Exact neural mass model for synaptic-based working memory

Halgurd Taher[1], Alessandro Torcini[2,3], Simona Olmi[1,3,*],

1 Inria Sophia Antipolis Méditerranée Research Centre, MathNeuro Team, 2004 route des Lucioles, Boîte Postale 93. 06902 Sophia Antipolis cedex, France
2 Laboratoire de Physique Théorique et Modélisation, Université de Cergy-Pontoise,CNRS, UMR 8089, 95302 Cergy-Pontoise cedex, France
3 CNR - Consiglio Nazionale delle Ricerche - Istituto dei Sistemi Complessi, 50019, Sesto Fiorentino, Italy

* simona.olmi@inria.fr

Abstract

A synaptic theory of Working Memory (WM) has been developed in the last decade to overcome several drawbacks related to the persistent spiking paradigm. Memory items are maintained in the synaptic facilitation and refreshed thanks to Population Bursts (PBs) promoted by synaptic depression. We have developed a neural mass model able to reproduce exactly the dynamics of heterogeneous spiking neural networks with short-term synaptic plasticity. This neural mass model gives access not only to the population firing rate, but also to the mean membrane potential. The latter is correlated to the Local Field Potentials and to electroencephalographic signals, usually measured during WM tasks to characterize the brain dynamics. The model is able to mimic several operations required by WM : memory loading, memory maintenance, competition of memory items and multi-item memory juggling. Memory access and loading is associated to self-sustained oscillations in the β-γ band and to transient evoked responses in the δ band, analougously to what observed in the prefrontal cortex of monkeys during object recognition tasks. More items can be stored in the WM by considering neural architectures composed by multiple excitatory populations and a common inhibitory pool. The items can be loaded and maintained at the same time: they are represented as trains of PBs with the same period, but delivered at evenly shifted phases. A memory competition is observable already by loading only two items. Depending on the stimulation features we can observe three different outcomes: maintenance of the first or second loaded item or their juggling in the memory. As reported in many experiments, the γ power increases with the number of loaded items. The power in the α-β instead shows a non monotonic behaviour. Finally, we report an expression for the maximal capacity adapted to the present model and architecture.

Author summary

Working Memory (WM) is the ability to temporary store and manipulate stimuli representations that are no longer available to the senses. WM can be represented in terms of spiking neural networks with synaptic facilitation and depression. Here, we present a neural mass model able to reproduce exactly the macroscopic dynamics of the network in terms of few collective variables. Memory items are transiently held as traces in the short-term synaptic facilitation. Memory refreshment is achieved thanks to bursts of neuronal activity controlled by synaptic depression. Memory operations are joined to

sustained or transient oscillations emerging in different frequency bands, in accordance with experimental results. We show that memory maintenance of several items can be achieved thanks to periodic trains of bursts delivered by different excitatory populations at preferential phases. This model can represent the first building block for the realization of more realistic multi-layer architectures.

Introduction

Working memory (WM) is the ability to keep recently accessed information, available for manipulation: it is fundamental in order to pass from reflexive input-output reactions to the organization of goal-directed behaviour [1–6]. Starting from the seminal work of Fuster and Alexander [7], several experiments have shown that neurons in higher-order cortex, including the prefrontal cortex (PFC), are characterized by elevated levels of spiking activity during memory delays related to WM tasks. Experimental results seem to suggest that WM maintenance is enhanced by delayed spiking activity, engaging executive functions associated with large part of the cortex, from frontal to posterior cortical areas [8, 9].

Therefore, classic models proposed persistent spiking for online maintenance of information, since neural ensembles in an active state appear more prone for fast processing. However, the evidence for persistent state results from specific data processings: neural spiking being averaged over time and across trials. In particular averaging across trials can create the appearance of persistent spiking even when, on individual trials, spiking is actually sparse [10, 11]. Anyhow there are examples of single neurons that show real persistent spiking on individual trial rasters, but they represent a small percentage, while the bulk of neurons spike sparsely during WM delays [12, 13].

A pioneering study [14] revealed that the interactions among pyramidal neurons in the PFC display synaptic facilitation lasting hundreds of milliseconds. This study paved the way for a development of an alternative model for the WM based on synaptic features. In particular, memory items can be stored by spiking-induced changes in the synaptic weights and these items can be maintained in the WM by brief bursts of spiking restoring the synaptic strengths [15–18]. On one hand this kind of model is more justifiable from a metabolic point of view, since the memory requires less resources to be held with respect to the persistent activity. On the other hand, multiple memory items can be maintained in the WM at the same time by having different neuronal populations emitting short bursts at different times. This would solve problems usually observables in models based on persistent spiking: namely, the interference among different stored items and the disruption of memories associated to new sensory inputs [18, 19].

Specifically, we focus on the WM model based on short-term synaptic plasticity (STP) introduced by Mongillo *et al.* in [15]. In this model synapses among pyramidal neurons display depressed and facilitated transmission, described by a phenomenological model of STP developed by Markram, Tsodyks and collaborators [20–22]. Synaptic facilitation allows to maintain an item stored for a certain period in the WM, without the need of an enhanced spiking activity. Furthermore, synaptic depression is responsible for the emergence of population bursts (PBs), which correspond to a sub-population of neurons firing almost synchronously within a short time window [23, 24]. This WM mechanism was implemented in [15] within a recurrent network of spiking neurons, while a simplified rate model was employed to gain some insight into the population dynamics. This rate model, as well as the most part of neural mass models investigated in the literature [15, 25, 26], is heuristic, i.e. the macroscopic description has not an exact correspondence with the microscopic neuronal evolutions. In particular, the neural variability is usually neglected.

Recent results for the PFC have revealed that the mnemonic stimuli are encoded at

a population level allowing a stable and robust WM maintenance, despite that single neurons show heterogeneous temporal dynamics [27]. Therefore, in order to shed further light on the synaptic mechanisms at the basis of the WM, the development of neural mass models able to reproduce exactly the macroscopic dynamics of heterogeneous spiking networks with STP can be extremely useful. This represents the main scope of our paper.

In particular, our analysis will focus on neural networks of Quadratic Integrate and Fire (QIF) neurons for two main reasons. First of all, the QIF model can be put in direct correspondence with the θ-neuron, which represents the normal form of Hodgkin's class I excitable membranes [28]. Therefore we expect that the QIF model reproduces the dynamical behaviour of all class I neurons in proximity to the saddle-node on a limit cycle bifurcation. Secondly, thanks to analytic techniques developed for a network of phase oscillators [29], the macroscopic dynamics of heterogeneous QIF networks can be reproduced exactly [30–32]. This new generation of neural mass models has been recently used to describe the emergence of collective oscillations (COs) in fully coupled networks [33–36] as well as in balanced sparse networks [37]. Furthermore, it has been also successfully employed to reveal the mechanisms at the basis of theta-nested gamma oscillations [38,39] and of the coexistence of slow and fast gamma oscillations [40]. However, to our knowledge, such models have been not yet generalized to spiking networks with short-term synaptic plasticity [23].

In this paper we revise the synaptic theory of working memory [15] by employing a neural mass model able to exactly reproduce the network dynamics of spiking neurons at any degree of synchronization [32]. Moreover, we extend such mean-field approach to networks where excitatory-excitatory synapses display short-term facilitation and depression [21]. The derived neural mass model describes the macroscopic evolution of the network not only in terms of the firing rate, as the standard heuristic models [25], but also of the mean membrane potential [32]. This allows for a more direct comparison with the results of electrophysiological experiments, where WM processes are often related to characteristics of electroencephalograms (EEGs) and Local Field Potentials (LFPs), which have the same information content as the mean membrane potentials [41].

The paper is organised as follows. Firstly, we will report clear evidences that the employed neural mass model reproduces exactly the dynamics of a heterogeneous network of QIF neurons with STP at a macroscopic level. Secondly, we will show that the model is able to reproduce operations required for the WM functioning. Then we will give a detailed analysis on how two consecutively loaded memory items compete in the model, whenever the memory maintenance is realised in terms of either synaptic facilitation or persistent spiking. Moreover, we will identify which frequency bands are excited during loading and maintenance of multiple items in the WM. We will also report an analytic estimation of the maximal working memory capacity of our model, by extending recent results reported in [26]. Finally, we will conclude with a discussion of the obtained results and of their possible relevance for the field of neuroscience.

Results

We develop a model for WM able to memorize discrete items by following [42], in particular we consider N_{pop} coupled excitatory populations each coding for one item, and a single inhibitory population connected to all the excitatory neurons. This architecture is justified by recent experimental results indicating that GABAergic interneurons in mouse frontal cortex are not arranged in sub-populations and that they densely innervate all pyramidal cells [43]. The role of inhibition is to avoid abnormal synchronization and to allow for a competition of different items once stored in the excitatory population activity. Furthermore, in order to mimic synaptic-based WM we

assume that only the excitatory-excitatory synapses are plastic displaying short-term depression and facilitation [15].

The macroscopic activity of $N_{\rm pop}$ excitatory and one inhibitory populations of heterogeneous QIF neurons can be exactly described in terms of the population firing rate $r_k(t)$ and mean membrane voltage $v_k(t)$ of each population k by the following set of ODEs [32]:

$$\tau_{\rm m}^n \dot{r}_k = \frac{\Delta_k}{\tau_{\rm m}^n \pi} + 2 r_k v_k \qquad k = 0, 1, \ldots, N_{\rm pop} \tag{1a}$$

$$\tau_{\rm m}^n \dot{v}_k = v_k^2 + H_k - (\pi \tau_{\rm m}^n r_k)^2 + I_{\rm B} + I_{\rm S}^{(k)}(t) + \tau_{\rm m}^n \sum_{l=0}^{N_{\rm pop}} \tilde{J}_{kl}(t) r_l \quad ; \tag{1b}$$

where $\tau_{\rm m}^n$ denotes the membrane time constant of excitatory (inhibitory) populations $n = {\rm e}$ ($n = {\rm i}$) and $I_{\rm B}$ is a constant background current common to all populations, whereas the (time dependent) stimulus current $I_{\rm S}^{(k)}(t)$ may be population specific. Excitatory populations are characterized by $k > 0$, the inhibitory by $k = 0$. In absence of STP and for instantaneous synapses the synaptic weights are constant in time $\tilde{J}_{kl}(t) = J_{kl}$ and their sign determines whether the connections are excitatory ($J_{kl} > 0$) or inhibitory ($J_{kl} < 0$). The heterogeneous nature of the neurons is taken into account by considering randomly distributed neural excitabilities, reflecting their biological variability. In particular, we assumed that the distribution of the neural excitabilities is a Lorentzian characterized by a median H_k and a half width half-maximum (HWHM) Δ_k. This choice allows to derive the neural mass model from the spiking QIF network analytically. However, the overall dynamics does not change substantially by considering other distributions for the neural excitabilities, like Gaussian and Erdös-Renyi ones [32, 40].

