Cognitive Computation (2024) 16:756–775
https://doi.org/10.1007/s12559-023-10234-4

# Synaptic Facilitation: A Key Biological Mechanism for Resource Allocation in Computational Models of Working Memory

**Marta Balagué-Marmaña[1] · Laura Dempere-Marco[1,2]**

Received: 13 April 2023 / Accepted: 2 December 2023 / Published online: 28 December 2023


**Abstract**
Working memory (WM) is a crucial cognitive function required to maintain and manipulate information that is no longer present through the senses. Two key features of WM are its limited capacity and the emergence of serial order effects. This study investigates how synaptic facilitation and diverse display dynamics influence the encoding and retention of multiple items in WM. A biophysically inspired attractor model of WM, endowed with synaptic facilitation, is considered in this study. The investigation delves into the behaviour of the model under both sequential and simultaneous display protocols. Synaptic facilitation plays a crucial role in establishing the response of the WM system by regulating resource allocation during the encoding stage. It boosts WM capacity and is a key mechanism in the emergence of serial order effects. The synaptic facilitation time constant ($\tau_F$) is critical in modulating these effects, and its heterogeneity in the prefrontal cortex (PFC) may contribute to the combination of primacy and recency effects observed experimentally. Additionally, we demonstrate that the WM capacity exhibited by the network is heavily influenced by factors such as the stimuli nature, and their display duration. Although the network connectivity determines the WM capacity by regulating the excitation-inhibition balance, the display protocol modulates its effective limit. Our findings shed light on how different stimulation protocol dynamics affect WM, underscoring the importance of synaptic facilitation and experimental protocol design in modulating WM capacity.

**Keywords** Working memory · Synaptic facilitation · Attractor networks · Serial order effects

## Introduction

Working memory (WM) is a cognitive function which is necessary to maintain and manipulate information that is not present physically through the senses. Its integrity is basic for higher cognitive functions, such as language, memory or reasoning. A hallmark property of WM is its limited capacity [1]. In the last decades, several theories have explored these capacity limits by not only paying attention to such absolute bounds but also to the accuracy with which the items are memorised [2–5]. Although all of these models have successfully accounted for a broad range of psychophysical results, little neurophysiological evidence about the neural mechanisms underlying these predictions is available.

Selectively enhanced activity of neurons in the prefrontal cortex (PFC) throughout the delay period of WM tasks has been traditionally regarded as a neural correlate of WM function [6]. Indeed, persistent activity has been reported in multiple studies (see [7] for a review on the topic). However, other authors rely on alternative mechanisms such as spiking rhythmicity [8], or short-term synaptic plasticity [9].

Although the neural mechanisms underlying the maintenance of multiple items in WM have not been clearly identified yet, several hypotheses have been considered from the computational perspective, which match the three mechanisms outlined earlier: (1) sustained neural activation (e.g. single-item WM [10, 11], and multi-item WM [12–16]), (2) neural oscillations (e.g. [8, 17]), or (3) patterns of synaptic strength [9].

Persistent neuronal activity can manifest in the brain through various mechanisms. It may arise at the network level, as a result of recurrent connections within a reverbera-

✉ Laura Dempere-Marco
laura.dempere@uvic.cat

Marta Balagué-Marmaña
marta.balague@uvic.cat

[1] Department of Engineering, University of Vic-Central University of Catalonia (UVic-UCC), Vic, Barcelona, Spain

[2] Institute for Research and Innovation in Life and Health Sciences in Central Catalonia (IRIS-CC), Vic, Barcelona, Spain

tory neural network, or at the cellular level, driven by intrinsic mechanisms that operate independently of network involvement [11, 18, 19]. In this study, delay activity is interpreted in terms of the attractor dynamics exhibited in neural networks. Within the attractor picture, persistent activity is characterised by the elevated firing of a subset of neurons, which remains even without external stimulation. The attractors, nonetheless, could also be realised in the synaptic dynamics [9] instead of the (persistent) neuronal activities. As such, attractor dynamics offer a fundamental framework to investigate both activity-based and activity-silent mnemonic regimes. In contrast to the activity-based models, Mongillo et al. [9] proposed a computational model which explores the hypothesis that persistent neural activity may not be the only mechanism underlying WM. They suggest that WM is sustained by calcium-mediated synaptic facilitation in the recurrent connections of neocortical networks, a mechanism that is metabolically efficient and robust. Unsurprisingly, however, persistent activity models have been largely prevalent in the field provided their direct relation with the available neurophysiological recordings. Although we do not claim that alternative mechanisms are not plausible (or even likely), in this work, we propose a persistent activity model and argue that it accounts for a variety of relevant experimental findings.

From a different perspective, it is clear that our perceptual reality is generally far from static. This is a consequence of the continuous influx of changing stimuli entering through our senses. While some of these stimuli appear simultaneously in the physical world, others appear asynchronously. Indeed, the storage and manipulation of serial order information is a critical element of WM, which is key for success in many different cognitive tasks, including numerical tasks [20], language production [21], or visual perception [22]. However, systematic investigations into the role of stimulation dynamics in determining both the capacity limits of WM and the characteristics of the resulting serial order effects have been relatively scarce [23–26].

In this research, our primary focus will be on the neuronal dynamics exhibited by the network. It is noteworthy that this temporal dimension contributes to the emergence of serial effects and carries significant implications not only for determining the quantity of displayed items retained in WM but also for identifying the items that are preferentially preserved. Throughout the remainder of the paper, the term *sequential stimulation* will refer to the serial display of stimuli, each presented individually with an interstimulus interval (ISI) between consecutive items. Notably, two serial effects, namely *primacy* and *recency*, have been experimentally reported (e.g. [27, 28]). In sequential display protocols, the recency effect implies that items presented in the final positions of a sequence are more likely to be retained in WM, while the primacy effect favours items shown earlier in the sequence.

Remarkably, serial effects have been observed across multiple WM tasks. These tasks encompass diverse test stages like free recall and recognition, various stimulus modalities such as verbal, spatial, and object-based WM, and different temporal dynamics in the presentation of items. A comprehensive examination of these serial effects reveals a clear primacy and recency pattern for verbal material [28]. Conversely, when dealing with spatial material, the recency effect is evident but primacy seems to be limited in this case [28]. In the visual domain, results remain inconclusive. Several studies report a recency effect for the last stimulus but do not observe primacy when using abstract visual patterns [29, 30], as well as other types of stimuli such as unfamiliar faces, inverted faces or non-work sequences [31–33]. In contrast, some studies demonstrate both primacy and recency effects for unfamiliar faces [31]. Indeed, Hurlstone et al. [27], in their review on this subject, emphasise the limited availability of information regarding the emergence of the primacy effect within the visual domain.

Similarly, distinct WM capacity limits have been documented across various WM tasks, with reported values ranging from 3 to 7 items [23, 34]. Altogether, these myriad experimental paradigms have contributed a wealth of empirical findings regarding WM. However, this heterogeneity has also posed new challenges, as it necessitates the identification of consistent findings across all of these paradigms to contribute fundamental insights into the mechanisms underlying WM [28].

In this work, we consider a minimal, biophysically inspired model of multi-item WM endowed with synaptic facilitation to investigate its underlying neural mechanisms in the context of these heterogeneous results. Furthermore, we evaluate its performance under different stimulation protocols. Hitherto, the influence of the stimulation protocol on the neuronal dynamics during the encoding stage, and the subsequent maintenance of items in WM has remained unclear, despite its likely significant impact on establishing an effective WM capacity ($K_e$) [13]. We hypothesise that the emergence of serial order effects is linked to the time scale of biophysically relevant processes involved in the neuronal dynamics, and that the design of the experimental protocols significantly influences the encoding stage of WM.

## Materials and Methods

### Experimental Protocols

Particular emphasis is placed on the dynamics of the encoding stage of the experimental protocols, which mainly encompasses the period when the memory set is displayed. In the proposed model, these dynamics are inherently embedded within the stimulation protocols. We consider both simulta-

neous and sequential stimulation protocols (see Fig. 1). In simultaneous protocols, all items are displayed simultaneously, while in sequential protocols, each item is presented individually with a time interval, known as ISI, separating contiguous items. Although a few studies have explicitly examined and compared both types of stimulation protocols [23–26], the prevailing trend in the literature is to predominantly focus on one of them. In this study, we consider an experimental paradigm derived from a delay-match-to-sample (DMS) task. In this paradigm, the memory set is displayed during a specific time interval ($\Delta t_{stim}$). In simultaneous protocols, $\Delta t_{stim}$ corresponds to the period during which all items are displayed concurrently, while in sequential protocols, each item is shown for $\Delta t_{stim}$, followed by an interstimulus interval (ISI) before the presentation of the next item. After the presentation of all stimuli, there is a delay period of $\Delta t_{delay}$. WM is probed during the final $\Delta t_{WM}$ of this delay period. In this study, $\Delta t_{delay}$ is set to 3 s, and $\Delta t_{WM}$ is 500 ms.

## Computational Model

In order to investigate the neural mechanisms underlying WM, we propose a biophysically inspired attractor network of object WM with spiking neurons. For simplicity, the level of accuracy with which an item is kept in WM has not been specifically addressed. Consequently, a discrete attractor network is considered. The spiking neural network considered in this study is based on the model proposed by Brunel and Wang [11] for single-item object WM, which was subsequently extended to multiple items [13, 14]. It consists of a network structured into statistically homogeneous neural populations. In particular, the statistical properties of the synaptic currents and the connection strengths are identical for all the neurons within the same population. There is one population of inhibitory cells and one population of excitatory cells, which is partitioned into $N$ subpopulations. The excitatory population is divided into 10 subpopulations, $P_1$ to $P_{10}$, with each subpopulation being selectively responsive to a particular object $i$. Each of them represents one memory by maintaining its activity during a delay period after a cue $\lambda_i$ has been applied during the encoding stage.

Recurrent connections between neurons from the same selective subpopulation are potentiated by a factor $\omega_+ > 1$ with respect to the baseline connectivity level, while connections between neurons from different selective subpopulations are weakened by a factor $0 < \omega_- < 1$. The strength of inhibitory-to-excitatory connections is denoted by the weight $\omega_{inh}$. The integrate-and-fire spiking network contains 1000 neurons (80% excitatory and 20% in the inhibitory pool). Each neuron in the network receives external Poisson inputs $\lambda_{ext}$ from 800 external neurons at a rate of 3.05 Hz/synapse (unless otherwise stated) to simulate the effect of inputs coming from other brain areas. The attractor network architecture is shown in Fig. 2.

