

$$\begin{aligned} \langle V(t) \rangle_T &= \frac{\tau_m}{C_m} \frac{\tau_s}{\tau_s - \tau_m} \omega \sum_{N_S=0}^{\infty} P(N_S) N_S \frac{1}{T^2} \int_0^T dt \int_0^t dt_1 \left[ e^{-\frac{t-t_1}{\tau_s}} - e^{-\frac{t-t_1}{\tau_m}} \right] \\ &= \frac{\tau_m}{C_m} \frac{\tau_s}{\tau_s - \tau_m} \omega \frac{\lambda}{T} \int_0^T dt \left[ \tau_s (1 - e^{-\frac{t}{\tau_s}}) - \tau_m (1 - e^{-\frac{t}{\tau_m}}) \right], \end{aligned} \quad (\text{H17})$$

where we used the relation  $\langle N_S \rangle = \lambda T$ . Since we are interested in the limit  $T \rightarrow \infty$ , the only surviving contribution is the one from the integral that is proportional to  $T$ , leading to

$$\lim_{T \rightarrow \infty} \langle V(t) \rangle_T = \frac{\tau_m}{C_m} \omega \lambda \tau_s. \quad (\text{H18})$$

So, we find that the average membrane potential is proportional to the mean current given by the Poisson spike train.

In order to obtain the variance of the membrane potential we now calculate the mean of  $V(t)^2$ , which we can write as

$$V(t)^2 = \alpha^2 \omega^2 \sum_{i=0}^{N_S} \sum_{j=0}^{N_S} \gamma_i(t) \gamma_j(t) \Theta(t - t_i) \Theta(t - t_j), \quad (\text{H19})$$

where we have defined  $\gamma_i(t) \equiv (e^{-(t-t_i)/\tau_s} - e^{-(t-t_i)/\tau_m})$  and  $\alpha \equiv (\frac{\tau_m}{C_m} \frac{\tau_s}{\tau_s - \tau_m})$ . By employing Eq. (H14) we get

$$\langle V(t)^2 \rangle_T = \alpha^2 \omega^2 \sum_{N_S=0}^{\infty} P(N_S) \int_0^T \frac{dt}{T} \int_0^T \frac{dt_1}{T} \cdots \int_0^T \frac{dt_{N_S}}{T} \left[ \sum_{i=1}^{N_S} \gamma_i^2(t) \Theta(t - t_i) + \sum_{i \neq j} \gamma_i(t) \gamma_j(t) \Theta(t - t_i) \Theta(t - t_j) \right]. \quad (\text{H20})$$

As noted earlier, since all  $t_i$  are integrated, we can write

$$\begin{aligned} \langle V(t)^2 \rangle_T &= \alpha^2 \omega^2 \sum_{N_S=0}^{\infty} P(N_S) \left[ N_S \frac{1}{T^2} \int_0^T dt \int_0^t dt_1 \gamma_1^2(t) + N_S(N_S - 1) \frac{1}{T^3} \int_0^T dt \left( \int_0^t dt_1 \gamma_1(t) \right) \left( \int_0^t dt_2 \gamma_2(t) \right) \right] \\ &= \alpha^2 \omega^2 \left[ \lambda \frac{1}{T} \int_0^T dt \int_0^t dt_1 \gamma_1^2(t) + \lambda^2 \frac{1}{T} \int_0^T dt \left( \int_0^t dt_1 \gamma_1(t) \right) \left( \int_0^t dt_2 \gamma_2(t) \right) \right], \end{aligned} \quad (\text{H21})$$

where the last step follows from using  $\langle N_S \rangle = \lambda T$  and  $\langle N_S^2 \rangle = \lambda T + \lambda^2 T^2$ . Keeping just the terms proportional to  $T$  from the integrations, we evaluate

$$\int_0^T dt \int_0^t dt_1 \gamma_1^2(t) \rightarrow \frac{(\tau_s - \tau_m)^2}{2(\tau_s + \tau_m)} T, \quad (\text{H22})$$

$$\int_0^T dt \left( \int_0^t dt_1 \gamma_1(t) \right) \left( \int_0^t dt_2 \gamma_2(t) \right) \rightarrow (\tau_s - \tau_m)^2 T, \quad (\text{H23})$$

so that

$$\lim_{T \rightarrow \infty} \langle V(t)^2 \rangle_T = \alpha^2 \omega^2 \frac{(\tau_s - \tau_m)^2}{2(\tau_s + \tau_m)} [\lambda + 2\lambda^2(\tau_s + \tau_m)]. \quad (\text{H24})$$