Plasticity enables synapses to modify their efficacy in response to spiking activity and hence to stimuli. As opposed to long-term plasticity, STP occurs on short time scales in the order of hundred milliseconds to a few seconds. In particular one distinguishes between synaptic depression, i.e. the weakening, and facilitation, i.e. the strengthening of synapses. The phenomenological model for neocortical synapses, introduced by Tsodyks and Markram to reproduce the synaptic depression of pyramidal neurons [20], assumes that each synapse is characterized by a finite amount of resources: each incoming pre-synaptic spike activates a fraction of resources, which rapidly inactivates in few milliseconds and recover from synaptic depression over a longer time-scale $\tau_{\rm d}$. The formulation of a facilitating mechanism has been subsequently developed by Tsodyks, Pawelzik and Markram [21]: in this case, the fraction of resources is increased at each pre-synaptic spike and returns to the baseline value with a time constant $\tau_{\rm f}$. The idea behind these phenomenological STP models is that the magnitude of post-synaptic potentials (PSPs) is weighted by the relative amount $X_i \in [0, 1]$ of available pre-synaptic resources and by the fraction $U_i \in [0, 1]$ of these resources utilized to generate the PSP. Whenever the neuron i emits a spike, neurotransmitters are released and X_i decreases, while at the same time calcium accumulates in the pre-synaptic terminals, increasing the neurotransmitter release probability at the next spike emission and hence the fraction U_i of utilized resources. This form of STP is purely pre-synaptic, i.e., depression and facilitation are influenced only by spikes emitted by the pre-synaptic neuron, this justifies why we can assume that the state of all the efferent synapses of a neuron i is described by only the single couple of variables (X_i, U_i) [44]. Furthermore, it is fundamental that the facilitation timescale $\tau_{\rm f}$ is longer than that associated to the recover from depression $\tau_{\rm d}$ [15]. In this way the information provided by the stimulus will be carried over a time $\tau_{\rm f}$ by the facilitated synapse.

The inclusion of the depression and facilitation STP mechanisms in the spiking

networks is reflected in the modification of the synaptic weights entering in the neural mass model (Eqs. (1)). In particular, these are rewritten as $\tilde{J}_{kl}(t) = J_{kl}u_l(t)x_l(t)$ if k and l are both excitatory populations, and simply as $\tilde{J}_{kl}(t) = J_{kl}$ if either k or l is inhibitory. The terms $x_k(t)$ and $u_k(t)$ represent the mean available resources and the mean utilization factor of the population k, respectively. If we neglect the very fast inactivation of the depression terms, the dynamical evolution of x_k and u_k is regulated by the following ODEs

$$\frac{\mathrm{d}x_k(t)}{\mathrm{d}t} = \frac{1 - x_k(t)}{\tau_\mathrm{d}} - u_k(t)x_k(t)r_k(t) \tag{2a}$$

$$\frac{\mathrm{d}u_k(t)}{\mathrm{d}t} = \frac{U_0 - u_k(t)}{\tau_\mathrm{f}} + U_0(1 - u_k(t))r_k(t) \qquad k = 1, \ldots, N_\mathrm{pop} \quad ; \tag{2b}$$

where $U_0 = 0.2$ is the baseline value of the utilization factor, while $\tau_\mathrm{d} = 200$ ms, $\tau_\mathrm{f} = 1500$ ms represent the depression and facilitation time scales, respectively. These parameter values are fixed for all the simulations reported in this paper.

The model (1) - (2) describes the dynamics of one inhibitory and N_pop excitatory coupled neuronal populations with STP in terms of their population firing rates r_k, mean membrane voltages v_k, mean available resources x_k and mean utilization factors u_k. Details on the derivation of the neural mass model Eqs. (1) - (2) and on the underlying spiking networks can be found in the sub-sections Spiking neuronal network model, Exact neural mass model with STP, Multi-populations models in Methods.

Network dynamics versus neural mass evolution

The neural mass model (1) reproduces exactly the population dynamics of a network of QIF spiking neurons in the infinite size limit as analytically demonstrated and numerically verified in [32]. In order to verify if this statement is valid also in presence of STP, we will compare for a single excitatory neural population the macroscopic evolution given by Eqs. (1) and (2) with those obtained by considering QIF networks with synaptic plasticity implemented at a microscopic (μ-STP) and at a mesoscopic (m-STP) level. The results of these comparisons are reported in Fig. 1: column A (B) for μ-STP (m-STP).

A spiking network of N neurons with μ-STP is described by the evolution of $3N$ variables, since each neuron is described in terms of its membrane voltage and the two synaptic variables $U_i(t)$ and $X_i(t)$ accounting for the dynamics of its N efferent synapses. The explicit dynamical model is reported in Eqs. (5), (9) and (10) in sub-section Short-term synaptic plasticity in Methods. In the m-STP model, the dynamics of all synapses is treated as a mesoscopic variable in terms of only two macroscopic variables, namely $u(t)$ and $x(t)$. In this case the network model reduces to a set of $N + 2$ ODEs, namely Eqs. (5) and (11) reported in sub-section Short-term synaptic plasticity in Methods.

In order to compare the different models we examine their response to the same delivered stimulus $I_\mathrm{S}(t)$, which consists of two identical rectangular pulses of height $I_\mathrm{S} = 2$ and duration 0.15 s separated by a time interval of 0.15 s (see Fig. 1 (A2-B2)). The stimulus is presented when the system is in a quiescent state, i.e. it is characterized by a low firing rate. For both comparisons the neural mass model has been integrated starting from initial values of r,v,x and u as obtained from the microscopic state of the considered networks (for more explanations see sub-section Exact neural mass model with STP in Methods). The response is the same in both network models: each pulse triggers a series of four PBs of decreasing amplitude followed by low activity phases in absence of stimuli (see Fig. 1 (A1-B1)). As shown in Fig. 1 (A3-B3), the population firing rate $r(t)$ obtained from the neural mass model (solid line) is almost coincident

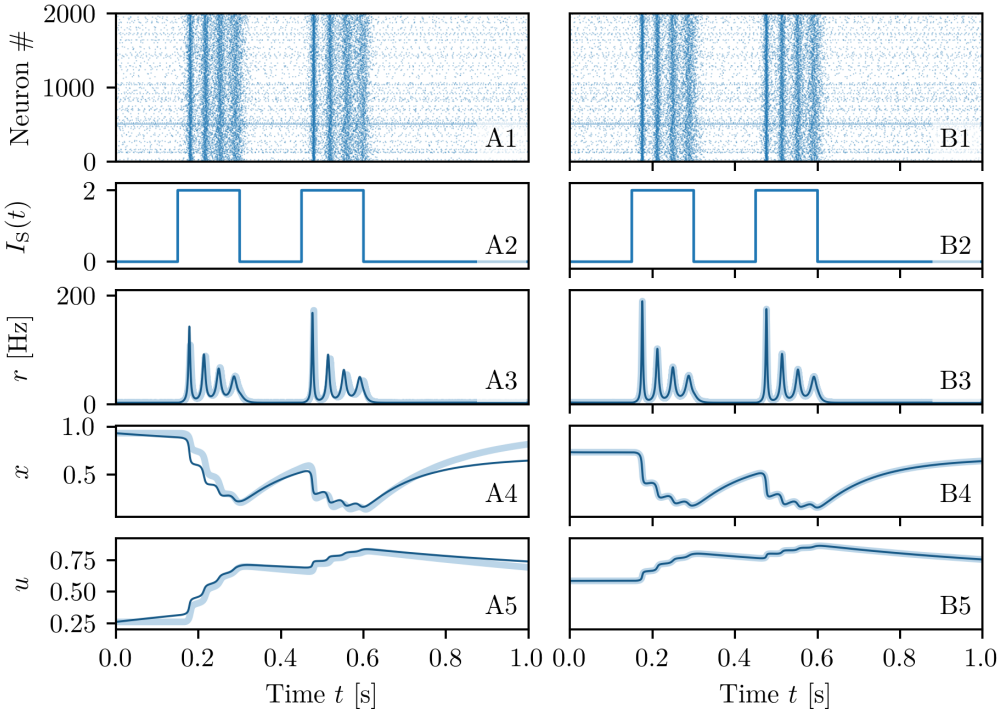


Fig 1. Comparison among neural mass and network models The results of the neural mass model (solid line) are compared with the network simulations (shaded lines) for a single excitatory population with μ-STP (column A) and with m-STP (column B). The corresponding raster plots for a subset of 2000 neurons are reported in (**A1-B1**), the profiles of the stimulation current $I_S(t)$ in (**A2-B2**), the instantaneous population firing rates $r(t)$ in (**A3-B3**), the available synaptic resources $x(t)$ in (**A4-B4**), and the utilization factors $u(t)$ in (**A5-B5**). Simulation parameters are $\tau_m^e = 15$ ms, $H = 0$, $\Delta = 0.25$, $J = 15$, $I_B = -1$ and network size $N = 200,000$.

with the results of the network simulations both with μ-STP and m-STP (shaded curves). For all models, the average synaptic variables x and u displayed in Fig. 1 (A4-B4) and (A5-B5) reveal the typical temporal evolution of STP upon excitatory stimulation: the available resources x (utilization factor u) decrease (increases) due to the series of emitted PBs in a step-wise fashion. Furthermore, in absence of stimulation x and u tend to recover to their stationary values $x = 1$ and $u = U_0$ for μ-STP ($x = 0.73$ and $u = 0.59$ for m-STP) over time-scales dictated by τ_d and τ_f, respectively. Due to the fact that $\tau_d \ll \tau_f$, the synapses remain facilitated for a time interval $\simeq 1$ s after the pre-synaptic resources have recovered from depression. The time courses of x and u for the neural mass compared to the network ones, shown in Fig. 1 (A4-B4) and (A5-B5), reveal an almost perfect agreement with the simulations of the network with m-STP, while some small discrepancies are observable when compared with the network with μ-STP. This is due to the fact that the correlations and fluctuations of the microscopic variables $\{X_i\}$ and $\{U_i\}$ are not taken into account in the mesoscopic model for STP [21], as pointed out in [45] (see also the discussion on this aspect reported in sub-section Short-term synaptic plasticity in Methods). We can conclude this sub-section by affirming that the developed neural mass model reproduces in detail the macroscopic dynamics of spiking networks of QIF neurons with both μ-STP and m-STP, therefore in the following we can safely rely on the mean-field simulations to explore the synaptic mechanisms at the basis of WM.