The behaviour of the neurons is modeled by means of the leaky integrate-and-fire (LIF) model, in which the membrane potential $V(t)$ obeys the following differential equation:

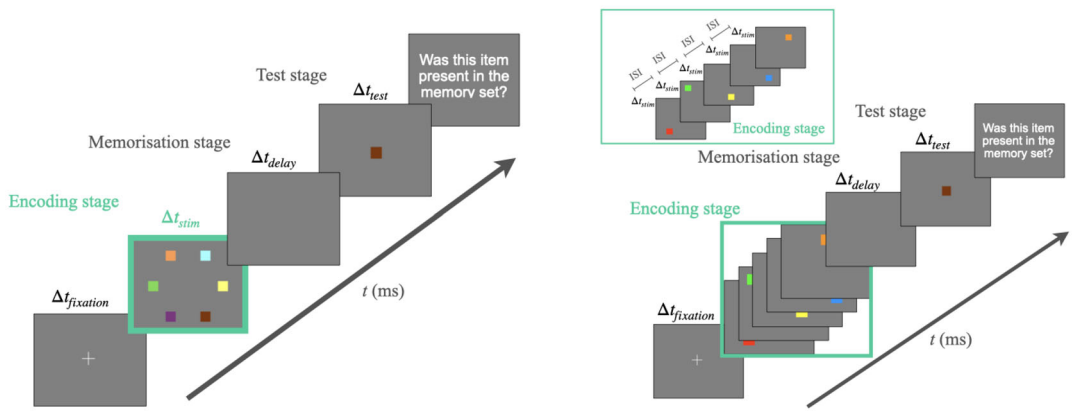


**Fig. 1** Multi-item WM experimental paradigms. Example of multi-item memory set displayed under two different experimental protocols in a DMS visual task. Following a $\Delta t_{fixation}$ period lasting 500 ms during which a central fixation cross is shown but no stimuli are displayed, the memory set is presented during the encoding stage. Then, the memory set is no longer shown during a $\Delta t_{delay}$ period lasting 3 s. This is the so-called memorisation stage. At the end of the memorisation stage, a test item is shown during $\Delta t_{test}$ (test stage). Experimentally, a memory test stage could be subsequently implemented by requesting the subjects to answer the following question: "Was this item present in the memory set?". The encoding stage of the two experimental protocols is different in the following way: (Left) Simultaneous protocol, all items are shown at the same time during $\Delta t_{stim}$, and (right) sequential protocol, the items are shown individually in a sequence during $\Delta t_{stim}$ (each item) with an ISI between items

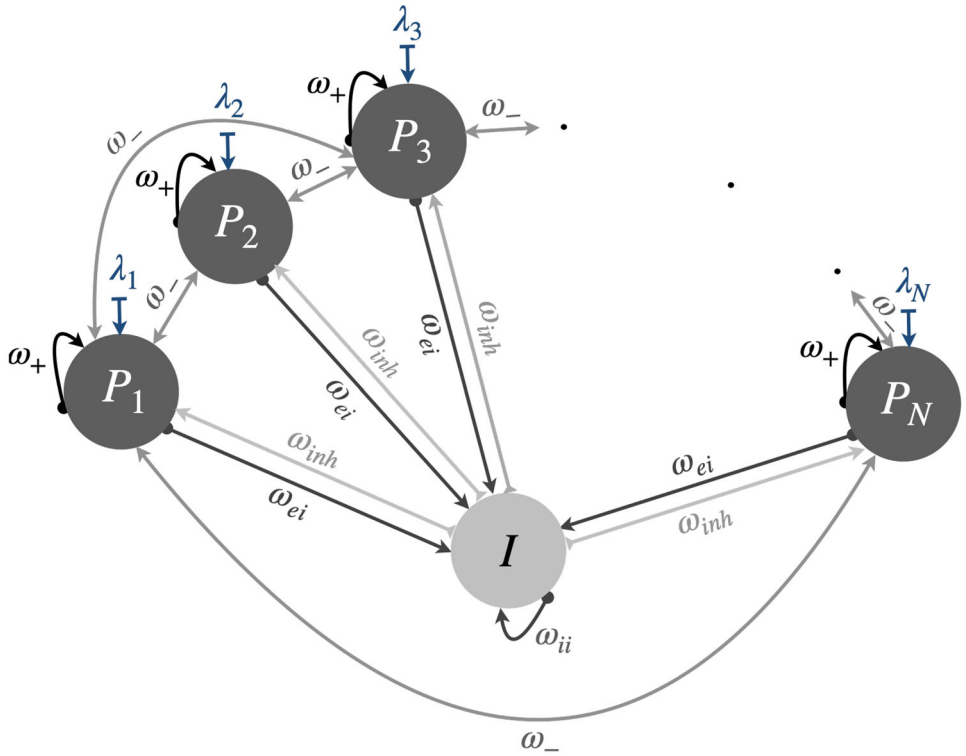


**Fig. 2** The attractor network model. The network is characterised by full connectivity, wherein the excitatory neurons are partitioned into $N$ selective pools or neuronal populations denoted as $P_1 - P_N$, with four of these pools illustrated in the diagram. Activation of these selective pools is driven by specific cues $\lambda_i$, where $i$ designates the individual pool. The synaptic connections within the pools have strengths consistent with principles of associative learning, particularly strong intra-pool connection strengths ($\omega_+$), and weaker between-pool synaptic connection strengths ($\omega_-$). The excitatory neurons receive inputs from the inhibitory neurons with synaptic connection strength $\omega_{inh}$. The remaining connection strengths are 1. $\omega_+$ was typically 2.3, $\omega_- = 0.87$, and $\omega_{inh} = 0.97$ (unless otherwise stated). The integrate-and-fire spiking network typically contained 1000 neurons, with 80 in each of the 10 non-overlapping excitatory pools, and 200 in the inhibitory pool. Each neuron in the network also receives external Poisson inputs $\lambda_{ext}$ from 800 external neurons at a typical rate of 3.05 Hz/synapse to simulate the effect of inputs coming from other brain areas

$$C_m \frac{dV(t)}{dt} = -g_L\Big(V(t) - V_L\Big) - I_{\text{syn}}(t) \tag{1}$$

where $C_m$ is the total membrane capacitance, $g_L$ is the passive conductance, $V_L$ is the resting potential, and $I_{\text{syn}}(t)$ is the synaptic current that charges the neuron. Four families of synapses are considered. The recurrent excitatory postsynaptic currents (EPSCs) have two components, which are mediated by AMPA and NMDA receptors. In contrast, only AMPA receptors mediate external EPSCs and GABA receptors mediate the inhibitory components. The total synaptic current is defined as follows:

$$\begin{aligned} I_{\text{syn}}(t) = {} & I_{\text{AMPA,ext}}(t) + I_{\text{AMPA,rec}}(t) \\ & + I_{\text{NMDA,rec}}(t) + I_{\text{GABA}}(t) \end{aligned} \tag{2}$$

The model is endowed with short-term plasticity [14], which relies on the properties of excitatory neurons in PFC and their facilitatory activity. STSF can be attributed, for instance, to the accumulation of residual calcium at the presynaptic terminals, which subsequently increases the likelihood of neurotransmitter release [35]. In this study, STSF has been implemented by adopting a phenomenological model based on calcium-mediated transmission principles [9]. The synaptic efficacy of recurrent connections among all excitatory neurons undergoes modulation by the utilisation parameter ($u$), which represents the fraction of resources used, and mirrors the calcium level. Upon the arrival of a spike at the presynaptic terminal, calcium influx leads to an increment in $u$. This, in turn, elevates the probability of transmitter release and thereby augments the synaptic strength at that particular synapse. The temporal decay of synaptic facilitation is regulated by the time constant $\tau_F$. This parameter determines how quickly the increased release probability returns to its baseline level after the spike has arrived. The model obeys the following equation [9]:

$$\frac{du_j(t)}{dt} = \frac{U - u_j(t)}{\tau_F} + U\Big(1 - u_j(t)\Big)\sum_k \delta(t - t_j^k) \tag{3}$$

The modulation by the utilisation factor is implemented by multiplying the synaptic weight of each recurrent excitatory synapse by $u$ to produce an effective synaptic weight $\omega_{eff} = u \cdot \omega_+$. The value for the baseline utilisation factor is $U$ ($U = 0.15$, in this study). Different time constants may reflect the diversity of mnemonic timescales observed in PFC. An essential aspect of the proposed mechanism is that STSF temporally (and dynamically) changes the functional connectivity of the network in a task-dependent manner.

The mean field analysis of the model, as described in [14], uncovers several critical aspects. On the one hand, the system can operate in a physiologically plausible regime in which multiple memories for those, and only those, items which have received external stimulation can be simultaneously held. On the other hand, STSF boosts the overall WM capacity of the network because of the effectively increased synaptic strengths of those pools to which the cues are applied, and the maintenance of this facilitation in just those pools when the cue is removed (as a result of their continuing neuronal firing). The network capacity $K$ is established by the maximum number of memories that can be simultaneously held upon stimulation. When the network is overloaded ($N_{stim} > K$) only some of the cued memories are correctly codified and held in WM. The network parameters considered throughout this work can be found in the Supplementary Material. The code that implements this model can be accessed at the following GitHub repository: https://github.com/mbalague-uvic/WMcode.

It is worth pointing out that we do not model the recognition stage, which takes place in the test stage. In this investigation, we consider that an item is maintained in WM if it shows persistent activity during the interval $\Delta t_{WM}$. The criterion employed to determine the retention of an item within WM is a mean firing rate $\bar{\nu} \geq 20$ Hz.

## Results

### Multi-item WM with Persistent Firing: Simultaneous *vs* Sequential Stimulation Protocols

The predictions of the proposed model are assessed in the light of the two stimulation protocols: simultaneous and sequential display of items. In the simultaneous stimulation protocol, after a 0.5 s period of spontaneous activity (with $\lambda_{ext}$ at the baseline level of 3.05 Hz/synapse), $\lambda_{item}$ cues $\lambda_1$ to $\lambda_{N_{stim}}$ are applied to excitatory neuron pools $P_1$ to $P_{N_{stim}}$ during the period 500–1500 ms. $\lambda_1 - \lambda_{N_{stim}}$ are applied by increasing the received external current from $\lambda_{ext}$ to $\lambda_{item}$ = 3.3125 Hz/synapse for just these $N_{stim}$ pools during the cue period. In the sequential stimulation protocol, also after a 0.5 s period of spontaneous activity with $\lambda_{ext}$ at the baseline level of 3.05 Hz/synapse, cues $\lambda_1$ to $\lambda_{N_{stim}}$ are sequentially applied to excitatory neuron pools $P_1$ to $P_{N_{stim}}$ during 1 s with an ISI of 1 s. The rest of pools remain at the baseline level of 3.05 Hz/synapse throughout the trial for both protocols. Note that the network parameters considered in this study, specifically those associated with the STSF mechanism, are similar to those considered in previous studies (e.g. [9, 14]). Figure 3 shows how the number of items successfully maintained in WM changes as a function of the memory set size ($S = N_{stim}$) for both stimulation protocols. Specifically, it

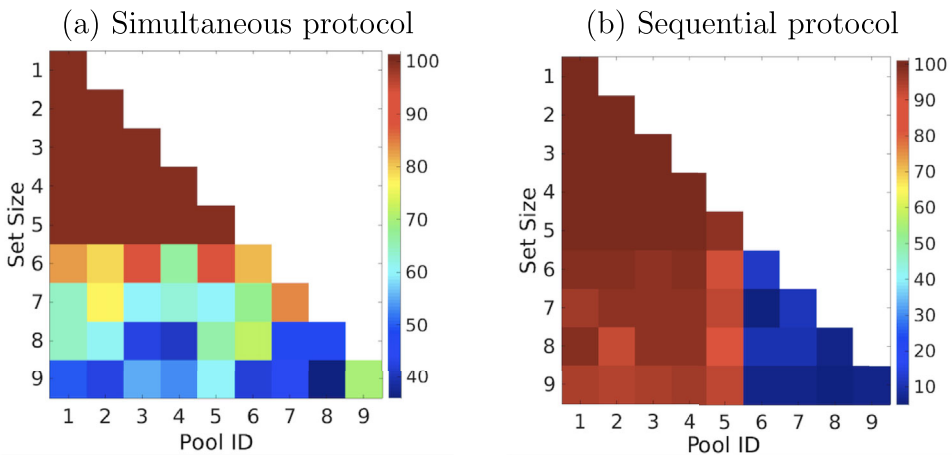


**Fig. 3** Cued memories held in WM under different stimulation protocols for various memory set sizes $S$ ($S = N_{stim}$). Histogram illustrating the percentage of trials in which each selective pool ($P_i$) shows persistent activity during the delay period. Pool ID identifies the selective pool $P_i$. The emergence of persistent activity is considered whenever the mean firing rate fulfils the condition $\bar{\nu} \geq 20$ Hz during a $\Delta t_{WM}$ period of 500 ms, 2.5 s after the last item of the memory set is displayed. **a** Simultaneous stimulation protocol, and **b** Sequential stimulation protocol. The results are obtained from 100 trials with the following network parameters $\omega_{inh} = 0.97$, $\tau_F = 1500$ ms, $\omega_+ = 2.3$ (see the Supplementary Material for the rest of network parameters)

depends on the number of items that have received external stimulation. As can be seen, consistent with previous studies (e.g. [12, 13]), the model predicts that the number of items that can be stored in WM reaches an upper limit. In these simulations, the capacity limit is $K \sim 5$.

Interestingly, although there is no clear preference as to which items are kept in WM during the delay period in the simultaneous stimulation paradigm, Fig. 3(b) shows that, when the sequential stimulation paradigm is considered, those items which are seen first in a sequence are much more likely to be maintained in WM. This is, however, at odds with the experimental observations suggesting a recency effect and indicates that the model, in its current setting, is unable to replicate such findings.