Now we can finally get the variance of the membrane potential from (H18) and (H24) as  $\Sigma_V^2 = \langle V(t)^2 \rangle - \langle V(t) \rangle^2$ :

$$\Sigma_V^2 = \left( \frac{\tau_m}{C_m} \right)^2 \frac{\tau_s^2}{2(\tau_s + \tau_m)} \omega^2 \lambda. \quad (\text{H25})$$

Thus, we can use Eqs. (H18) and (H25) to derive the rate of the Poisson process  $\lambda$  and the amplitude of the spikes  $\omega$  needed to elicit an average membrane potential with determined mean and standard deviation.

We can easily derive that

$$\lambda = \frac{\langle V \rangle^2}{\Sigma_V^2} \frac{1}{2(\tau_s + \tau_m)} \quad \text{and} \quad \omega = \frac{\langle V \rangle C_m}{\lambda \tau_m \tau_s}. \quad (\text{H26})$$

In our simulations using NEST, each neuron is connected to a `poisson_generator` device having a rate equal to  $\lambda$  and connected with the neuron through a synapse with weight  $\omega$ .

- 
- [1] G. A. Miller, E. Galanter, and K. H. Pribram, *Plans and the Structure of Behavior* (Henry Holt and Co., New York, NY, 1960).
  - [2] A. D. Baddeley and G. Hitch, *Working Memory* (Academic Press, 1974), pp. 47–89.
  - [3] N. Cowan, *Attention and Memory* (Oxford University Press, Oxford, UK, 1998).
  - [4] B. Golosio, A. Cangelosi, O. Gamotina, and G. L. Masala, A cognitive neural architecture able to learn and communicate through natural language, *PLoS One* **10**, e0140866 (2015).
  - [5] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex, *J. Neurophysiol.* **61**, 331 (1989).
  - [6] P. Goldman-Rakic, Cellular basis of working memory, *Neuron* **14**, 477 (1995).
  - [7] M. D'Esposito and B. R. Postle, The cognitive neuroscience of working memory, *Annu. Rev. Psychol.* **66**, 115 (2015).
  - [8] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, NY, 1949).
  - [9] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982).
  - [10] N. Brunel, Persistent activity and the single cell frequency-current curve in a cortical network model, *Netw. Comput. Neural Syst.* **11**, 261 (2000).
  - [11] M. G. Stokes, 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework, *Trends Cognit. Sci.* **19**, 394 (2015).
  - [12] R. Honkanen, S. Rouhinen, S. H. Wang, J. M. Palva, and S. Palva, Gamma oscillations underlie the maintenance of feature-specific information and the contents of visual working memory, *Cereb. Cortex* **25**, 3788 (2015).
  - [13] M. Lundqvist, J. Rose, P. Herman, S. L. Brincat, T. J. Buschman, and E. K. Miller, Gamma and beta bursts underlie working memory, *Neuron* **90**, 152 (2016).
  - [14] O. Barak and M. Tsodyks, Persistent activity in neural networks with dynamic synapses, *PLoS Comput. Biol.* **3**, e35 (2007).
  - [15] G. Mongillo, O. Barak, and M. Tsodyks, Synaptic theory of working memory, *Science* **319**, 1543 (2008).