Multi-Item Architecture

In order to illustrate the architecture employed to store more items at the same time in the WM, we will analyse the simplest non-trivial situation where we want to store two items. For this purpose, we consider two excitatory populations with STP employed to store one item each, and an inhibitory one, necessary to allow for item competition and a homoeostatic regulation of the firing activity. In particular, we restrict to non-overlapping memories: neurons belonging to a certain excitatory population encode only one working memory item. Furthermore, incoming information on a WM item, in form of an external stimulus, target only the excitatory population which codes for that item, hence making the response selective

As previously mentioned, only excitatory-excitatory synapses are plastic, therefore we will have no time dependence for synaptic strength involving the inhibitory population. Moreover, the WM items should be free to compete among them on the same basis, therefore we assume that the synaptic couplings within a population of a certain type (inhibitory or excitatory) and among populations of the same type are identical. In summary, we have

$$\begin{aligned} &\tilde{J}_{00}(t) = J_{ii}, \;\; \tilde{J}_{0k}(t) = J_{\mathrm{ie}}, \;\; \tilde{J}_{k0}(t) = J_{\mathrm{ei}}, \;\; \text{for } \; k > 0 \quad ; \\ &\tilde{J}_{kk}(t) = J_{\mathrm{ee}}^{(\mathrm{s})} x_k(t) u_k(t), \;\; J_{kj}(t) = J_{\mathrm{ee}}^{(\mathrm{c})} x_j(t) u_j(t), \quad \forall k, j > 0 \end{aligned} \tag{3}$$

where we have denoted inhibitory (excitatory) populations by i (e) index. Furthermore, synaptic connections within each excitatory population, coding for a certain memory, will be stronger than connections between different excitatory populations, thus we assume $J_{\mathrm{ee}}^{(\mathrm{s})} > J_{\mathrm{ee}}^{(\mathrm{c})}$. A common background input current I_{B} impinges on all three populations, thus playing a crucial role in controlling the operational point of the network. The selective storage and retrieval of memory items induced by incoming stimuli is mimicked via time dependent external currents $I_{\mathrm{S}}^{(k)}(t)$ applied to the excitatory populations. The discussed multi-item architecture is displayed in Fig. 2.

Due to the fact that each excitatory population is indistinguishable from the others and we have only one inhibitory pool, we set for clarity $H_0 = H^{(\mathrm{i})}$, $H_k = H^{(\mathrm{e})} \;\; \forall k > 0$ and $\Delta_0 = \Delta^{(\mathrm{i})}$, $\Delta_k = \Delta^{(\mathrm{e})} \;\; \forall k > 0$. Furthermore, in order to better highlight the variations of the synaptic parameters, in the following we report normalized available resources $\tilde{x}_k$ and normalized facilitation factors $\tilde{u}_k$ [1].

Memory maintenance with synaptic facilitation

To show the capability of our model to reproduce WM activity we report in this sub-section three different numerical experiments. In particular, we will show that the model is able to reproduce specific computational operations required by the WM: namely, memory load, memory maintenance during a delay period, selective or spontaneous reactivation of the memory and memory clearance.

Let us start from an example of selective reactivation of the target item via a non specific read-out signal, targeting both excitatory populations, as shown in the left column (A) of Fig. 3. Since the observed dynamics does not depend on the population into which an item is loaded, without loss of generality we always load the item in population one (coded by the blue colour in Fig. 3). The system is initialized in an asynchronous state of low firing activity common to both populations (as shown in panel (A5)), therefore the synaptic variables have equal values (namely, $\tilde{x}_1 = \tilde{x}_2 = 0.79$ and $\tilde{u}_1 = \tilde{u}_2 = 0.26$) dictated by the average firing rate (see panels (A6-A7)).

[1] These parameters are normalized with respect to the difference of their maximal and minimal value taken within the displayed time interval.

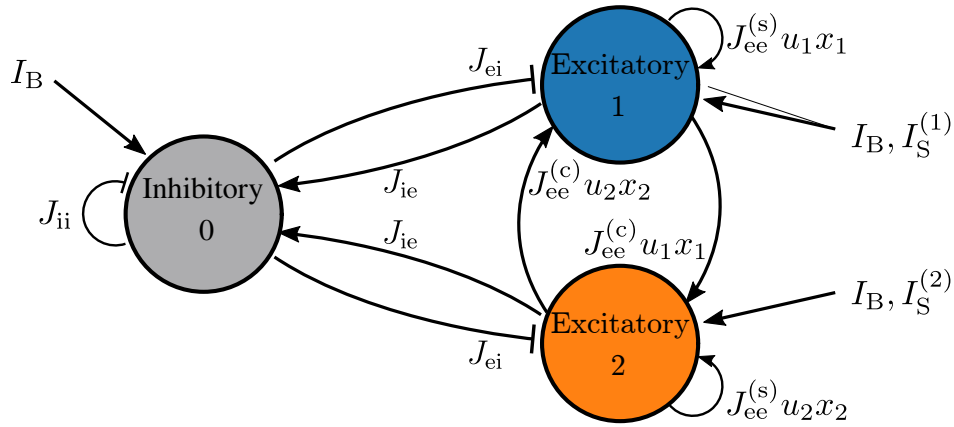


Fig 2. Two-item architecture The reported network is composed of two identical and symmetrically coupled excitatory populations and an inhibitory one: this architecture can at most store two WM items, one for each excitatory population.

The memory load is performed at time $t = 0$ s, by delivering a stimulation current in the form of a step with amplitude $\Delta I^{(1)} = 0.2$ and width $\Delta T_1 = 350$ ms only to population one (panel (A2)). In response to this stimulus, the population displays PBs, as shown by the raster plot in panel (A1). The frequency of these collective oscillations (COs) is in the β band (namely, $\simeq 21.6$ Hz), as evident from the spectrogram in panel (A8). The PBs are clearly generated via a pyramidal interneuronal network gamma (PING) mechanism [46], since the PBs in population one trigger also PBs in the inhibitory population with a delay $\simeq 1 - 2$ ms, while the same frequency is maintained, as evident from the inhibitory population spectrogram in panel (A10).

The intense firing of neurons in population one changes the internal state of its efferent synapses, thus leading to an initial drop (increase) of x_1 (u_1) (panels (A6-A7)). Depression prevails on short time scales $\simeq \tau_{\rm d}$, while at longer times $t > 0.5$ s, $\tilde{x}_1$ recovers to its initial value and the synapses remain facilitated, with $\tilde{u}_1 > 0.73$, for one or two seconds (see panel (A7)).

Population two (denoted by the orange colour in Fig. 3) is not particularly affected by the memory loading. Indeed, population two shows only a slight reduction in its asynchronous activity, which is reflected in a small increase (decrease) of $\tilde{x}_2$ ($\tilde{u}_2$) (see panels (A3) and (A6-A7)). This is due to the action of the inhibitory bursts, which are unable to support PBs in population two, but sufficient to modulate it's asynchronous activity as revealed by the spectrogram in panel (A9).

Since the information on the initial stimulus is stored in the facilitation of the synapses, we expect that, by presenting a weak non-specific read-out signal to both populations, the memory will be reactivated, even if the population activity is back to the spontaneous level. Indeed, if after a delay period of 1.2 s from the initial memory load, we inject a weak stimulation current of amplitude 0.1 for a time interval $\Delta T = 250$ ms to all the excitatory neurons, we observe that only neurons of the blue population respond by emitting a PB of brief duration. The other neurons not associated to the loaded item remain at the baseline activity level, despite the stimulus (panel (A3)). The emission of the PB in population one has also the effect to refresh the memory, i.e. the utilization factor u_1 which has decreased towards the initial value during the delay period, returns to the value reached when the item was loaded. Therefore the memory can be maintained for a longer time period.

In a second example, shown in the central column (B) of Fig. 3, we consider a case where the target population reactivates spontaneously by emitting a regular series of PBs even in absence of read-out signals. To obtain such situation we increase the