For instance, Kool et al. [36] proposed a WM task in which different coloured squares (set sizes 3 to 5) were observed during 50 ms. The results revealed a solid recency effect in the case of larger set sizes, indicating that $K < 5$ in their experiments. The emergence of the recency effect remained independent of the retention interval time, while a weaker primacy effect was observed only for the first item. Yakovlev et al. [34] consider a DMS visual task in which a variable number of items are presented with an ISI of 1 s, and followed by a variable (1 to 3 s) delay period. Their results demonstrate that performance decreases with increasing sequence lengths. In alignment with the findings of [36], a clear recency effect emerges, while no primacy effects are observed. Their results suggest that up to 6–7 stimuli can be retained in WM.

In the following sections, we embark on an empirical investigation with the primary objective of validating the core hypotheses underpinning our research. We have postulated that the manifestation of serial order effects in WM is closely interconnected with the temporal dynamics of the biophysically relevant processes governing the neuronal activity. One such pivotal process is STSF, with its temporal dynamics regulated by the time constant $\tau_F$. Furthermore, we assert that the design of experimental protocols exerts a substantial influence on the encoding stage of WM, thereby wielding a significant impact on both the effective WM capacity exhibited by the network and the emergence of serial effects. Notably, two key facets of experimental protocols are the nature of stimuli, closely related to the intensity of the stimulation $\lambda_{stim}$, and the stimulation duration, denoted as $\Delta t_{stim}$. In our investigation, we employ a computational model, systematically manipulating these variables to elucidate the fundamental relationships between STFS, the experimental design parameters, and the key features observed in WM function.

## The Critical Role of Synaptic Facilitation in WM

Synaptic facilitation is a ubiquitous feature throughout the brain and it is thought to play a key role in neural processing [37]. Indeed, facilitating excitatory synapses are a major feature in PFC [38]. Interestingly, in an experimental study carried out on ferrets, Wang et al. [38] suggest that the distribution of time constants ($\tau_F$) associated with synaptic facilitation in mPFC is in the range of a few hundred milliseconds, in contrast to the larger values ($\tau_F \sim$ 1500–2000 ms) considered in [9, 14].

The time constant $\tau_F$ plays a critical role in establishing the dynamics of the facilitation mechanism through the temporal evolution of the utilisation parameter $u$. As can be seen from Eq. 3, two antagonistic terms govern the evolution of $u$. On the one hand, a relaxation term, mediated by $\tau_F$, establishes the temporal scale in which $u$ relaxes back to baseline ($U$). On the other hand, for every incoming spike, $u$ increases. In this section, we explore the role that $\tau_F$ plays on the encoding and maintenance of items in WM. Particularly, our next focus of inquiry is the neurodynamical origin of the limited WM capacity exhibited by the network, as well as the mechanisms responsible for establishing which stimuli are preferentially maintained in WM when only some of the cued items can be simultaneously held in memory. To this end, we consider a memory set size of $N_{stim} = 9$, which is significantly larger than most capacity limits commonly reported in the literature ($K \sim$ 4–5) [1].

The results of the simulations for various $\tau_F$ values are presented in Fig. 4. Figure 4(a) depicts the temporal evolution of the firing rate for all the excitatory selective pools in the network, along with the corresponding inhibitory activity. Figure 4(b) illustrates the temporal evolution of the utilisation parameter ($u$). Note that synaptic facilitation exerts an effect akin to the augmentation of synaptic connection weights within each neural population activated by a specific cue during a given trial, in contrast to the non-cued pools. As a consequence, only the synaptic weights of the cued pools experience this augmentation during a particular trial. This results in a significant difference from the synaptic weights associated with the uncued pools, for which the utilisation parameter remains relatively low [14].

For low time constants (e.g. $\tau_F < 500$ ms, results not shown), the system remains quiescent during the delay period, and no memories are retained in WM. This is because the emergence of persistent activity in the network requires a minimal level of effective feedback excitation. The rapid relaxation of the utilisation parameter to baseline prevents $u$ from reaching or remaining at a high asymptotic value ($u_\infty \sim 1$). This condition is crucial for achieving and maintaining the persistent firing regime.

As $\tau_F$ increases, see $\tau_F \sim 550$ ms in Fig. 4, the overall excitability of the system also grows, enabling the network to sustain some memories. This is because the relaxation term diminishes, causing the utilisation parameter to approach high asymptotic values. Consequently, the number of pools coexisting in a high firing rate state increases. However, due to concurrent growth in inhibition, the firing rates of pools

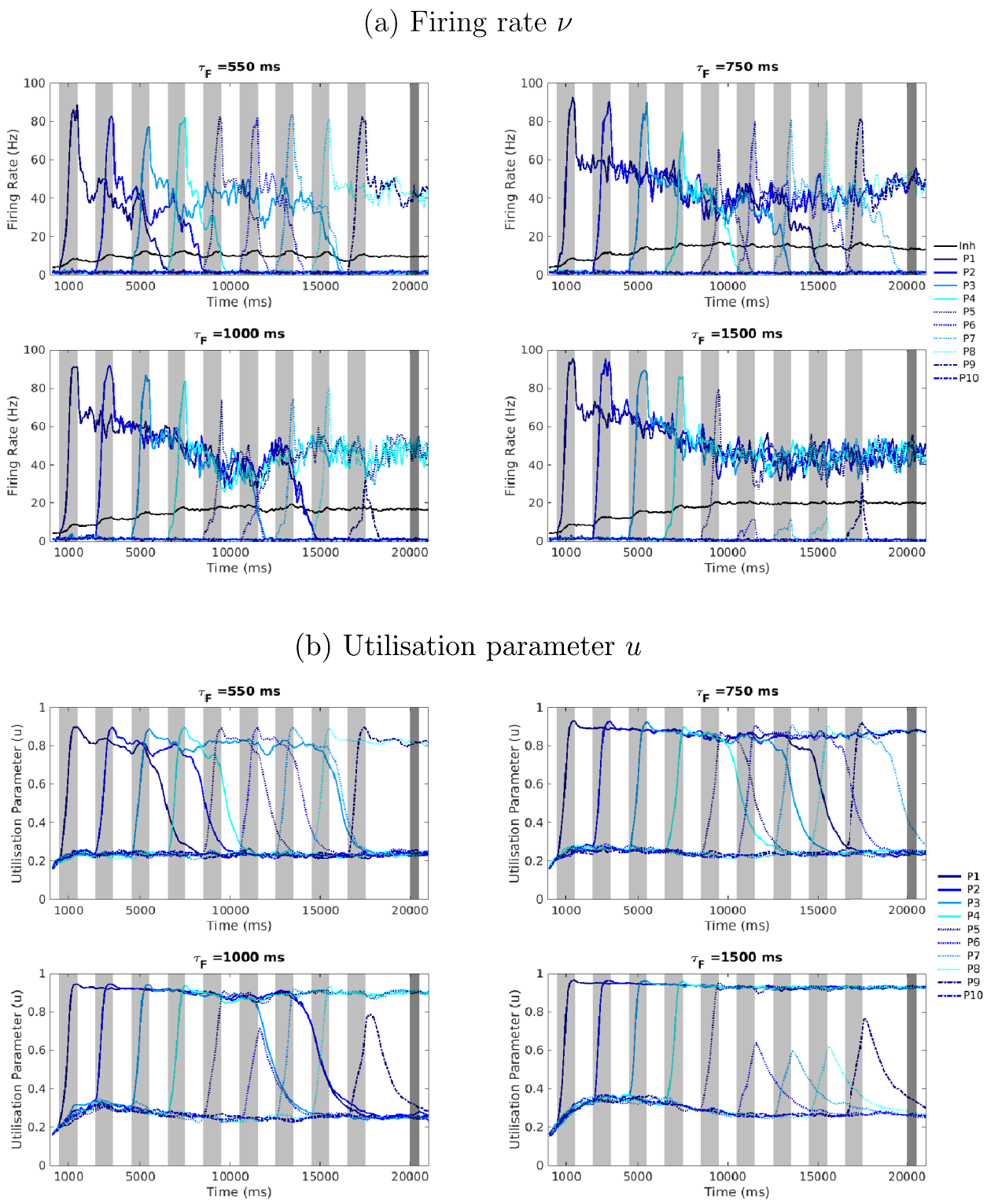


**Fig. 4** Effect of the STSF time constant ($\tau_F$) in the utilisation parameter and firing rate dynamics of sequential stimulation protocols. The stimulation period is depicted in pale grey whereas $\Delta t_{WM}$ (within the delay period) is depicted in dark grey. The network parameters considered in these simulations are the same as in Fig. 3. The stimulation protocol corresponds to the sequential paradigm also described in Fig. 3 ($N_{stim} = 9$). The panels in **a** show the temporal evolution of the firing rates corresponding to the different pools, whereas the panels in **b** illustrate the associated utilisation parameters ($u$), for different $\tau_F$ values: $\tau_F = 550$ ms, $\tau_F = 750$ ms, $\tau_F = 1000$ ms, and $\tau_F = 1500$ ms

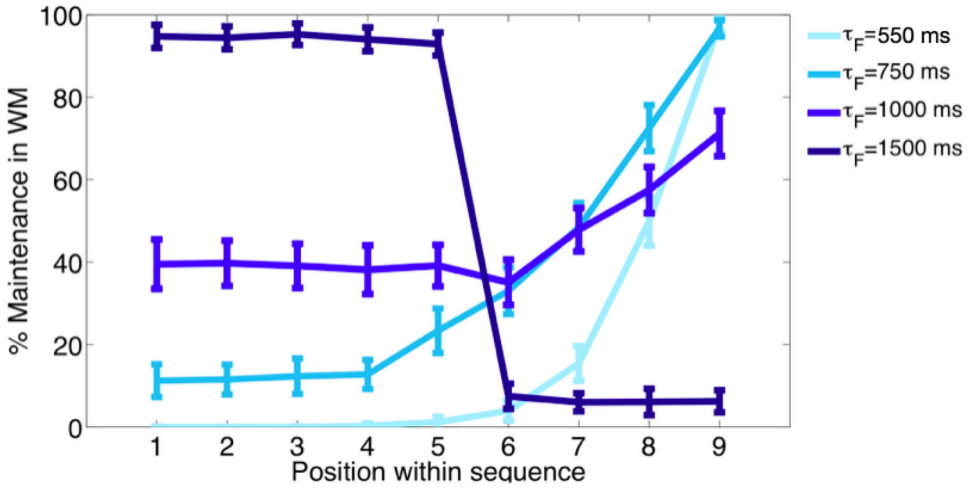


**Fig. 5** Serial order effects in sequential stimulation protocols. The maintenance of an object in WM is estimated by assuming that the item is held in memory if its associated selective pool exhibits a mean persistent activity, $\bar{\nu} > 20$ Hz, for a $\Delta t_{WM}$ period of 500 ms, 2.5 s after the last cue is removed. For each block, this measure provides an estimate of the probability of maintenance of each item within the series, which is subsequently averaged over 100 blocks. The standard deviation of the probability estimate (across blocks) is also illustrated. The results are derived from computational simulations (100 blocks of 100 trials) when different $\tau_F$ values (550 ms, 750 ms, 1000 ms, and 1500 ms) are considered

entering WM gradually decrease. As a result of the competition between activated pools, some of the initial memories that accessed WM become destabilised and are erased from it. The subsequent reduction in global inhibition frees up resources, permitting new memories to access WM, which is key for the recency effect to emerge. Accordingly, the associated $K$ remains low, as only a few items can be maintained in WM.

As the value of $\tau_F$ further increases, see $\tau_F \sim$ 750–1000 ms in Fig. 4, the utilisation parameter $u$ corresponding to each pool maintained in WM during the delay period reaches higher $u_\infty$ values. This leads to more pools that received cues during the stimulation period attaining and maintaining a state characterised by a persistently high firing rate, resulting in a progressive increase in $K$. Gorgoraptis et al. [26] argue that, from the perspective of a shared resource model, the recall advantage for the final item in a sequence results from an uneven distribution of resources, with the largest proportion allocated to the most recently presented item. Notably, the proposed model provides a mechanistic explanation for such allocation of resources, which results from the inherent competition–cooperation mechanisms intrinsic to the network dynamics.