- [16] E. T. Rolls, L. Dempere-Marco, and G. Deco, Holding multiple items in short term memory: A neural mechanism, *PLoS ONE* **8**, e61078 (2013).
- [17] D. Hansel and G. Mato, Short-term plasticity explains irregular persistent activity in working memory tasks, *J. Neurosci.* **33**, 133 (2013).
- [18] N. E. Myers, M. G. Stokes, and A. C. Nobre, Prioritizing information during working memory: Beyond sustained internal attention, *Trends Cognit. Sci.* **21**, 449 (2017).
- [19] J. Barbosa, H. Stein, R. L. Martinez, A. Galan-Gadea, S. Li, J. Dalmau, K. C. S. Adam, J. Valls-Solé, C. Constantinidis, and A. Compte, Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory, *Nat. Neurosci.* **23**, 1016 (2020).
- [20] R. Douglas Fields, The enigma of working memory: Changing views, *Neuroscientist* **28**, 420 (2022).
- [21] D. Trübtschek, S. Marti, H. Ueberschär, and S. Dehaene, Probing the limits of activity-silent non-conscious working memory, *Proc. Natl. Acad. Sci. USA* **116**, 14358 (2019).
- [22] A. Seeholzer, M. Deger, and W. Gerstner, Stability of working memory in continuous attractor networks under the control of short-term plasticity, *PLoS Comput. Biol.* **15**, e1006928 (2019).
- [23] Y. Mi, M. Katkov, and M. Tsodyks, Synaptic correlates of working memory capacity, *Neuron* **93**, 323 (2017).
- [24] G. Tiddia, B. Golosio, V. Fanti, and P. S. Paolucci, Simulations of working memory spiking networks driven by short-term plasticity, *Front. Integr. Neurosci.* **16**, 972055 (2022).
- [25] H. Taher, A. Torcini, and S. Olmi, Exact neural mass model for synaptic-based working memory, *PLoS Comput. Biol.* **16**, e1008533 (2020).
- [26] Y. Wang, H. Markram, P. H. Goodman, T. K. Berger, J. Ma, and P. S. Goldman-Rakic, Heterogeneity in the pyramidal network of the medial prefrontal cortex, *Nat. Neurosci.* **9**, 534 (2006).
- [27] N. Perez-Nieves, V. Leung, P. Dragotti, and D. Goodman, Neural heterogeneity promotes robust learning, *Nat. Commun.* **12**, 5791 (2021).
- [28] R. Gast, S. A. Solla, and A. Kennedy, Neural heterogeneity controls computations in spiking neural networks, *Proc. Natl. Acad. Sci. USA* **121**, e2311885121 (2024).
- [29] J. Hass, S. Ardid, J. Sherfey, and N. Kopell, Constraints on persistent activity in a biologically detailed network model of the prefrontal cortex with heterogeneities, *Prog. Neurobiol.* **215**, 102287 (2022).
- [30] A. Renart, P. Song, and X.-J. Wang, Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks, *Neuron* **38**, 473 (2003).
- [31] M.-O. Gewaltig and M. Diesmann, NEST (NEural Simulation Tool), *Scholarpedia* **2**, 1430 (2007).
- [32] M. Shafi, Y. Zhou, J. Quintana, C. Chow, J. Fuster, and M. Bodner, Variability in neuronal activity in primate cortex during working memory tasks, *Neuroscience* **146**, 1082 (2007).
- [33] S. Raghavachari, M. J. Kahana, D. S. Rizzuto, J. B. Caplan, M. P. Kirschen, B. Bourgeois, J. R. Madsen, and J. E. Lisman, Gating of human theta oscillations by a working memory task, *J. Neurosci.* **21**, 3175 (2001).
- [34] H. Lee, G. V. Simpson, N. K. Logothetis, and G. Rainer, Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex, *Neuron* **45**, 147 (2005).
- [35] Y. Xie, Y. Li, H. Duan, X. Xu, W. Zhang, and P. Fang, Theta oscillations and source connectivity during complex audiovisual object encoding in working memory, *Front. Hum. Neurosci.* **15**, 614950 (2021).
- [36] J.-P. Thivierge and P. Cisek, Nonperiodic synchronization in heterogeneous networks of spiking neurons, *J. Neurosci.* **28**, 7968 (2008).