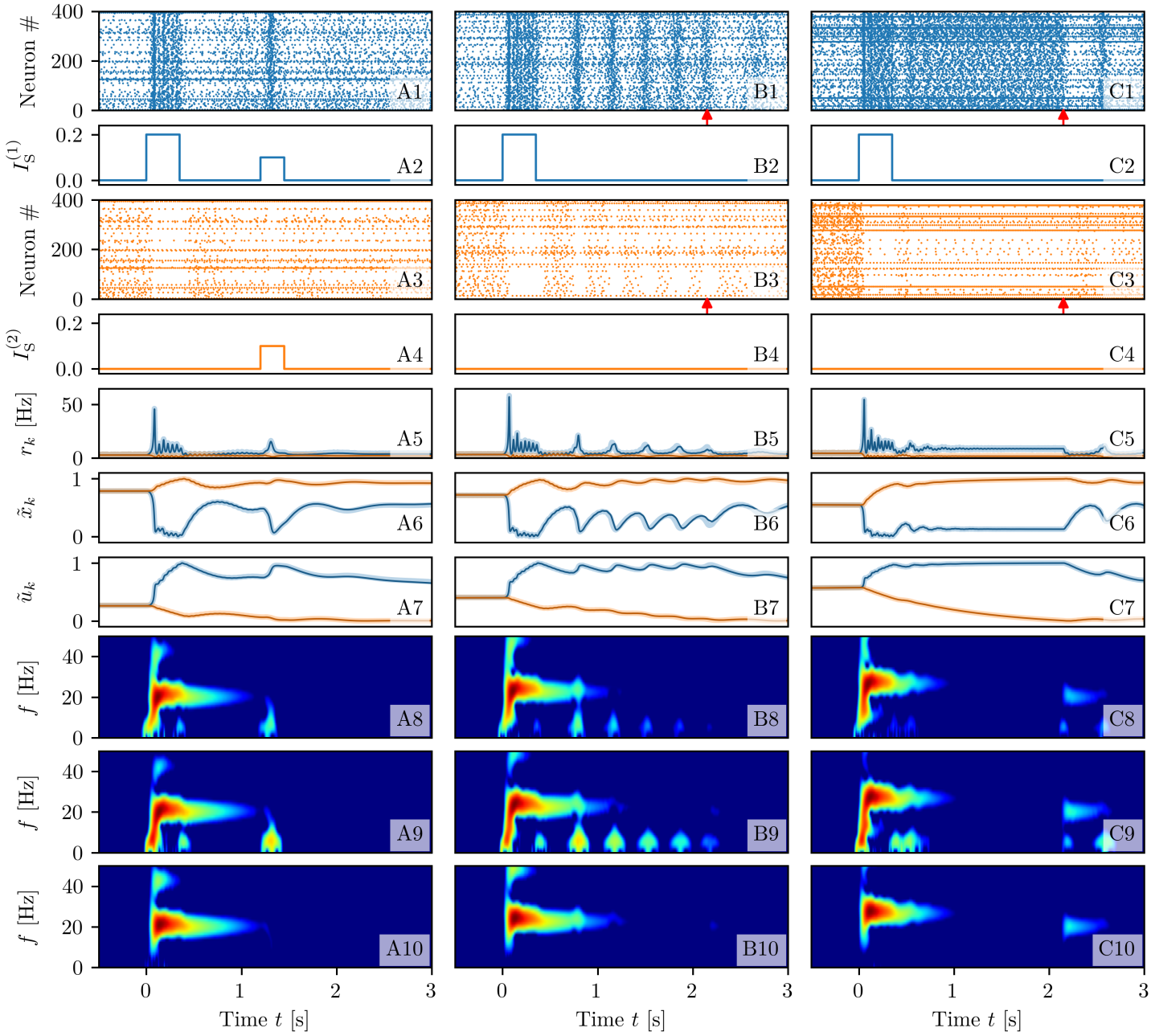


Fig 3. Memory loading, maintenance and rehearsal The results of three experiments are reported here for different background currents: selective reactivation of the target population ($I_B = 1.2$, A); WM maintenance via spontaneous reactivation of the target population ($I_B = 1.532$, B) and via a persistent asynchronous activity ($I_B = 2$, C). Raster plots of the network activity for the first (blue, **A1-C1**) and second (orange, **A3-C3**) excitatory population. Profiles of the stimulation current $I_S^{(k)}(t)$ for the first (**A2-C2**) (second (**A4-C4**)) excitatory population. Population firing rates r_k(t) (**A5-C5**), normalized available resources $\tilde{x}_k(t)$ (**A6-C6**) and normalized utilization factors $\tilde{u}_k(t)$ (**A7-C7**) of the excitatory populations calculated from the simulations of the neural mass model (solid line) and of the network (shading). Spectrograms of the mean membrane potentials $v_1(t)$ (**A8-C8**), $v_2(t)$ (**A9-C9**), and $v_0(t)$ (**A10-C10**) obtained from the neural mass model. Red arrows in columns (B) and (C) indicate the time $t = 2.15$ s at which the background current is set to the value $I_B = 1.2$ employed in column (A). Other parameters: $H^{(i)} = H^{(e)} = 0$, $\Delta^{(i)} = \Delta^{(e)} = 0.1$, $J_{ee}^{(c)} = 5\sqrt{a}$, $J_{ee}^{(s)} = 35\sqrt{a}$, $J_{ie} = 13\sqrt{a}$, $J_{ei} = -16\sqrt{a}$, $J_{ii} = -14\sqrt{a}$ with $a = 0.4$.

background signal up to $I_B = 1.532$, thus the system is in a regime where the spontaneous low activity state coexists with states in which one of the excitatory populations periodically emits PBs (for more details on the emerging states see the sub-section Bifurcation analysis within the section Methods). As in the previous experiment, the system is initialized in a low firing activity state, where only spontaneous activity is present. The memory item is loaded in population one (see panels (B1-B5)). In this case population one encodes the item by emitting a series of PBs with a slightly higher frequency ($\simeq 24.1$ Hz) with respect to the previous case, due to the increase in I_B, but still in the β range (as shown in panels (B8-B10)). After a short time delay, PBs emerge regularly in a self-sustained manner: each reactivation leads to an increase of $\tilde{u}_1$ and a decrease of $\tilde{x}_1$. The time interval between two PBs is dictated by the time required to recover sufficient synaptic resources in order to emit a new PB, i.e. by the time scale τ_d controlling the depression [23, 24] (as shown in panel (B6)). Thus, we have memory maintenance through synaptic facilitation, which is refreshed by PBs. As one can appreciate from the spectrogram of population one reported in panel (B8), whenever a PB is delivered, a transient oscillation in the δ band, at a frequency $\simeq 3$ Hz, is observed. Similar transient oscillations have been observed when items are loaded in the PFC of monkeys during WM tasks [19]. Signatures of these oscillations are present also in the spectrogram of the other excitatory population (panel (B9)) but, in this case, they are due to a modulation of the subthreshold membrane potentials and not to PBs. It is interesting to note that δ-oscillations are only sustained by the activity of the excitatory populations, since in the inhibitory spectrogram (panel (B10)) there is no trace of them.

The sequence of PBs can be terminated by reducing the excitation I_B: this operation is performed at time $t = 2.15$ s after the initial stimulation and signalled by the red arrow in panels (B1) and (B3). Also in this experiment the memory load and retrieval is selective, since the second population activity is almost unaffected by these operations as it turns out from the raster plot reported in panel (B3).

As a third experiment, we consider a situation where the memory item is maintained via the persistent activity of the target population, as shown in the right column (C) of Fig. 3. This is possible when the non-specific background input I_B is increased further such that we can observe multistability among three asynchronous states: one corresponding to the spontaneous activity at low firing rate of both excitatory populations and the other two corresponding to population one (two) in a persistent state with increased firing activity and population two (one) at a spontaneous level of activity. More details can be found in Bifurcation analysis. When at time $t = 0$ the memory item is loaded in population one, the population responds by emitting a series of PBs at a frequency $\simeq 27.2$ Hz in the β-γ band (see panels (C8-C10)). As shown in panels (C1) and (C5), once the loading is terminated the blue population enters into a regime of persistent firing characterized by an almost constant firing rate $r_1 \simeq 8.6$ Hz joined to constant values of the synaptic variables. In particular, the available resources of population one remain at a constant and low value due to the continuous firing of the neurons ($\tilde{x}_1 = 0.13$) (see panel C6). However, the lack of resources is compensated by the quite high utilization factor $\tilde{u}_1 = 0.98$ (reported in panel (C7)). Indeed in this case the memory is continuously refreshed by the persistent spiking activity. At time $t = 2.15$ s, indicated by red arrows in panels (C1) and (C3), the value of I_B is reduced to $I_B = 1.2$ and the persistent activity is interrupted. This would correspond to a memory clearance if the WM would be based on the spiking activity only. Instead in this case, it represents only the interruption of the persistent activity, the memory clearance occurring when the facilitation returns to its original value after few more seconds (the decay of u_1 is evident in panel (C7)). The second population, not involved in the memory loading, remains always in a low firing rate regime (panels (C3) and (C5)).

The same numerical experiments have been done by simulating a network with three populations for a total of 600,000 neurons and the same topological architecture displayed in Fig. 2. As expected, the neural mass dynamics reproduces almost perfectly the network dynamics with m-STP for all the three considered experiments. In particular this can be appreciated through the agreement among shaded lines, corresponding to network simulations, and solid ones, referring to the neural mass evolution, reported in Fig. 3 (A5-C5),(A6-C6), (A7-C7).

Competition among two memory items

In this sub-section we verify the robustness of the investigated set-up when two memory items are loaded, one for each excitatory population. In particular, we will examine the possible outcomes of the competition among two loaded items with emphasis on the mechanisms leading to memory juggling [47].

First we analyse the two-item memory juggling in presence of a periodic unspecific stimulation to the excitatory populations. As shown in the left column (A) of Fig. 4, at time $t = 0$ we load the first item in population one by stimulating the population for $\Delta T_1 = 350$ ms, with an excitatory step of amplitude $\Delta I^{(1)} = 0.2$ (see panel (A2)). Population one encodes the item via the facilitation of its efferent synapses. This is a consequence of a series of PBs emitted via the PING mechanism at a frequency $\simeq 21.6$ Hz in the β-range, as evident from the spectrograms in panels (A8-A10). Subsequently a periodic sequence of unspecific stimulations of small amplitude and brief duration is delivered to both populations (namely, step currents of amplitude 0.1 and duration 150 ms applied at intervals of 400 ms). These are sufficient to refresh the memory associated to population one, that reacts each time, by emitting a brief PB able to restore $\tilde{u}_1$ to a high value (panel (A7)). On the contrary population two remains in the low activity regime (panels (A3) and (A5)).

The second item is loaded at time $t = 2.65$ s, by presenting to population two a signal equal to the one presented at time $t = 0$ (panel (A4)). Also the loading of this item is associated to PBs emitted in the β range and involving the inhibitory population. During the presentation of the second item, the previous item is temporarily suppressed. However, when the unspecific stimulations are presented again to both population, PBs are triggered in both populations, resulting to be in anti-phase, i.e. PBs alternate between the two populations at each read-out pulse. The period related to each item is therefore twice the interval between read-out signals (see panels (A1) and (A3)). This is due to the fact that a PB in one excitatory population stimulates the action of the inhibitory neurons in population zero which, as a consequence, suppresses the activity of the other excitatory population, leading to the observed juggling between the two items in working memory. As clear from the spectrograms in panels (A8-A10), the unspecific stimulations of the excitatory populations induce localized peaks in their spectrograms around 2-3 Hz, not involving the inhibitory population. The latter instead oscillates at a frequency $\simeq 21.6$ Hz, thus inducing a modulation of the neural activity of the two excitatory populations. This explain the presence of the series of peaks in the three spectrograms around 20-25 Hz.