For $\tau_F \sim 1000$ ms, it is interesting to note that the stability of the memories cued earlier in the sequence is greater than that of the pools cued later because inhibition progressively becomes stronger. This effect becomes more prominent when considering items presented in a serial order close to the WM capacity limit of the system (see also Figs. 5 and 6). As a result, items presented in the middle of the sequence are more likely to decay from the WM system if they had accessed it, or to be prevented from accessing it altogether. As these items decay from memory, inhibition decreases thereby creating new opportunities for later items in the sequence to be successfully encoded into WM. While a recency effect remains observable, it is worth noting the concurrent emergence of a primacy-like effect.

For larger $\tau_F$ values ($\tau_F \gtrsim 1200$ ms), only the pools cued early in the sequence achieve the state of persistent firing. From the fifth stimulated pool onwards, none of the subsequently stimulated pools reaches (not even transiently) a high firing rate state during the encoding stage. These memories are highly stable, resistant to interference, and able to efficiently recruit inhibition. This is why memories can be successfully encoded up to the capacity limit $K$, beyond which no additional items are encoded in the WM system. Consequently, the recency effect is diminished and the serial order effects are mainly determined by a primacy-like behaviour.

To evaluate the generality of the previous findings, Fig. 5 presents an estimation of the likelihood of retaining an item in WM as a function of its serial position in the sequence. The results are estimated from 10000 trials (100 blocks with 100 trials each). The simulations indicate that within the range $\tau_F \sim$ 550–1000 ms, the probability of keeping the last cued item in WM is higher compared to earlier-cued items. However, as $\tau_F$ surpasses 1000 ms, the prevalence of the last stimulated item in WM diminishes, although the recency effect remains. As previously discussed, a qualitative shift occurs in the system when $\tau_F \gtrsim 1200$ ms, driven by the increased inhi-

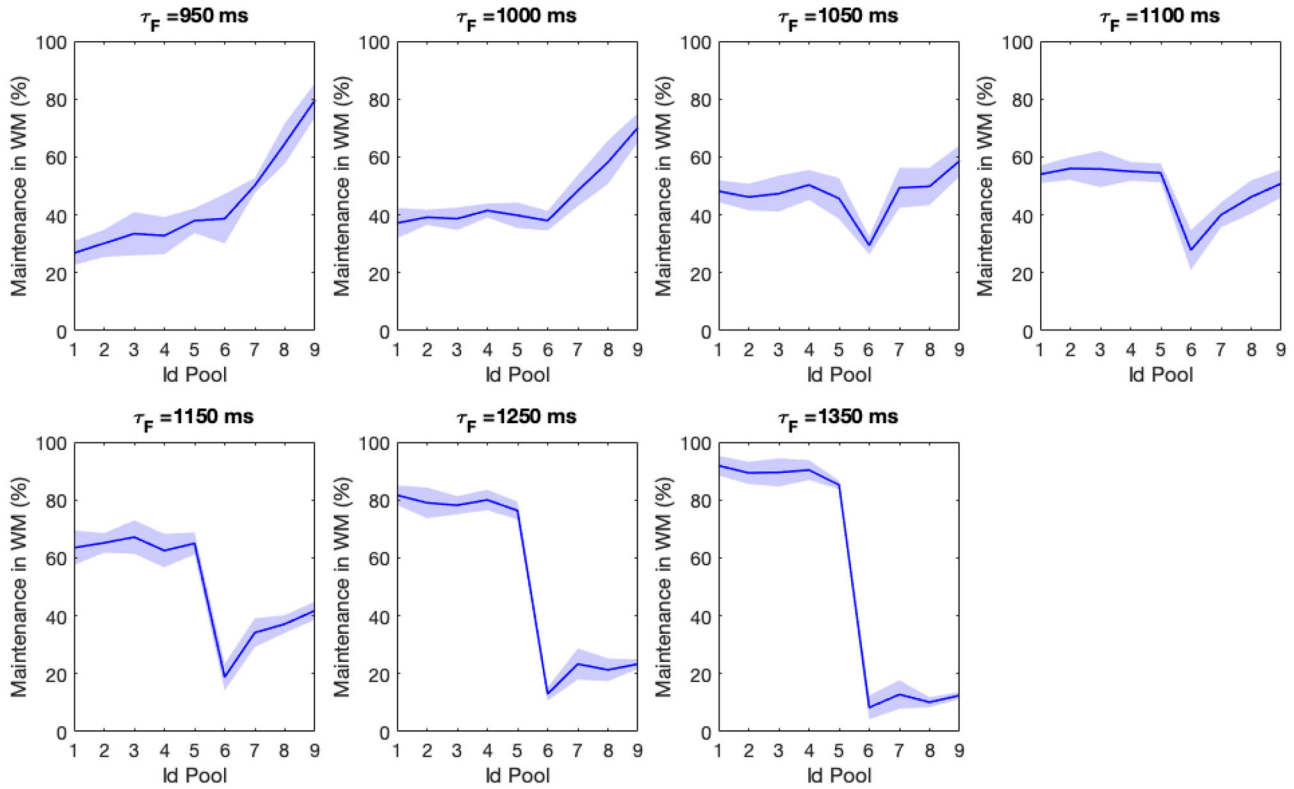


**Fig. 6** Serial order effects in sequential stimulation protocols. As in Fig. 5 for $\tau_F$ values: 950 ms, 1000 ms, 1000 ms, 1050 ms, 1100 ms, 1150 ms, 1250 ms, and 1350 ms; which correspond to the transition between two distinctive regimes. The shaded area represents the standard deviation obtained across blocks

bition, ultimately leading to the disappearance of the recency effect. In this regime, cues presented later in the sequence fail to attain a high self-sustained firing rate, resulting in the absence of stable memories for these items. In the light of these results, we suggest that the inherent heterogeneity of $\tau_F$ values in the brain may partially account for the variability in findings reported in the literature regarding serial order effects, specifically the presence of recency effects alone versus the coexistence of primacy and recency effects.

After identifying different dynamical regimes in which serial order effects emerge with the sequential display of stimuli, we now reconsider the simultaneous stimulation condition. Figure 7 clearly illustrates the scaling of WM capacity with $\tau_F$, aligning with the observations made with sequential stimulation protocols. This boost in capacity is achieved by effectively increasing the synaptic strengths only for those selective pools that are cued, and by maintaining the synaptic facilitation through continued neuronal firing in those pools when the cue is removed. In contrast to sequential stimulation, when all items are cued simultaneously, only the stochastic nature of the received stimulation, along with that of the spontaneous background activity, will determine which items are successfully encoded and retained in WM. As a result, on average, all stimuli are equally likely to be memorised from trial to trial.

Overall, the cooperation-competition mechanisms inherent to the attractor networks dynamics, along with the synaptic dynamics that transiently modulate network connectivity, establish working regimes compatible with experimental results reported in the literature for both simultaneous and sequential stimulation protocols.

## WM Capacity and Serial Order Effects: Sequential *vs* Simultaneous Encoding

Now, we analyse and compare the predictions of our model for both stimulation protocols while varying the memory set size. The dynamics of the stimulation protocols remain consistent with the description in the previous section, featuring $\Delta t_{stim} = 1$ s in both protocols. Considering the consistent reporting of the recency effect in the literature for sequential protocols, as opposed to the occasional emergence of the primacy effect, we choose the value $\tau_F = 750$ ms that places the network in a working regime where recency is prominent.

As seen in Fig. 8, when the number of cued pools is below the WM capacity limit ($K \sim 4$), the network successfully maintains all the stimulated items in WM. In agreement with Gorgoraptis et al. [26], in these circumstances, no differences are found between both display protocols. In contrast, when the memory set size surpasses the network capacity (i.e.

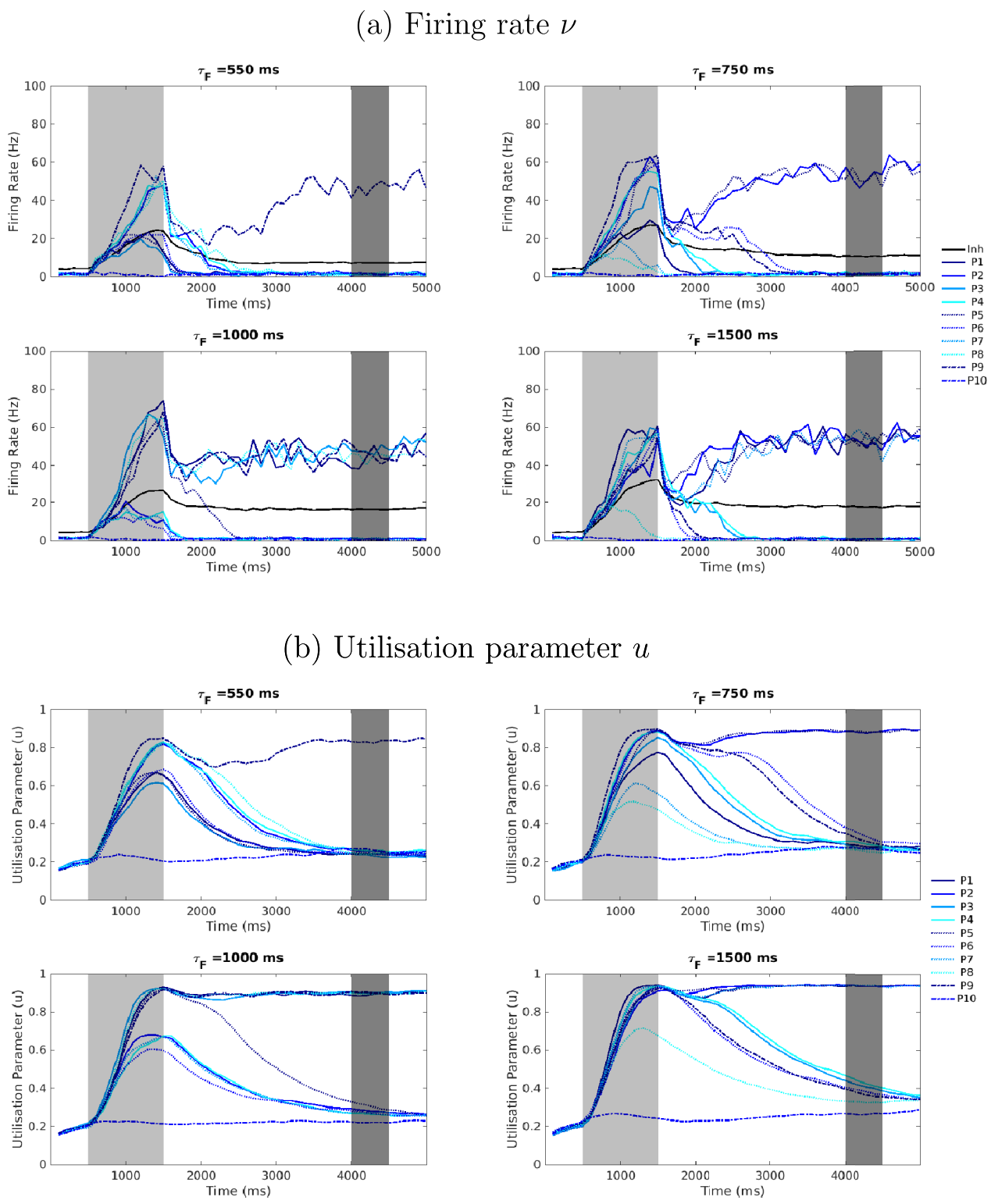


**Fig. 7** Effect of the STSF time constant ($\tau_F$) in the utilisation parameter and firing rate dynamics of simultaneous stimulation protocols. The stimulation period is depicted in pale grey whereas $\Delta t_{WM}$ (within the delay period) is depicted in dark grey. The network parameters considered in these simulations are the same as in Fig. 3. The stimulation protocol corresponds to the simultaneous paradigm also described in Fig. 3 ($N_{stim} = 9$). The panels in **a** show the temporal evolution of the firing rates corresponding to the different pools, whereas the panels in **b** illustrate the associated utilisation parameters ($u$), for different $\tau_F$ values: $\tau_F = 550$ ms, $\tau_F = 750$ ms, $\tau_F = 1000$ ms, and $\tau_F = 1500$ ms