- [37] J.-N. Teramae and T. Fukai, Computational implications of lognormally distributed synaptic weights, *Proc. IEEE* **102**, 500 (2014).
- [38] Y. Loewenstein, A. Kuras, and S. Rumpel, Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex *in vivo*, *J. Neurosci.* **31**, 9481 (2011).
- [39] G. Buzsáki and K. Mizuseki, The log-dynamic brain: How skewed distributions affect network operations, *Nat. Rev. Neurosci.* **15**, 264 (2014).
- [40] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, Highly nonrandom features of synaptic connectivity in local cortical circuits, *PLoS Biol.* **3**, e68 (2005).
- [41] S. E. Motley, Relationship between neuromodulation and working memory in the prefrontal cortex: It's complicated, *Front. Neural Circuits* **12**, 31 (2018).
- [42] S. Raghavachari, J. E. Lisman, M. Tully, J. R. Madsen, E. B. Bromfield, and M. J. Kahana, Theta oscillations in human cortex during a working-memory task: Evidence for local generators, *J. Neurophysiol.* **95**, 1630 (2006).
- [43] T. Elbert, W. J. Ray, Z. J. Kowalik, J. E. Skinner, K. E. Graf, and N. Birbaumer, Chaos and physiology: Deterministic chaos in excitable cell assemblies, *Physiol. Rev.* **74**, 1 (1994).
- [44] J. F. Mejías and X.-J. Wang, Mechanisms of distributed working memory in a large-scale network of macaque neocortex, *eLife* **11**, e72136 (2022).
- [45] L. Ye, J. Feng, and C. Li, Controlling brain dynamics: Landscape and transition path for working memory, *PLoS Comput. Biol.* **19**, e1011446 (2023).
- [46] E. Hunsberger, M. Scott, and C. Eliasmith, The competing benefits of noise and heterogeneity in neural coding, *Neural Comput.* **26**, 1600 (2014).
- [47] E. Marder and J.-M. Goaillard, Variability, compensation and homeostasis in neuron and network function, *Nat. Rev. Neurosci.* **7**, 563 (2006).
- [48] C. M. Hempel, K. H. Hartman, X.-J. Wang, G. G. Turrigiano, and S. B. Nelson, Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex, *J. Neurophysiol.* **83**, 3031 (2000).
- [49] G. Tiddia, Large-scale neuronal networks: From simulation technology to the study of plasticity-driven cognitive processes, Ph.D. thesis, University of Cagliari, 2024.
- [50] J. Senk, B. Kriener, M. Djurfeldt, N. Voges, H.-J. Jiang, L. Schüttler, G. Gramelsberger, M. Diesmann, H. E. Plesser, and S. J. van Albada, Connectivity concepts in neuronal network modeling, *PLoS Comput. Biol.* **18**, e1010086 (2022).
- [51] E. Nordlie, M.-O. Gewaltig, and H. E. Plesser, Towards reproducible descriptions of neuronal network models, *PLoS Comput. Biol.* **5**, e1000456 (2009).
- [52] C. Linssen, P. N. Babu, J. M. Eppler, L. Koll, B. Rumpel, and A. Morrison, NESTML: A generic modeling language and code generation tool for the simulation of spiking neural

- networks with advanced plasticity rules, *Front. Neuroinform.* **19**, 1544143 (2025).
- [53] C. Linssen, P. N. Babu, Y. Bouhadjar, L. Ewert, W. Wybo, M. Lober, F. Feller, B. Rumpe, and A. Morrison, NESTML 8.0.0 (Version 8.0.0) [Computer software], Zenodo, 2024, <https://doi.org/10.5281/zenodo.12191059>.
- [54] G. Tiddia, L. Sergi, S. Rubiu, A. Incollu, and B. Golosio, Short-term plasticity-based working memory spiking model is resilient to synaptic heterogeneity, Zenodo, 2025, <https://doi.org/10.5281/zenodo.17991881>.
- [55] M. Tsodyks, K. Pawelzik, and H. Markram, Neural networks with dynamic synapses, *Neural Comput.* **10**, 821 (1998).