Let us now consider a higher background current (namely, $I_{\mathrm{B}} = 1.532$), where self-sustained periodic PBs can be emitted by each excitatory population. In this case it is not necessary to deliver unspecific periodic stimuli to refresh the synaptic memory. The loading of the two items is done by applying the same currents as before ($\Delta I^{(1)} = \Delta I^{(2)} = 0.2$, $\Delta T_1 = \Delta T_2 = 350$ ms) to the two populations, as shown in Fig. 4 (B2) and (B4). The memory loading occurs, once more, via series of bursts emitted with frequency in the β range, through a PING mechanism. Once the item is loaded, a series of subsequent PBs is emitted regularly, at a frequency $\simeq 2.90$ Hz (see Fig. 4 (B1), (B5) and (B8-B10)).

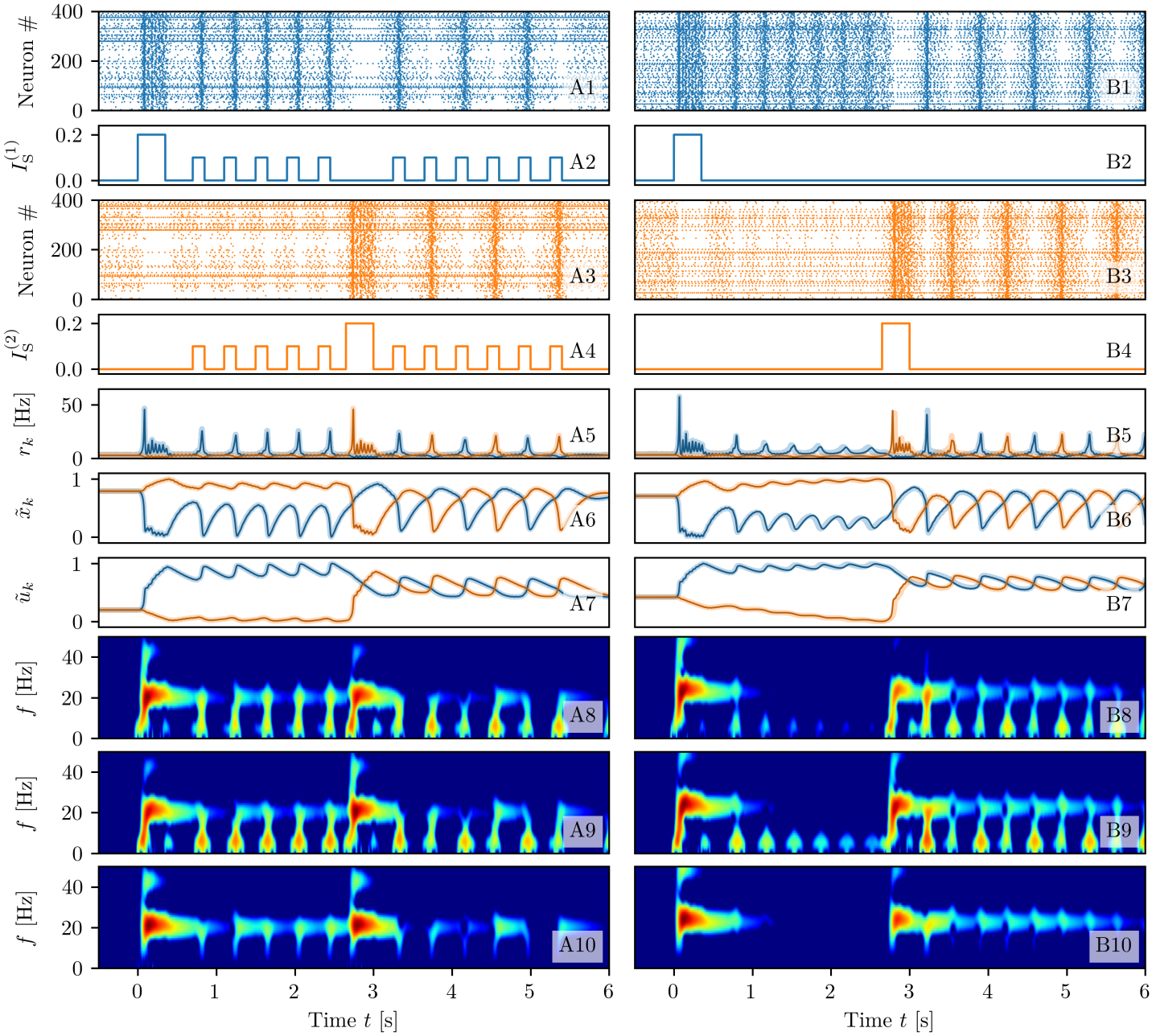


Fig 4. Juggling among two memory items The memory juggling is obtained in two experiments with different background currents: in presence of a periodic unspecific stimulation ($I_B = 1.2$, A) and in the case of spontaneous WM reactivation ($I_B = 1.532$, B). Raster plots of the network activity for the first (blue, **A1-B1**) and second (orange, **A3-B3**) excitatory population. Profiles of the stimulation current $I_S^{(k)}(t)$ for the first (**A2-B2**) (second (**A4-B4**))) excitatory population. Population firing rates r_k(t) (**A5-C5**), normalized available resources $\tilde{x}_k(t)$ (**A6-C6**) and normalized utilization factors $\tilde{u}_k(t)$ (**A7-C7**) of the excitatory populations calculated from the simulations of the neural mass model (solid line) and of the network (shading). Spectrograms of the mean membrane potentials $v_1(t)$ (**A8-C8**), $v_2(t)$ (**A9-C9**), and $v_0(t)$ (**A10-C10**) obtained from the neural mass model. All the other parameters as in Fig. 3.

The second population remains quiescent until the second item is loaded (panels (B3) and (B4)). During the loading period, the activity of the first population is

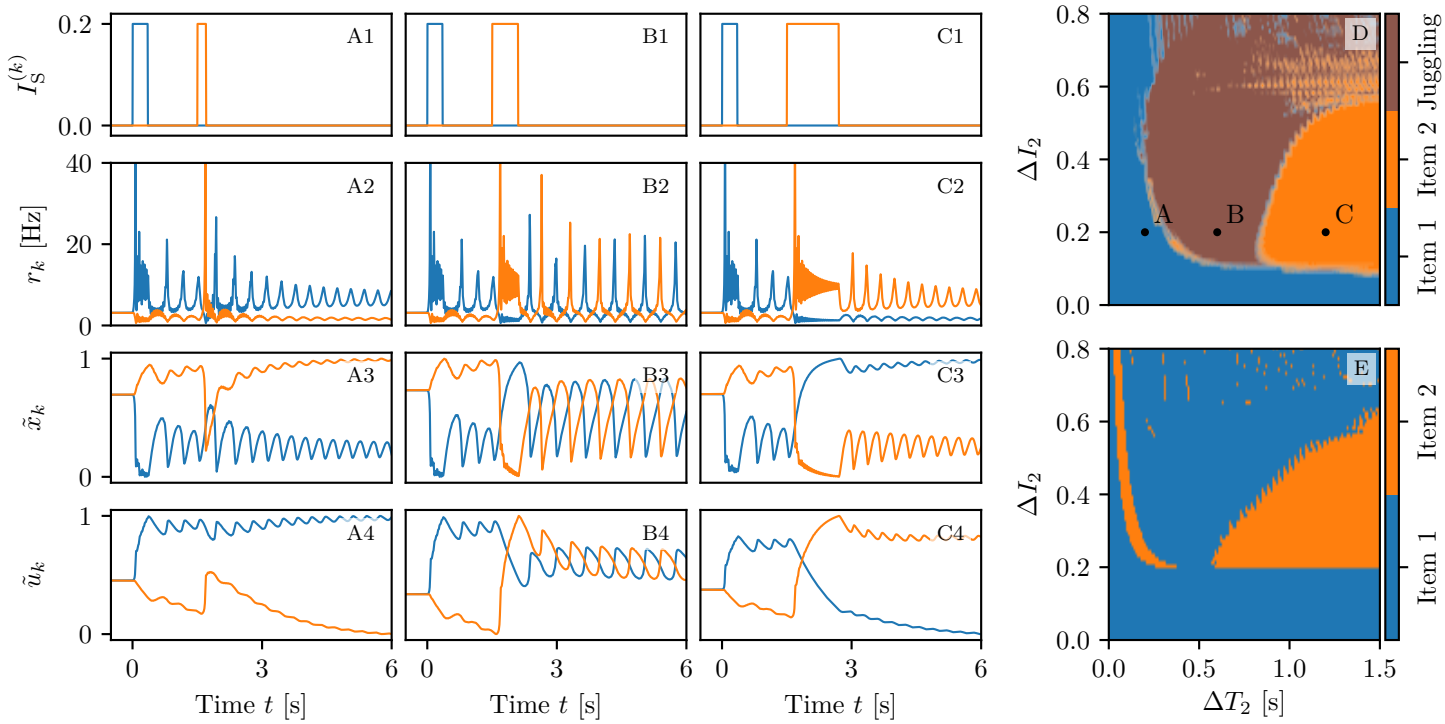


Fig 5. Competition among two memory items. The loading of two memory items is performed in three different ways via two step currents of equal amplitudes $\Delta I^{(1)} = \Delta I^{(2)} = 0.2$, the first one delivered at $t = 0$ and the second one at $t = 1.5$ s. The first step has always a duration of $\Delta T_1 = 0.35$ s; the second one $\Delta T_2 = 0.2$ s (column A), $\Delta T_2 = 0.6$ s (column B) and $\Delta T_2 = 1.2$ s (column C). Stimulation currents $I_S^{(k)}(t)$ (**A1-C1**), instantaneous population firing rates $r_k(t)$ (**A2-C2**), normalized available resources $\tilde{x}_k(t)$ (**A3-C3**) and normalized utilization factors $\tilde{u}_k(t)$ (**A4-C4**) for the first (blue) and second (orange) excitatory population. Final memory states as a function of the amplitude $\Delta I^{(2)}$ and width ΔT_2 of the stimulus to population two for two different background currents: namely, $I_B = 1.532$ (**D**) and $I_B = 2$ (**E**). In (**D**) the three letters A, B, C refer to the states examined in the corresponding columns. The data reported in panel (**D**) has been obtained by delivering the second stimulation at different phases ϕ with respect to the period of the CO of the first population: 20 equidistant phases in the interval $\phi \in [0, 2\pi)$ have been considered. For each of the three possible outcomes an image has been created with a level of transparency corresponding to the fraction of times it has been measured. The three images have been merged resulting in panel (**D**). All the other parameters as in Fig. 3.

temporarily suppressed while, at later times, both items are maintained in the memory by subsequent periodic reactivations of the two populations. In particular it turns out that periodic reactivations are in anti-phase, with frequency $\simeq 1.5$ Hz, i.e., one observes juggling among the two working memory items (see panel (B5)). Also in this case, the inhibitory population is responsible for sustaining the β rhythm, while the excitatory ones for the slow oscillations, as revealed by the spectrograms in panels (B8-B10).