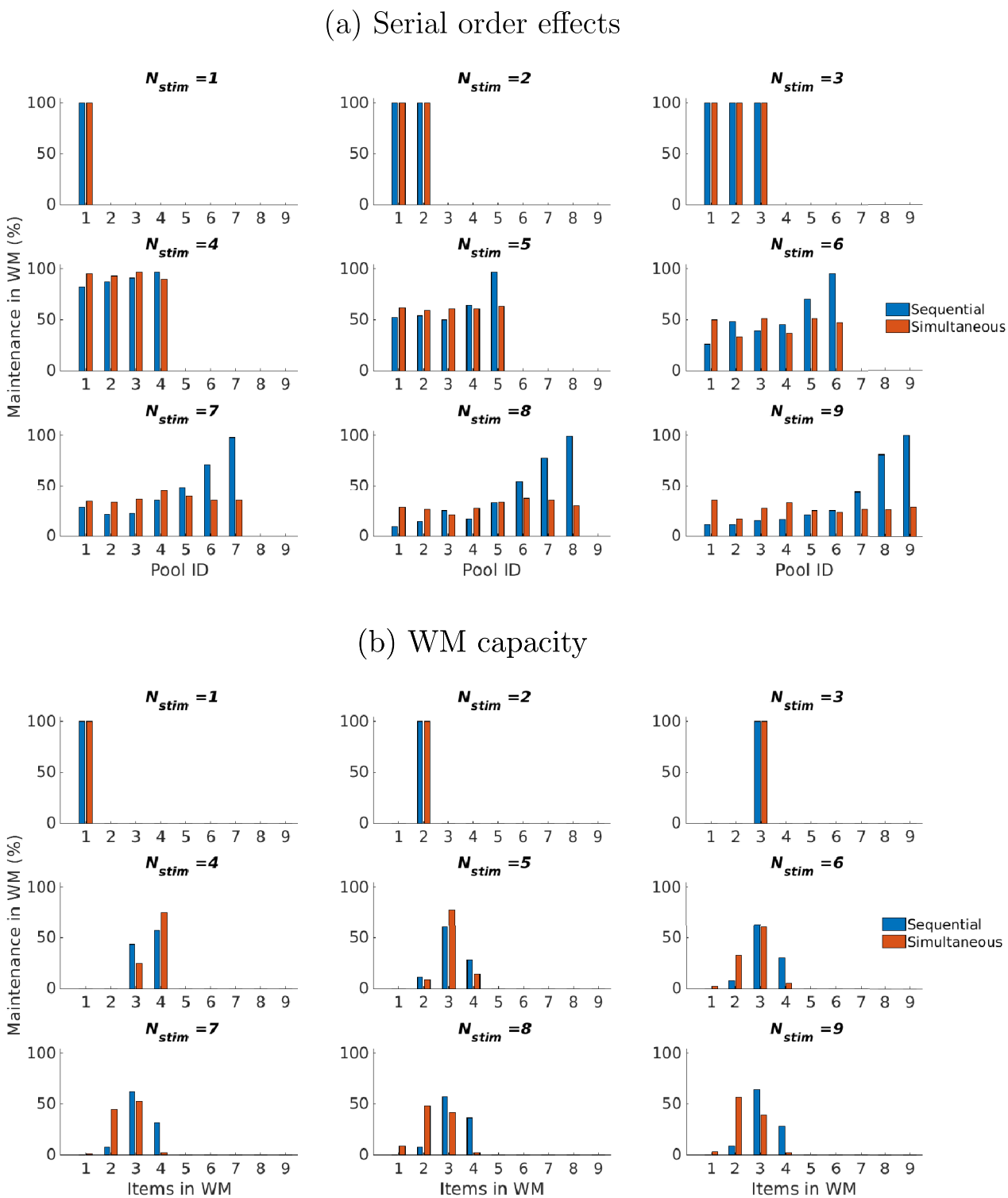


**Fig. 8** Serial order effects and WM capacity in sequential and simultaneous stimulation protocols. Results obtained from 100 simulated trials with the following network parameters: $\omega_+ = 2.3$, $\omega_{inh} = 0.97$, $N = 1000$ neurons, and $\tau_F = 750$ ms. The specific stimulation protocol considered in these simulations makes use of: $\lambda_{stim} = 3.31$ Hz/synapse, $\lambda_{ext} = 3.05$ Hz/synapse, $\Delta t_{stim} = 1000$ ms (for all the simulations), and ISI = 1000 ms (for the sequential protocol only). **a** Serial order effects, and **b** WM capacity

$N_{stim} = S > K$), only a subset of the stimulated items can be retained in WM. In particular, in the sequential condition, a prominent recency effect is evident for $S > 4$.

For larger memory set sizes, increased competition among the stimulated pools leads to a reduction in memories held in WM, falling below 4 for both experimental conditions. However, this phenomenon appears to be particularly emphasised in the simultaneous condition, giving rise to predictions that are strongly aligned with those presented in Edin et al. [12]. Indeed, the results presented in Fig. 8(b) suggest that the overall network capacity is higher for sequential stimulation protocols when $S > 7$ ($K_{8,seq} = 3.29$, $K_{9,seq} = 3.2$, $K_{8,sim} = 2.36$, $K_{9,sim} = 2.4$).

The average WM capacity ($K$) exhibited by the network can be estimated as follows:

$$K = \sum_{i=1}^{S} p_i \cdot i \tag{4}$$

where $p_i$ is the probability (estimated from the simulations) that $i$ memories are simultaneously held in WM, and $S$ is the memory set size.

The literature presents inconclusive findings regarding which stimulation protocol leads to higher WM capacities. In [23], a capacity $K \sim 3$ was reported for both protocols. In contrast, Lecerf and Ribaupierre [24] reported an advantage in recall performance for simultaneous presentations compared to sequential ones in a visuospatial working memory task. Similarly, the results in [25] showed improved performance with simultaneous presentation of stimuli and higher WM capacities compared with sequential protocols. Notably, these experiments used various types of stimulus and different $\Delta t_{stim}$ intervals. For instance, in [23], $\Delta t_{stim}$ was set 200 ms for the simultaneous condition and 50 ms for the sequential one, while in [24], $\Delta t_{stim}$ was 1 s for both protocols, and in [25], $\Delta t_{stim}$ was 1 s for the sequential protocol and 4 s for the simultaneous one.

These observations reinforce our hypothesis that the design of experimental protocols exerts a substantial influence on the encoding stage of WM, thereby modulating the effective WM capacity exhibited by the network. The stimulation time $\Delta t_{stim}$ plays a key role in establishing the overall excitatory activity of the network. To investigate this hypothesis, we have conducted a simulation study with varying $\Delta t_{stim}$ values for both the sequential and the simultaneous conditions.

Figure 9 compares the results obtained when items displayed simultaneously are shown for 1 s, while sequentially stimulated items are displayed for 250 ms. This adjustment aims to ensure comparable levels of attention allocation to all items in the memory set for both stimulation protocols. As can be seen, in terms of the emergence of serial order effects, similar outcomes are observed when the stimulation time for the sequentially presented items is reduced. Noteworthingly, for the network to reach a sufficiently high level of excitation, given the much shorter stimulation time ($\Delta t_{stim}$) in the sequential stimulation protocol, the $\lambda_{stim}$ applied to each pool must be increased. Specifically, we have increased it from 3.31 Hz/synapse to 3.63 Hz/synapse for both protocols. Naturally, when the stimulation period is extended, the network reaches comparable levels of excitation and engages more recurrent activity with lower $\lambda_{stim}$ levels. This is in agreement with the notion of effective WM ($e$WM) introduced in [13], which is also supported by [18].

The discrepancy between the outcomes depicted in Figs. 8(b) and 9(b) can be attributed to the heightened competition among stimulated pools for larger values of $\lambda_{stim}$, leading to a reduced $K$ in the simultaneous condition for $S > 7$ ($K_{8,seq} = 3.14$, $K_{9,seq} = 3.17$, $K_{8,sim} = 1.71$, $K_{9,sim} = 1.34$). Importantly, $\lambda_{stim}$ can be related, at the behavioural level, to the stimulus type as well as to its intensity or conspicuity, all of them critical aspects in the design of experimental protocols.

Finally, since the increase of the synaptic strength due to the facilitation mechanism is mostly triggered during the stimulation phase, modifying the synaptic dynamics during this stage might clearly have an impact on the overall behaviour of the network. This can be achieved by varying specific aspects of the protocols such as the type of stimuli employed, the time during which each item is displayed, or the specific recall task (e.g. recall a complete sequence *vs* presence of particular item within a sequence). As emphasised throughout this work, the conclusions drawn from investigations into WM capacity limits may vary depending on the specific experimental design of the stimulation protocol. This underscores the need to establish benchmarks for WM assessment [28]. It has also inspired the following section, in which we revisit the notion of $e$WM [13] and assess how variations in the experimental protocols, such as stimulation strength $\lambda_{stim}$ and stimulation time $\Delta t_{stim}$, affect $K$ and the emergence of serial effects in sequential protocols.

## Effective WM Capacity ($K_e$)

In this section, we explore the coordinated influence of the synaptic facilitation time constant $\tau_F$ alongside various factors intricately linked to the specific experimental protocol. These factors encompass the level of external stimulation relative to the spontaneous background activity ($\Delta\lambda = \lambda_{stim} - \lambda_{ext}$), and the duration of the stimulation $\Delta t_{stim}$. It is essential to highlight that both of these factors are critically involved in the encoding stage of WM. In fact, as previously discussed, $\Delta\lambda$ is likely to reflect the type of stimulus in terms of complexity and conspicuity, whereas $\Delta t_{stim}$ denotes the duration for which the stimuli are presented.

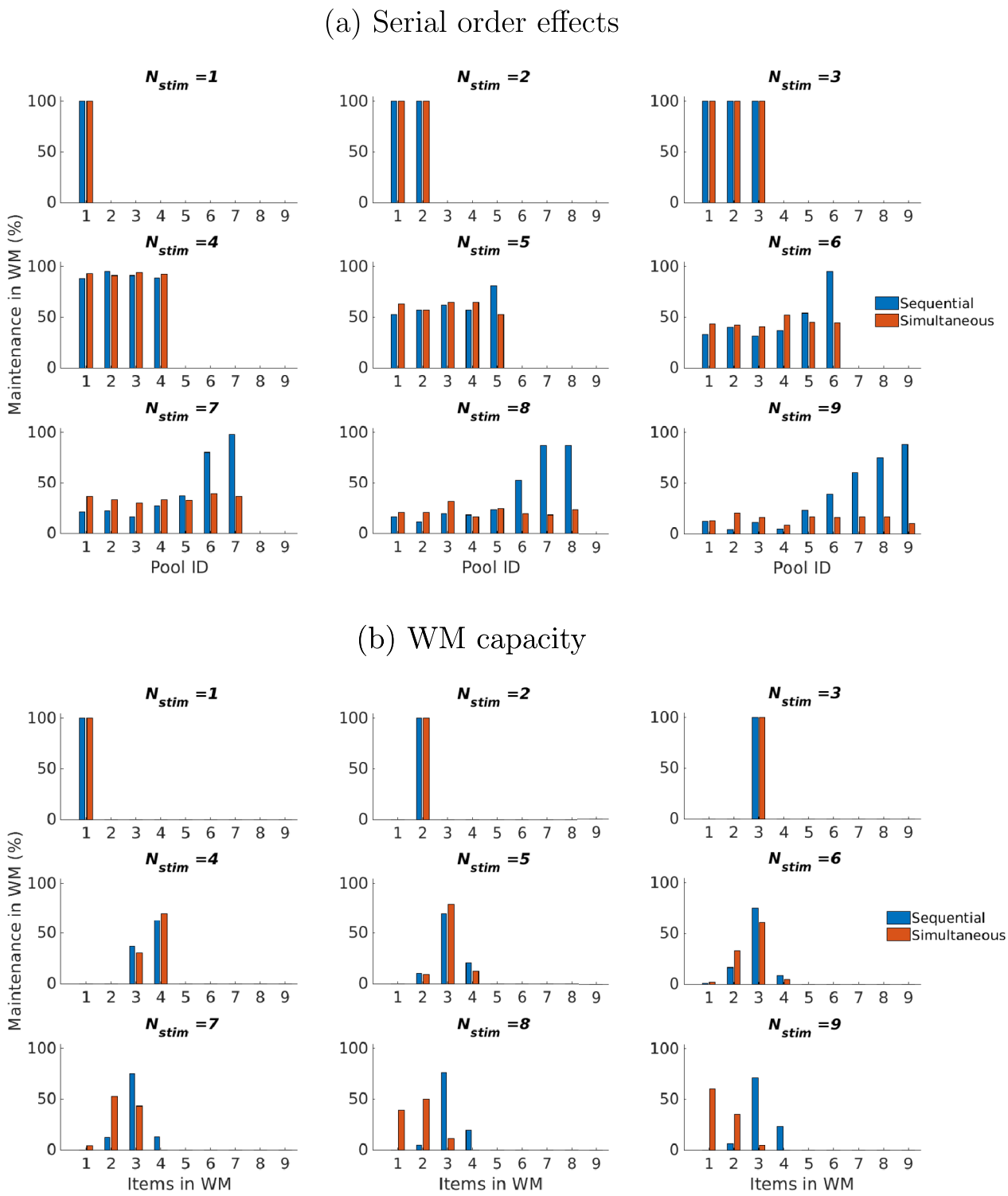


**Fig. 9** Serial order effects and WM capacity in sequential and simultaneous stimulation protocols. Results obtained from 100 simulated trials with the following network parameters: $\omega_{+} = 2.3$, $\omega_{inh} = 0.97$, $N = 1000$ neurons, and $\tau_F = 750$ ms. The specific stimulation protocol considered in these simulations makes use of: $\lambda_{stim} = 3.63$ Hz/synapse and $\lambda_{ext} = 3.05$ Hz/synapse. $\Delta t_{stim}$ varies depending on the stimulation protocol. For simultaneous stimulation $\Delta t_{stim} = 1000$ ms, whereas for sequential stimulation $\Delta t_{stim} = \text{ISI} = 250$ ms. **a** Serial order effects, and **b** WM capacity