As before, analogous numerical experiments have been performed using the m-STP QIF network with 600,000 neurons and the same topological architecture (see Fig. 2). The agreement among the network simulations (solid lines) and the neural mass results (shaded lines) is impressive for both experiments, as testified by the comparisons reported in Fig. 4 (A5-A7) and (B5-B7).

For the setup analysed in column (B) of Fig. 4, the competition among the two memory items can have different final outcomes, depending on the characteristics of the stimuli, namely their amplitude and duration. For simplicity we fix these parameters for the first stimulation and vary those of the second stimulation. As shown in Fig. 5 (A-C), three outcomes are possible for stimuli with the same amplitude $\Delta I^{(2)} = 0.2$ but

different pulse duration: item one wins, item two wins or they both coexist in the memory (juggling). If ΔT_2 is too short, the facilitation u_2 has no time to become sufficiently large to compete with that of population one (column A). Therefore, once the second stimulus is removed, population one recovers its oscillatory activity and population two returns to the low firing asynchronous activity (see panel A2). For intermediate durations ΔT_2, at the end of the stimulation period, the facilitation u_2 reaches the value u_1 (see panel B4), thus leading the two items to compete. Indeed, as shown in Fig. 5 (B2), the two populations display COs of the same period but in phase opposition, analogously to the case reported in Fig. 4 (B). The stimulation of the second population suppresses the activity of the first, leading to a relaxation of the facilitation to the baseline. Hence, whenever ΔT_2 is sufficiently long, the facilitation of population two prevails on that of population one (see panel (C4)) and, as a final outcome, population two displays COs, while population one has a low activity (panel (C2)).

The results of a detailed analysis for different amplitudes $\Delta I^{(2)}$ and durations ΔT_2 are summarized in Fig. 5 (D). Population one always receives the same stimulus at time $t = 0$ ($\Delta I^{(1)} = 0.2$ and $\Delta T_1 = 0.35$ s), while $\Delta I^{(2)}$ and ΔT_2 of the second stimulus, delivered at time $t = 100$ s, are varied [2]. In particular, the amplitude has been varied in steps of 0.01 within the interval $\Delta I^{(2)} \in [0, 0.8]$, while the duration in steps of 0.01 s for $\Delta T_2 \in [0, 1.5]$ s. The classification of the final states has been performed three seconds after the second stimulation, by estimating the time averaged firing rates $\langle r_1 \rangle$ and $\langle r_2 \rangle$ over an interval of one second. From this we get the indicator $P := \frac{\langle r_1 \rangle}{\langle r_1 \rangle + \langle r_2 \rangle}$. If $P > 0.7$ ($P < 0.3$) item one (item two) has been memorized, otherwise we have juggling between the two items. From Fig. 5 (D), we see that, for sufficiently small amplitudes $\Delta I^{(2)} < 0.11$ or durations $\Delta T_2 < 0.2$ s, the second stimulation is unable to change the memory state, that remains on item one. A frequent outcome of the experiments is the coexistence between the two memory items, that we observe in a large parameter region for intermediate pulse durations, namely $0.2 \text{ s} < \Delta T_2 < 0.8$ s. For sufficiently long perturbations $\Delta T_2 > 0.8$ s and intermediate amplitudes $0.11 < \Delta I^{(2)} < 0.53$, item two is memorized. It is unexpected that, for larger amplitudes, we observe a multistability among all the three possible outcomes. Indeed, small variations of amplitude, duration or delivering time can induce a completely different outcome, thus suggesting that for these parameter values, some transient chaotic behaviour is observable [48, 49].

We have performed the same analysis for the higher current value $I_{\mathrm{B}} = 2$, which supports a persistent state activity of one of the two populations. We set initially population one in the persistent state and we deliver a stimulus in form of a step current to population two. In this case we can obtain only two possible final outcomes: item one (two) loaded corresponding to population one (two) in the persistent state, while the other in the low activity regime. No memory juggling has been observed with persistent states in our model. The results of this investigation are summarized in Fig. 5 (E): for what concerns the final prevalence of item two, the results are similar to the ones obtained in the previous analysis, where the item was encoded in self-sustained periodic COs. This apart for a narrow stripe in the $(\Delta T_2, \Delta I^{(2)})$ - plane where item two is also finally selected. We can understand the origin of this stripe by examining the results reported in Fig. S1, for three stimulations with same amplitude $\Delta I^{(2)} = 0.4$ and different durations. The effect of the stimulation is to increase the activity of population two and indirectly, also the inhibitory action on population one. However, for short $\Delta T_2 < 70$ ms, the activity of population two is not sufficient to render its efferent synapses stronger than those of population one. As a matter of fact, when the stimulations is over, population one returns into the persistent state. For longer $\Delta T_2 \geq 70$ ms, population one is not able to recover its resources before population two

[2] The long interval of 100 seconds between the two stimuli is chosen in order to ensure that the outcome is independent of the characteristic of the first stimulus: namely, $\Delta I^{(1)}$ and ΔT_1.

elicits a PB. Population two takes over and the dynamics relax back, maintaining item two in WM via the persistent activity (as shown in column (A) of Fig. S1). For $\Delta T_2 \geq 130$ ms we observe the emission of two or more successive PBs in population two during the stimulation period. These PBs depress noticeably x_2, which is unable to recover before population one, characterized by a larger amount of available resources x_1, emits a PB and takes over to item one (see column (B) of Fig. S1). However, for much longer $\Delta T_2 \geq 850$ ms, the absence of activity in population one leads to a noticeable decrease of the utilization factor u_1. As a matter of fact at the end of the stimulation period, the item two is finally selected (as shown in column (C) of Fig. S1).

To summarize, item two can be selected when the duration of the stimulation is restricted to a narrow time interval, sufficiently long to render the efferent synapses of population two stronger than those of population one, but not so long to highly deplete the resources of the same synapses. Furthermore, item two can be selected when ΔT_2 is sufficiently long to allow for the decay of the synaptic facilitation of population one towards its baseline value.

Multi-Item Memory Loading

In order to be able to load a higher number of memory items in the WM, we consider the neural mass model (Eqs. (1)) with m-STP (Eqs. (2)) arranged in a more complex architecture composed by $N_{\rm pop} = 7$ excitatory and one inhibitory populations, each excitatory population coding for one memory item. The system is initialized with all the populations in the silent state. Then, each item is loaded by delivering an excitatory pulse of amplitude $\Delta I^{(k)} = 1$ and duration of 0.2 s to the chosen population, while successive items are loaded at intervals of 1.25 s, as shown in Fig. 6 (C).

Let us first consider the successive loading of three items shown in Fig. 6. As one can appreciate from the spectrogram reported in panel (D), during the loading phase, each stimulated excitatory population emits a sequence of PBs in the $\beta - \gamma$ range ($\simeq 27$ Hz) via a PING mechanism mediated by the inhibitory population. This is joined to stimulus locked transient oscillations in the δ-band around 2 Hz. Furthermore, during each loading period, the activity of the other populations is interrupted and it recovers when the stimulation ends. These results resemble the LFP measurement performed in PFC of primates during coding of objects in short term-memory [11, 19]. In particular, the experimentally measured power spectrum of the LFP displays transient oscillations at $\simeq 2 - 4$ Hz in the δ range, phase-locked to stimulus presentation, together with tonic oscillations at $\simeq 32$ Hz.

As shown in Fig. 6 (A), the loading of each item is followed by the emission of PBs from the stimulated population, at regular time intervals $T_{\rm c}$. The PBs of all the excited populations arrange in a sort of *splay state* with inter-burst periods $T_{\rm b} = T_{\rm c}/N_{\rm I}$, where $N_{\rm I}$ is the number of loaded items [50]. The period $T_{\rm c}$ depends on the number of loaded items: for $N_{\rm I} = 3$ we have $T_{\rm c} \simeq 0.2035$ s. The non stimulated populations display a low firing activity modulated by the slow PB emission in the excitatory populations and by the fast $\beta - \gamma$ oscillations driven by the inhibitory activity (see Fig. 6 (B)).

A more detailed analysis of the spectrogram in Fig. 6 (D) reveals that, after each loading phase, several harmonics of the fundamental frequency $f_0 \equiv 1/T_{\rm c} \simeq 5$ Hz are excited. In particular after the complete loading of the three items the most enhanced harmonics are those corresponding to $3f_0 \simeq 15$ Hz and $6f_0 \simeq 30$ Hz. This is due to the coincidence of the frequency $3f_0 = f_{\rm b} \equiv 1/T_{\rm b}$ with the inter-burst frequency $f_{\rm b}$, while the main peak in the spectrogram, located at 30 Hz, is particularly enhanced due to the resonance with the proximal $\beta - \gamma$ rhythm associated to the PING oscillations.