Considering that serial effects manifest exclusively in sequential protocols when $N_{items} > K$, our study remains focused on a stimulation protocol involving 9 items. Figure 10 illustrates several relevant features. First and foremost, as previously discussed, $\tau_F$ exerts a pronounced influence on network behaviour, resulting in a transition of emerging serial effects from recency to primacy-like patterns. This qualitative shift in behaviour stems from the dynamic interplay between excitation and inhibition in the network. Thus, similar effects may be elicited by altering the balance between excitation and inhibition through alternative means. This is unsurprising in the light of the differential equation governing the dynamics of the utilisation parameter $u$ (see Eq. 3) since $\Delta\lambda$ crucially modulates the generation of spikes trains during the stimulation period.

The intensity of the stimulation received by the selective pools ($\lambda_{stim}$) relative to the spontaneous background activity of the network ($\lambda_{ext}$, often linked with the arousal level) also appears to have a significant impact on the behaviour of the network. Indeed, increasing $\Delta\lambda$ leads to a transformation of sequential effects from primacy-like to recency patterns. This transformation is evident in the results corresponding to $\Delta t_{stim}$ = 500 ms and $\tau_F \geq 750$ ms in Fig. 10(b), where a sudden transition appears. Interestingly, this transition occurs at various $\Delta\lambda$ levels for different $\tau_F$ values, highlighting the deep interdependence between these parameters.

Low $\Delta\lambda$ values trigger a gradual engagement of inhibition as the network encodes new stimuli. Therefore, those memories entering WM first are more likely to remain stable during the delay period, benefitting from the facilitated recurrent connections. As the network approaches its capacity limit, the encoding of additional memories becomes increasingly challenging.

In contrast, when $\Delta\lambda$ values are large, inhibition is rapidly recruited, which can destabilise previously encoded memories. As a result of the subsequent decrease in global inhibition, the WM system is more likely to successfully encode and retain the last items in the sequence. These effects strongly depend on the time scale established by $\tau_F$ for synaptic facilitation. Specifically, larger $\tau_F$ values promote the stability of initially encoded items. As a consequence, increasing $\Delta\lambda$ values are required for the emergence of serial effects. The results displayed in Fig. 10, correspond to a fixed $\lambda_{ext} = 3.05$ Hz/synapse value. Similar results are obtained when $\lambda_{ext} = 3.1$ Hz/synapse for equivalent $\Delta\lambda$ values (see Supplementary Material).

It is worth noting that $\Delta t_{stim}$ and $\Delta\lambda$ exhibit a conjugate-like behaviour in terms of the overall excitatory input received by the network following stimulation (in agreement with [18]). Similar results are obtained for smaller values of $\Delta\lambda$ when $\Delta t_{stim}$ is increased, and conversely. Note, for instance, how incrementing the stimulation time from $\Delta t_{stim}$ = 250 ms to $\Delta t_{stim}$ = 1000 ms for $\tau_F$ = 750 ms results in a change of regime from primacy-like ($\Delta t_{stim}$ = 500 ms) to recency ($\Delta t_{stim}$ = 1000 ms) in an initially quiescent network ($\Delta t_{stim}$ = 250 ms) for $\Delta\lambda = 0.26$ Hz/synapse. Once more, this reinforces the idea that encoding is contingent upon the particular WM task, especially the stimulation protocol, as predicted by the concept of $e$WM capacity [13].

Notably, the parameters $\tau_F$, $\Delta\lambda$ and $\Delta t_{stim}$ not only affect the emergence and nature of the serial effects in sequential stimulation protocols but also play a crucial role in establishing the overall effective network capacity ($K_e$), as depicted in Fig. 11. In particular, increasing $\Delta\lambda$ initially leads to larger $K_e$ values, which is a result of the overall increase in network excitability, while the recruited inhibition does not yet dominate the behaviour of the network. Nonetheless, for large $\Delta\lambda$ values, the competition between the different pools (as mediated by the ramping inhibition) is so strong that the overall network capacity may be again reduced. See, for instance, the results obtained for $\Delta t_{stim}$ = 1000 ms when $\Delta\lambda$ changes from 0.26 to 0.64 Hz/synapse, and notice how $K_e$ decreases when $\tau_F \geq 1000$ ms.

Naturally, the functioning of WM in general, including its capacity, is also highly dependent on the connectivity parameters of the network. So far, we have examined a predetermined set of values that enable the network to function in accordance with the observed experimental outcomes. Notably, among these parameters, $\omega_{inh}$ plays a prominent role, as discussed in [14]. To evaluate the robustness of our proposed model to variations in $\omega_{inh}$, and its interplay with the time scale of the STFS process, Fig. 12 illustrates the influence of modifying $\omega_{inh}$ on WM capacity and the occurrence of serial order effects for networks with and without STFS. Two different $\tau_F$ values ($\tau_F = 750$ ms and $\tau_F = 1500$ ms) have been considered for the network endowed with STFS.

In Fig. 12(a), the shaded areas between the continuous lines delineate the region of the parameter space ($\omega_{inh}$, $N_{stim}$) where all the stimulated items are concurrently maintained in WM, for each of the networks. Above these lines, only a subset of the cued pools persists in memory, while below them ($\omega_{inh} < 1.025$ for the network without STFS, and $\omega_{inh} < 0.94$ for the networks endowed with STFS), non-cued memories may enter WM. Undoubtedly, when there is a need to simultaneously retain several memories, it becomes imperative to restrict the level of inhibition. This prerequisite imposes an upper threshold on $\omega_{inh}$, which decreases as the demand for concurrent memory activation escalates. The precise value of $\omega_{inh}$ at which the upper and lower limit coincide determines the largest WM capacity achievable by the network. In this study, $K_{max} = 4$ if the system does not incorporate STFS, whereas for the networks endowed with STFS, $K_{max} = 6$ when $\tau_F = 750$ ms, and $K_{max} = 10$ when $\tau_F = 1500$ ms.

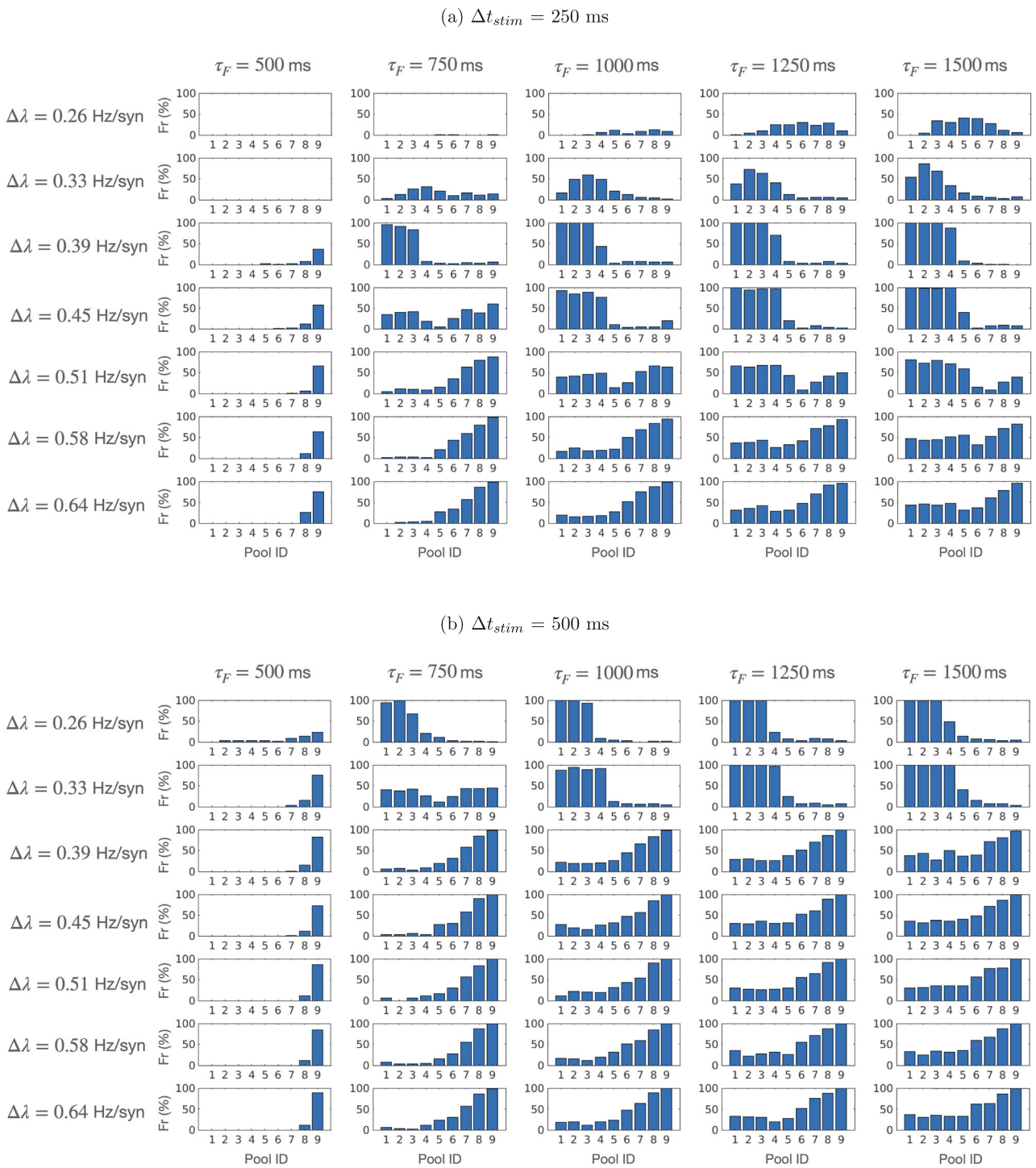


**Fig. 10** Serial order effects: Analysis of the serial order effects as a function of the concerted action of $\tau_F$, $\Delta\lambda$ ($\Delta\lambda = \lambda_{stim} - \lambda_{ext}$), and $\Delta t_{stim}$. In all the simulations ($N = 100$), $\lambda_{ext} = 3.05$ Hz/synapse. Each bar indicates the frequency with which an item in a particular position in the sequential memory set (as characterised by its pool ID) is held in WM. **a** $\Delta t_{stim}$ = 250 ms, **b** $\Delta t_{stim}$ = 500 ms, and **c** $\Delta t_{stim}$ = 1000 ms