It is interesting to notice that the transient δ oscillations related to the stimulation phase are only supported by the excitatory population dynamics, as evident by comparing the spectrograms for $v_1(t)$ and for the membrane potential averaged over all

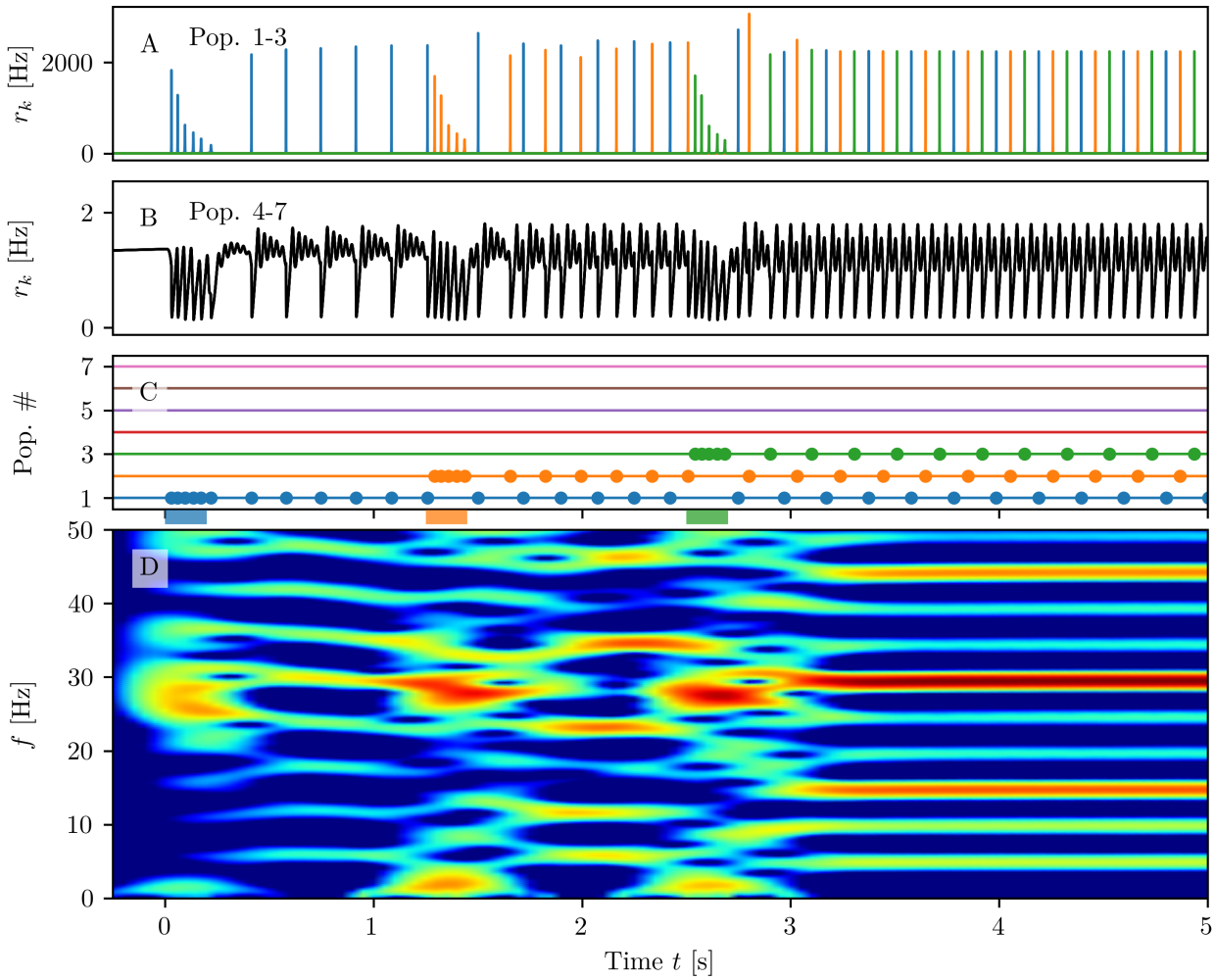


Fig 6. Multi-item memory loading Firing rates $r_k(t)$ of excitatory populations: blue, orange and green curves corresponds to $k = 1, 2, 3$ (**A**), while the black curve refers to $k = 4, \ldots, 7$ (**B**). PBs of excitatory populations are shown in (**C**): horizontal lines in absence of dots indicate low activity regimes; coloured dots mark the PBs' occurrences. The coloured bars on the time axis denote the presence of a stimulation pulse targeting the corresponding population. Only populations $k = 1, \ldots, 3$ are stimulated in the present example. The spectrogram of the average membrane potential of population one $v_1(t)$ is reported in panel (**D**). Parameters: $N_{\text{pop}} = 7$, $\tau_{\text{m}}^{\text{e}} = 15$ ms, $\tau_{\text{m}}^{\text{i}} = 10$ ms, $J_{\text{ee}}^{(\text{s})} = 154$, $J_{\text{ee}}^{(\text{c})} = \frac{4}{7} \cdot 18.5$, $J_{\text{ei}} = -26$, $J_{\text{ie}} = \frac{4}{7} \cdot 97$, $J_{\text{ii}} = -60$, $I_{\text{B}} = 0$, $H^{(\text{e})} = 0.05$, $H^{(\text{i})} = -2$, $\Delta^{(\text{e})} = \Delta^{(\text{i})} = 0.1$

excitatory populations (shown in panels (B) and (D) of Fig. 7, Fig. S2, and Fig. S3) with those for the inhibitory population (shown in panel (C) of Fig. 7, Fig. S2, and Fig. S3). Another important aspect is that only the harmonics of $f_{\text{b}} \equiv N_{\text{I}} f_0$ are present in the spectrogram of the inhibitory population, and in that associated to the average excitatory membrane potential, but not all the other harmonics of f_0. This is related to the fact that, in the average of the mean excitatory membrane potentials, the bursts of all populations will be present with a period T_{b}, therefore in this time signal there is no more trace of the periodic activity of the single population. This is due to the fact that $T_{\text{c}} \equiv N_{\text{I}} T_{\text{b}}$. Furthermore, in the dynamics of the inhibitory population, the activity of all excitatory populations are reflected with the same weight, therefore also in this case there is no sign of the fundamental frequency f_0.

If we continue to load more items, we discover that up to five items are loaded and retained by the WM (see Fig. S2 (A)). However, if we load a sixth item, the WM is able to maintain all the loaded $N_{\mathrm{I}} = 6$ items only for a short time interval $\simeq 1$ s, after which population five stops delivering further PBs. The corresponding item is not in WM anymore and the memory load returns to $N_{\mathrm{I}} = 5$ (as shown in Fig. 7 (A)). By loading seven items we can induce more complex instabilities in the WM, indeed the experiment reported in Fig. S3 reveals that, in the end, only $N_{\mathrm{I}} = 4$ items can be maintained, suggesting that a too fast acquisition of new memory items can compromise also already stored WM items. As a matter of fact, only the oldest and newest items are retained by the WM: this in an example of the so-called *primacy and recency effect*, which has been reported in many contexts when a list of items should be memorized [51, 52].

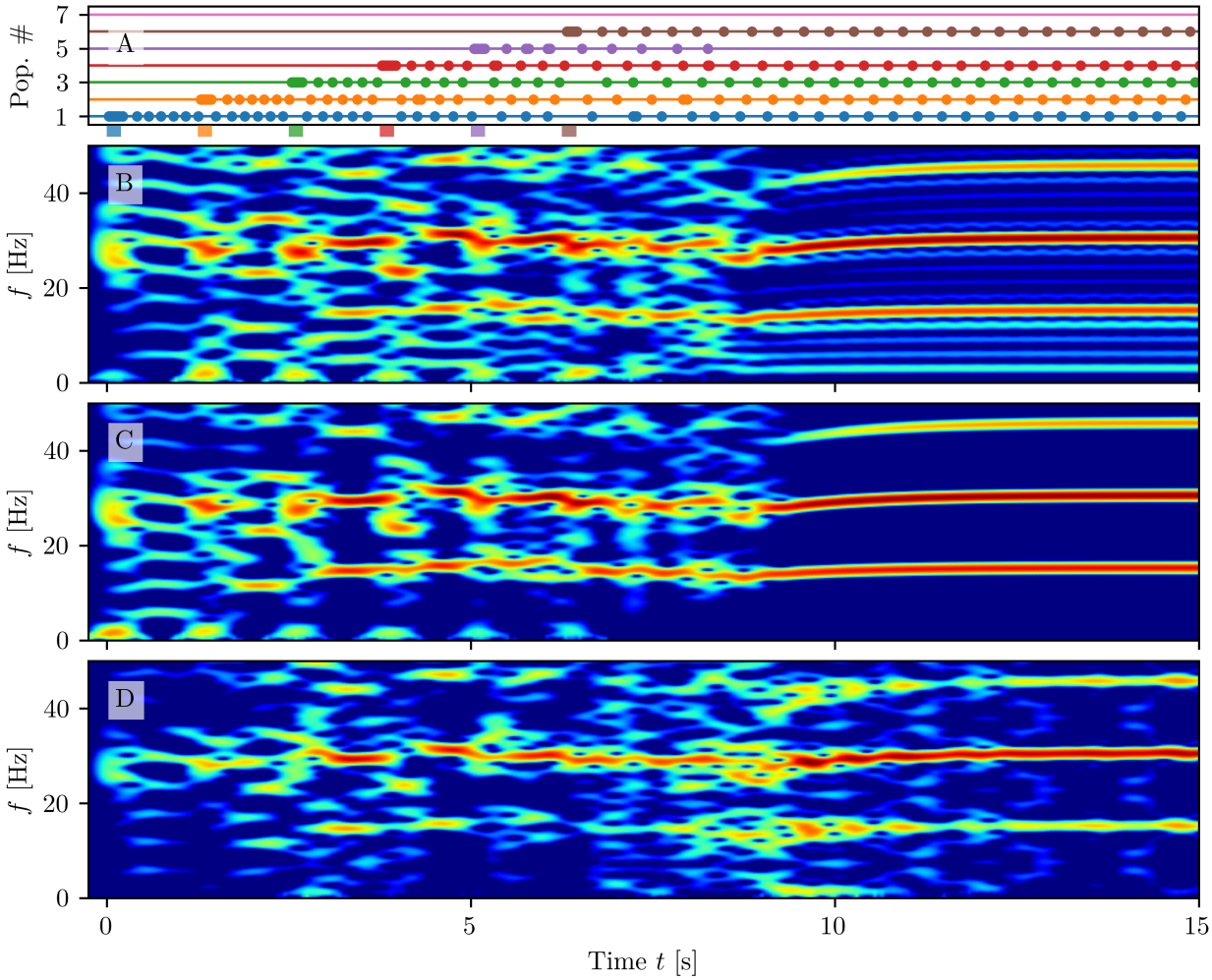


Fig 7. Maximal capacity Population spikes of excitatory populations: horizontal lines in absence of dots indicate quiescence phases at low firing rates r_k for populations $k = 1, ...7$. Dots mark population spikes of the corresponding population. The coloured bars on the time axis mark the starting and ending time of stimulating pulses, targeting each population. Spectrograms of the mean membrane potential $v_1(t)$ (**B**), $v_0(t)$ (**C**) and of the mean membrane potentials averaged over all the excitatory populations (**D**). Parameters as in Fig. 6.