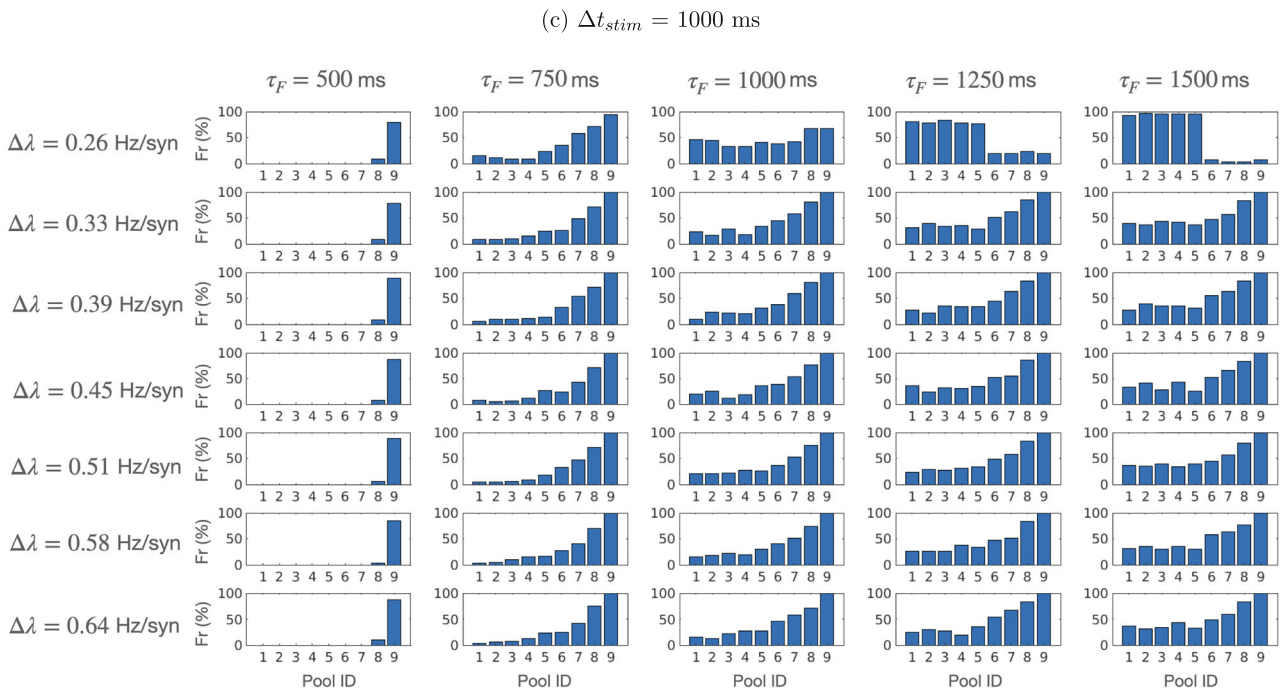


**Fig. 10** continued

Figure 12(b) shows the histograms depicting the percentage of trials in which each item, stimulated in the sequence in the order denoted by its Pool ID, is retained in WM. These histograms are generated from simulations of the network using the $\omega_{inh}$ values indicated in the figure. It is worth noting that these values are also highlighted in Fig. 12(a) to depict their locations within the $(\omega_{inh}, N_{stim})$ parameter space and their association with the number of memories concurrently held in WM. Although, as can be seen, STSF is not a strict requirement for the recency effect to emerge, we uphold the view that it exerts a distinct and crucial influence on the behaviour of the network, and emphasise the particular significance of $\tau_F$ in this endeavour. The inclusion of STFS renders the system less susceptible to minor fluctuations in $\omega_{inh}$. For instance, the range $\Delta\omega_{inh}$ within which the network without STFS can sustain three items simultaneously in WM is significantly narrower when compared to those of the networks endowed with STFS. This robustness to small changes is a significant advantage, as it better reflects the way the brain operates. Moreover, STSF amplifies the overall WM capacity of the network, bringing it into greater alignment with the capacities documented in the literature. The histograms in the figure, resembling those in Fig. 10, illustrate the nature of the serial order effects observed for the specific $\omega_{inh}$ under consideration, once again underscoring the relevance of $\tau_F$ in promoting the transition from recency to primacy-like effects.

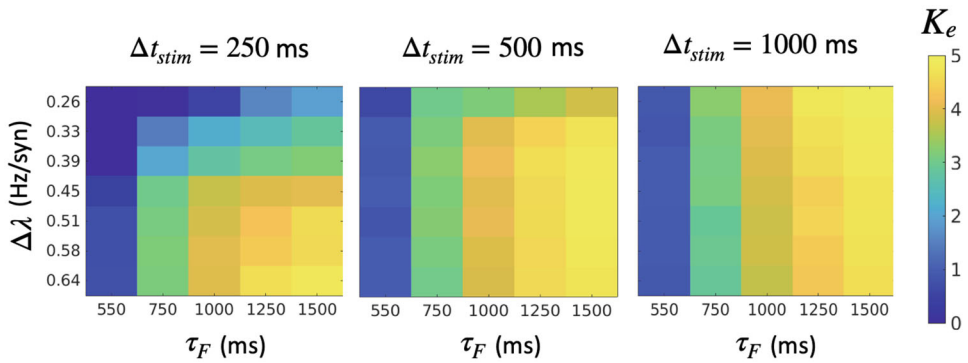


**Fig. 11** WM capacity ($K_e$): Analysis of $K_e$ as a function of the concerted action of $\tau_F$, $\Delta\lambda$ ($\Delta\lambda = \lambda_{stim} - \lambda_{ext}$), and $\Delta t_{stim}$. In all the simulations ($N = 100$), $\omega_+ = 2.3$, $\lambda_{stim} = 3.31$ Hz/synapse, $\lambda_{ext} = 3.05$ Hz/synapse

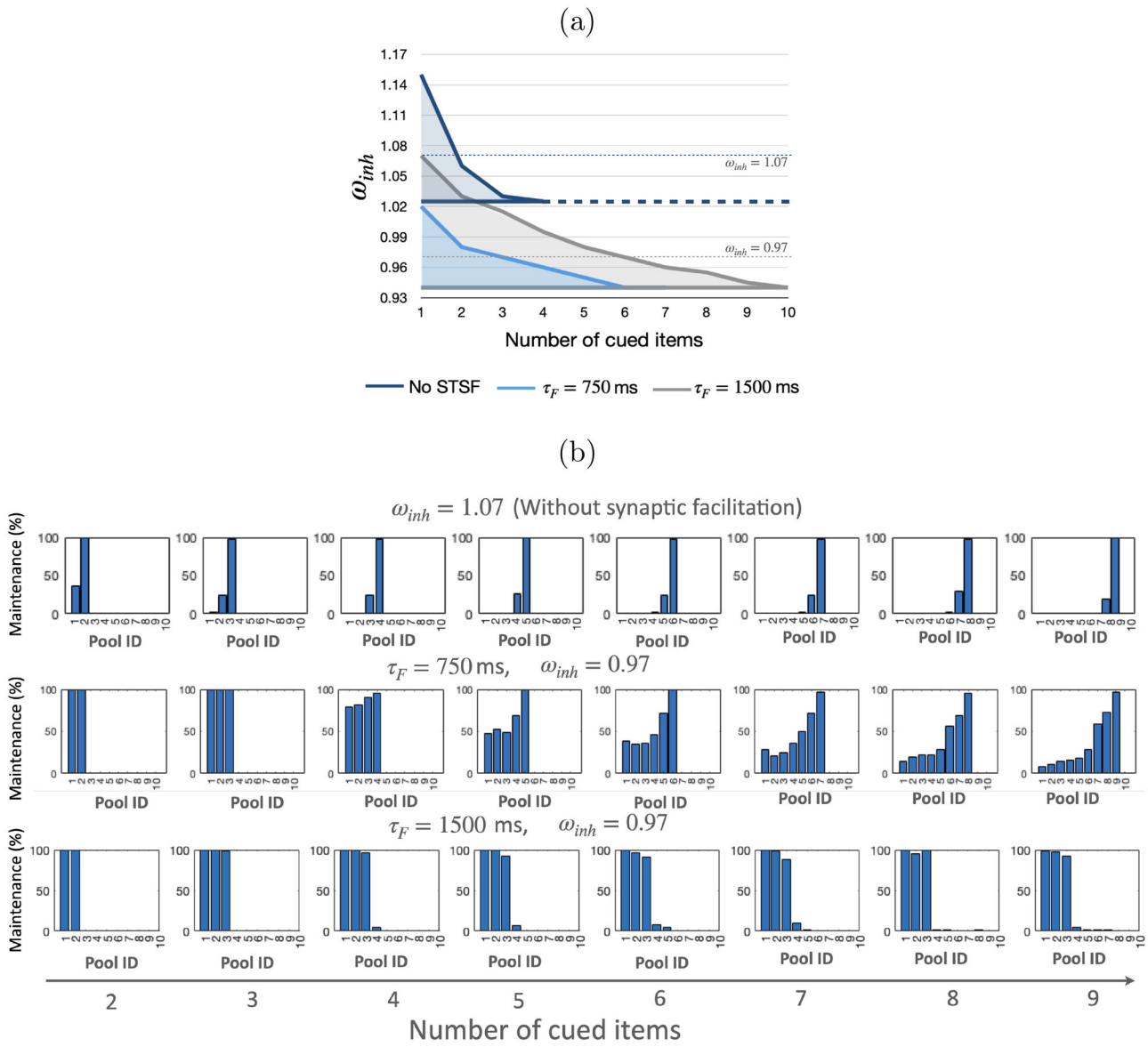


**Fig. 12** Interplay between the strength of the inhibitory-to-excitatory synapses ($\omega_{inh}$) and the time constant of the STSF process ($\tau_F$). In panel **a**, the shaded areas between the continuous lines delineate the region of the parameter space ($\omega_{inh}$, $N_{items}$) where all stimulated items are concurrently maintained in WM. In the computational simulations, both the number of sequentially cued pools ($N_{items}$) and $\omega_{inh}$ are varied to assess their effect on WM capacity, and the emergence of serial order effects. Different colours correspond to networks with different properties in terms of the synaptic facilitation mechanism. Above the continuous lines, only a subset of the cued pools persists in memory, while below them, non-cued memories may enter WM. In panel **b**, histograms depict the percentage of trials in which each item, stimulated in the sequence in the order denoted by its Pool ID, is retained in WM for varying $\omega_{inh}$ values: $\omega_{inh}$ = 1.07 for the network without STFS, and $\omega_{inh}$ = 0.97 for the networks with STFS. The parameters considered in these simulations are: $\omega_+ = 2.3$, $\lambda_{stim} = 3.31$ Hz/synapse, $\lambda_{ext} = 3.05$ Hz/synapse

## Conclusions and Discussion

Of particular interest to our study is the work by Edin et al. [12], in which an attractor network is used to model visuospatial WM and the mechanisms underlying WM capacity are analysed in depth. The authors show that there exists an upper boundary to the capacity limit arising from lateral inhibition. Our model aims to preserve these dynamics while also accounting for the serial effects reported in the literature on sequential protocols. In contrast to the model proposed by Mi et al. [39], which is based on the synaptic theory of WM and keeps a single memory representation active at any given moment, our model displays persistent firing activity in all objects sustained in WM (as in [40]).

In agreement with previous studies [9, 14], we have found that synaptic facilitation boosts the WM capacity limit by effectively increasing the synaptic strengths only for those pools to which a cue is applied. After the cue is removed, the continued firing of neurons in these same pools maintains synaptic facilitation. We suggest that STSF is a neurophysiological mechanism with a key role in establishing the WM capacity limits while also linking the system response to the intrinsic dynamics of the experimental protocols. In particular, the time constant of the synaptic facilitation process ($\tau_F$) has been identified as a crucial factor in modulating WM capacity and the emergence of serial order effects in sequential protocols.

Notably, the design of the experimental protocol, which includes aspects such as the specific recall task, the nature of the stimuli (related to $\Delta\lambda$), and the duration of the stimuli display ($\Delta t_{stim}$) together with the interstimulus interval (ISI), plays also a critical role on establishing the effective limits to WM capacity. While the network connectivity structure determines the excitation–inhibition balance that sets the maximum WM capacity the network may exhibit, the specific display protocol establish whether such limit may be reached. Accordingly, we have investigated the impact of stimulation protocols on the key features of WM function, particularly in determining the effective capacity limits of the network for a specific experimental task, and the emergence of serial order effects. Unlike most computational studies that focus mainly on the behaviour of the network during the steady-state, we emphasise the importance of considering the encoding stage during the display of the memory set. We begin our analysis by evaluating both sequential and simultaneous displays of multi-item memory sets, establishing a foundational level of analysis. We then delve deeper by examining the impact of the stimulation time on WM function, in conjunction with the level of external excitation received by the network. We hereby recall the notion of effective WM ($e$WM) [13], which is a construct that considers the critical constraints imposed on the WM system during the encoding stage. This approach offers a comprehensive framework to investigate WM function and a plausible explanation of the neuronal mechanisms underlying WM capacity limits.

Consistent with the findings in [13], the resulting excitation achieved by the network in response to the stimulation displays a conjugate-like behaviour concerning the variables intensity level ($\lambda_{stim}$) and time ($\Delta t_{stim}$). This behaviour persists even when the system is endowed with STSF. These results are in accordance with the research presented in [18], which outlines a multiscale neocortical model spanning from molecular to network levels. The primary emphasis of this study is to assess HCN modulation mechanisms through a second-messenger signalling cascade across these levels. This investigation demonstrates that prolonging the stimulation duration, even at low intensities, leads to an augmentation in the firing rate of the stimulated neuron population, and this firing rate continues to rise with extended stimulation durations. Indeed, our findings suggest that these two variables have similar effects on encoding. Specifically, we observed that increasing $\lambda_{stim}$, and thus $\Delta\lambda$, leads to a reduction in the required $\Delta t_{stim}$ to achieve a given effective WM capacity ($K_e$). Conversely, longer $\Delta t_{stim}$ values require smaller $\Delta\lambda$ cue inputs to overcome the quiescent state and reach the same $K_e$. Therefore, it is the combination of both parameters that will establish the specific $K_e$ for a particular WM task.

Our results are arguably compatible with a resource model of WM and with the dynamic coding perspective outlined in [41], which states that WM capacity limits are closely tied to limits in encoding and/or readout. Indeed, as the number of items in memory increases, the proportion of resources dedicated to each item declines. In our model such distribution of resources may be interpreted in terms of the steady-state firing rate, which is lower for larger memory set sizes, thus degrading the stability of the memories and (likely) the fidelity of storage [26]. In the proposed computational model, the inhibition recruited as a result of the increasing excitatory activity of the cued pools mediates the competition between the different pools and determines the effective capacity limit. In agreement with Gorgoraptis et al. [26], when the memory set size is below the capacity limit no differences are found between both display protocols. However, shall the dynamics of the inherent network competition be altered, for instance by manipulating the relative saliency of the items in the memory set, our prediction is that $K_e$ will reflect such effects. Indeed, as discussed in [13], when one item is prioritised (e.g. either through top down modulation or conspicuity), its recall is enhanced, an aspect which is accompanied with corresponding decrements in recall rate for other objects. The proposed model thus paves the way to further investigate the interplay between attention and WM, which becomes critical shall more realistic scenarios be considered.

Although ours is a minimal model which does not attempt to explain any aspects related to WM accuracy—something which can only be tackled through the consideration of bump attractor models—we argue that it offers valuable insights into the critical role that STSF plays in WM function. Notably, it is able to reconcile numerous experimental results stemming from broadly varying task designs, and offers a principled explanation for the observed behaviour. For instance, it provides an explanation to the fact that some task designs lead to values of $K_e \sim$ 6–7 items while others show strict limits around $K_e \sim$ 2–3 items. Synaptic facilitation has been shown to improve the stability of WM in continuous attractor networks by decreasing both diffusion and directed drifts [42], which underlie the melting of bump attractor of overlapping representations. The introduction of

STSF might then be beneficial to overcome this issue for moderately overlapping representations while it is likely to result in increasing $K_e$ values as observed when discrete representations are considered.

The literature also shows certain variability with regard to the nature of the serial effects reported in sequential protocols. Whereas some authors highlight the emergence of the recency effect and show no reliable primacy effects [26, 29, 32], other authors report experimental results whereby both of them emerge [30, 31, 33, 36]. Interestingly, Lee et al. [43] propose the existence of two types of codings (i.e. stable and flexible), which take the distinct roles of retaining and updating information in WM. In their view, stable encoding of information favours the retention of previous memory and, hence, results in the primacy effect. In contrast, flexible encoding enables update of recent memory, thus resulting in the recency effect. They claim that it is the combination of both types of codings that induces the emergence of the joint appearance of the primacy and recency serial-position effects. We suggest that the heterogeneity in the $\tau_F$ values observed in PFC may underlie the combination of serial effects. Indeed, large $\tau_F$ ($\tau_F > 1200$ ms) values lead to a type of coding akin to the stable one whereas smaller values ($\tau_F < 1000$ ms) lead to the so-called flexible coding. The concerted action of $\tau_F$ (a parameter broadly grounded on neurophysiology) and the variables $\Delta t_{stim}$ and $\Delta\lambda$ (closely related to the specific WM task and its associated experimental protocol) leads to the broad repertoire of results found in the literature.

To summarise, we propose that the neurodynamical basis of WM involves persistent activity during the delay period. However, we acknowledge that this is not the only mechanism underlying WM. In fact, we do believe that most likely both activity-silent and activity-based mechanisms coexist and play complementary roles in WM function. Nevertheless, it is appealing that, by itself, the proposed model is able to account for a large variety of experimental results in a highly principled way. Moreover, we claim that STSF is a key mechanism to account for the dynamic allocation of resources in multi-item WM. Further research will be needed to explore the relevance of each type of mechanism and how they cooperate in the brain to shape WM function.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12559-023-10234-4.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work is partially supported by the European Cooperation in Science and Technology COST Action CA1806.

**Data Availability** Data availability is not applicable to this article as no new data were created.

## Declarations

**Ethics Approval** This article does not contain any studies with human participants performed by any of the authors.

**Conflict of Interest** The authors declare no competing interests.



## References

1. Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci. 2000;24(1):87–114.
2. Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. Nature. 2008;453:233–5.
3. Bays PM, Husain M. Dynamic shifts of limited working memory resources in human vision. Science. 2008;321(5890):851–4.
4. Adam KC, Vogel EK, Awh E. Clear evidence for item limits in visual working memory. Cogn Psychol. 2017;97:79–97.
5. Fougnie D, Suchow JW, Alvarez GA. Variability in the quality of visual working memory. Nat Commun. 2012;3(1):1–8.
6. Fuster JM, Alexander G. Neuron activity related to short-term memory. Science. 1971;173(3997):652–4.
7. Constantinidis C, Funahashi S, Lee D, Murray JD, Qi XL, Wang M, Arnsten AFT. Persistent Spiking Activity Underlies Working Memory. J Neurosci. 2018;38(32):7020–8.
8. Lundqvist M, Herman P, Miller EK. Working memory: delay activity, yes! Persistent activity? Maybe not. J Neurosci. 2018;38(32):7013–9.
9. Mongillo G, Barak O, Tsodyks M. Synaptic theory of working memory. Science. 2008;319(5869):1543–6.
10. Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb Cortex. 2000;10(9):910–23.
11. Brunel N, Wang XJ. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comp Neurol. 2001;11(1):63–85.
12. Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A. Mechanisms for top-down control of working memory capacity. Proc Natl Acad Sci USA. 2009;106(16):6802–7.

13. Dempere-Marco L, Melcher DP, Deco G. Effective visual working memory capacity: an emergent effect from the neural dynamics in an attractor network. PLoS ONE. 2012;7(10). https://doi.org/10.1371/journal.pone.0042719
14. Rolls ET, Dempere-Marco L, Deco G. Holding multiple items in short-term memory: a neural mechanism. PLoS ONE. 2013;8(4):e61078. https://doi.org/10.1371/journal.pone.0061078.
15. Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci. 2014;17:431–9.
16. Almeida R, Barbosa JA, Compte A. Neural circuit basis of visuospatial working memory precision: a computational and behavioral study. J Neurophysiol. 2015;114(3):1806–18.
17. Lisman JE, Idiart MAP. Storage of 7 ± 2 short-term memories in oscillatory subcycles. Science. 1995;267:1512–5.
18. Neymotin SA, McDougal RA, Bulanova AS, Zeki M, Lakatos P, Terman D, Hines ML, Lytton WW. Calcium regulation of HCN channels supports persistent activity in a multiscale model of neocortex. Neuroscience. 2016;316.
19. Zylberberg J, Strowbridge BW. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. Annu Rev Neurosci. 2017.
20. Majerus S, Attout L. Working memory for serial order and numerical cognition: what kind of association? In: Henik A, Fias W, editors. Heterogeneity of Function in Numerical Cognition. Academic Press; 2018. p. 409–31.
21. Baddeley AD, Hitch GJ, Allen R. A multicomponent model of working memory. Working memory: State of the Science. 2021;10–43.
22. Irwin DE, Brown JS, Sun JS. Visual masking and visual integration across saccadic eye movements. J Exp Psychol Gen. 1988;117(3):276–87.
23. Chun MM, Xu Y. Dissociable neural mechanisms supporting visual short-term memory for objects. Nature. 2006;440(7080):91–5.
24. Lecerf T, Ribaupierre A. Recognition in a visuospatial memory task: the effect of presentation. Eur J Cogn Psychol. 2005;17(1):47–75.
25. Blalock LD. Clegg BA. Encoding and representation of simultaneous and sequential arrays in visuospatial working memory. Q J Exp Psychol. 2010;63:856–62.
26. Gorgoraptis N, Catalao RFG, Bays PM, Husain M. Dynamic updating of working memory resources for visual objects. J Neurosci. 2011;31(23): 8502–8511.
27. Hurlstone MJ, Hitch GJ, Baddeley AD. Memory for serial order across domains: an overview of the literature and directions for future research. Psychol Bull. 2014;140(2):339–73.
28. Oberauer K, Lewandowsky S, Awh E, Brown GDA, Conway A, Cowan N, Donkin C, Farrell S, Hitch GJ, Hurlstone MJ, Ma WJ, Morey CC, Nee DE, Schweppe J, Vergauwe E, Ward G. Benchmarks for models of short-term and working memory. Psychol Bull. 2018;144(9):885–958. https://doi.org/10.1037/bul0000153. PMID: 30148379.
29. Broadbent DE, Broadbent MHP. Recency effects in visual memory. Q J Exp Psychol. 1981;33A:1–15.
30. Avons SE. Serial report and item recognition of novel visual patterns. Br J Psychol. 1998;89:285–308.
31. Smyth MM, Hay DC, Hitch GJ, Horton NJ. Serial position memory in the visual-spatial domain: reconstructing sequences of unfamiliar faces. Q J Exp Psychol A. 2005;58:909–30.
32. Hay DC, Smyth MM, Hitch GJ, Horton NJ. Serial position effects in short-term visual memory: A simple explanation? Mem Cognit. 2007;35(1):176–90.
33. Ward G, Avons SE, Melling L. Serial position curves in short-term memory: functional equivalence across modalities. Memory. 2005;13(3–4):308–17.
34. Yakovlev V, Bernacchia A, Orlov T, Hochstein S, Amit D. Multi-item working memory - a behavioral study. Cereb Cortex. 2005;15:602–15.
35. Zucker RS, Regehr WG. Short-term synaptic plasticity. Annu Rev Physiol. 2002;64:355–405.
36. Kool W, Conway ARA, Turk-Browne NB. Sequential dynamics in visual short-term memory. Atten Percept Psychophys. 2014;76:1885–901.
37. Jackman SL, Regehr WG. The mechanisms and functions of synaptic facilitation. Neuron. 2017;94(3):447–64.
38. Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat Neurosci. 2006;9;534–542
39. Mi Y, Katkov M, Tsodyks M. Synaptic correlates of working memory capacity. Neuron. 2017;93:32–330.
40. Buschman TJ, Siegel M, Roy JE, Miller E. Neural substrates of cognitive capacity limitations. Proc Natl Acad Sci USA. 2011;108:11252–5.
41. Stokes MG. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci. 2015;19(7):394–405.
42. Seeholzer A, Deger M, Gerstner W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. PLoS Comput Biol. 2019;15(4):1–48.
43. Lee H, Choi W, Park Y, Paik SB. Distinct role of flexible and stable encoding in sequential working memory. Neural Netw. 2020;121:419–29.