Another feature usually measured during the loading or juggling of more items is the power associated to different frequency bands. In particular, a recent experiment has

examined the frequency spectra of LFPs measured from cortical areas of monkeys while they maintained multiple visual stimuli in WM [53]. These results have revealed that higher-frequency power (50–100 Hz) increased with the number of loaded stimuli, while lower-frequency power (8–50 Hz) decreased. Furthermore, the analysis of a detailed network model, biophysically inspired by the PFC structure, has shown that θ and γ power increased with memory load, while the power in the α-β band decreased, in accordance with experimental findings [16].

We considered the loading of $N_{\rm I}$ items and estimated the power spectrum, after the loading of all the considered items, for the following variables: i) the mean membrane potential $v_1(t)$ of the excitatory population one; ii) the mean membrane potential v_0 of the inhibitory population; iii) the mean membrane potential averaged over all the excitatory populations. The power has been estimated in the α-β (8-25 Hz) and in the γ-bands (25-100 Hz) by integrating the spectra in such frequency intervals. The results for $1 \leq N_{\rm I} \leq 5$ shown in Fig. 8 reveal that the power in the γ-band is essentially increasing with $N_{\rm I}$ for all the considered signals, as reported experimentally for several brain areas involved in WM [54–56]. Moreover, the power in the α-β band increases monotonically with $N_{\rm I}$ for v_1, while it displays a non monotonic behaviour for the inhibitory population and for the average excitatory membrane potential. This difference can be explained by the fact that the fundamental frequency f_0 is present in the spectrum of v_1, but not in those of v_0 and of the average excitatory membrane potential.

A theoretical estimation of the maximal capacity $N_{\rm c}^{\rm max}$ for WM, based on short-term synaptic plasticity, has been recently derived in [26]. For our neural mass model we have obtained an analogous expression by following [26], namely

$$N_{\rm c}^{\rm max} \simeq \frac{\tau_{\rm d}}{\tau_{\rm m}^{\rm e}} \ln \left[\frac{\tau_{\rm f}/\tau_{\rm d}}{1-U_0} \right] \frac{\sqrt{C}}{\pi} \quad ; \tag{4}$$

where the expression of C is explicitly reported in Maximal Working Memory Capacity. As shown in [26], the value of $N_{\rm c}^{\rm max}$ is essentially dictated by the recovery time of the synaptic resources $\tau_{\rm d}$ and has a weaker (logarithmic) dependence on the ratio between the facilitation and depression time scales. Furthermore, for our model, the maximal capacity increases by increasing $H^{(\rm e)}$, $I_{\rm B}$ and the self and cross excitatory synaptic couplings, while it decreases whenever the coupling from inhibitory to excitatory population is augmented, see Eq. (22).

By employing the theoretical estimation (4) we get $N_{\rm c}^{\rm max} \in [3.6, 4.8]$, in pretty good agreement with the measured maximal capacity. We can affirm this in view of the results reported in [26] for a different mean field model, where the analytical predictions overestimate the maximal capacity by a factor two.

Discussion

In this paper we have reported a new generation neural mass model for WM based on short-term depression and facilitation. The model reproduces exactly the dynamics of heterogeneous QIF networks with STP. The novelty of this neural mass model, besides being not heuristic, but derived in an exact manner from the microscopic underlying dynamics, is that it reproduces the macroscopic evolution of the population firing rate as well as of the mean membrane potential. This allows to get insight not only on the synchronized spiking activity, but also on the sub-threshold dynamics and to extract information correlated to the LFPs and to electroencephalographic signals.

The model is extremely flexible since, depending on the operational point, it allows to mimic WM maintenance both in terms of persistent spiking and of plastic footprints

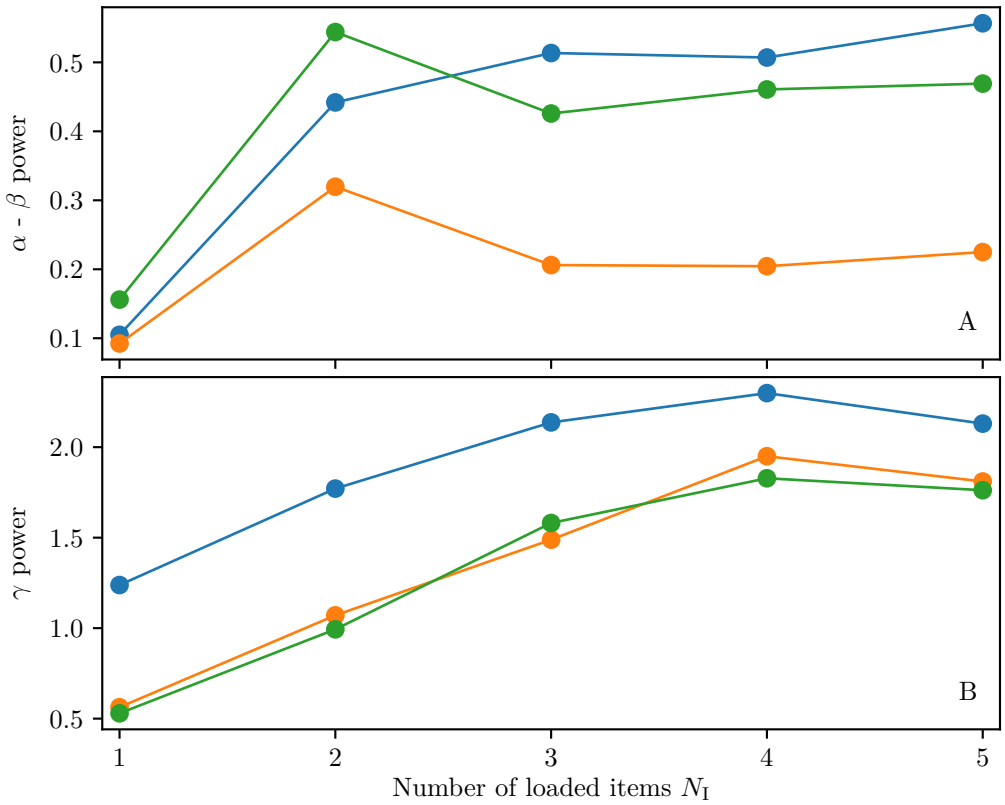


Fig 8. Dependence of the power on the number of loaded items Power in the $\alpha - \beta$-band (8-25 Hz) (**A**) and in the γ-band (25-100 Hz) (**B**) as a function of the number N_I of loaded items. The integral of the power spectra in the specified frequency bands are displayed for the mean membrane potential $v_1(t)$ (blue symbols), $v_0(t)$ (orange symbols) and for the mean membrane potentials averaged over all the excitatory populations (green symbols). The power spectra have been evaluated over a 10 s time window, after the loading of N_I items, when these items were juggling in the WM. Parameters as in Fig. 6.

into the synapses. In particular, short-term depression and facilitation complement each other to obtain an efficient synaptic-based WM model. Facilitation allows to maintain trace of the loaded items stored in the WM, while depression is responsible for the bursting activity that refreshes the WM content.

Memory loading is characterized by collective oscillations (PBs) with frequencies $\simeq 22 - 28$ Hz in the β-γ band. These oscillations emerge naturally in the considered model thanks to a PING mechanism involving the selectively stimulated excitatory population and the common inhibitory pool. The memory loading is also phase-locked to a transient oscillation of frequency $\simeq 2 - 3$ Hz. These results resemble the population activity of the PFC measured during a behavioural task of object recognition performed by monkeys [11, 19]. In particular, in these experiments the analysis of the LFP spectrograms revealed an evoked response around 2-4 Hz and tonic oscillations around 32 Hz.

Memory maintenance in the synaptic-based model is ensured by facilitation over a time scale of few seconds, even in absence of spiking activity. The memory can be refreshed, and therefore maintained for a longer period, thanks to the reactivation of the synaptic resources of the excitatory population storing the item. This can be obtained,

on one side, via brief non specific stimulations given to all the excitatory populations, analogously to the reactivation of latent working memories performed with single-pulse transcranial magnetic stimulation realised in humans [17]. In particular, to maintain the memory for long time intervals, the stimulations should be delivered with a period smaller than the decay time of the facilitation. On the other side the memory maintenance over long periods can be achieved thanks to the spontaneous emergence of periodic PBs delivered by the excitatory population coding for the loaded item.

An interesting aspect that we have investigated, concerns the competition among two items loaded in non-overlapping populations. When the WM load consists already of one item coded by an excitatory population, the stimulation of the other population can be regarded as a distractor. If the items are stored as repetitive PBs, we can observe three outcomes: for brief stimulations the distractor has no influence on the WM, and the first item remains loaded; for sufficiently long stimulations the second item is loaded in the WM; for intermediate situations both items are maintained in the WM. In this latter case both populations deliver periodic trains of PBs arranged in anti-phase. Such behaviour can be seen as a neural correlate to two items juggling in WM [47]. For sufficiently long and strong stimulations one observe a chaotic scenario [49], where the final outcome can be any of the three described above and it depends on small differences in the perturbation deliverance. Our findings can help in elucidating the results reported in [47], where it has been shown that, when an attended and an unattended item are juggling for a long period, the unattended item can prevail leading to a loss of the stored memory.

If the items are stored in the WM as persistent states, the memory juggling is no more observable. As in the previous case, for sufficiently long stimulations, the second item substitutes the first one in the WM. However, due to synaptic depression and facilitation, even a short stimulation can lead to the loading of the second item in the WM, whenever its duration falls in a narrow time interval ($\simeq \mathcal{O}(\tau_{\rm d}/2)$) and the amplitude of the stimulation is sufficiently large. These suggests that there are optimal stimulation strategies to ensure a fast learning of a new item in the WM.

As our results demonstrate, more items can be maintained at the same time in the WM as periodic trains of PBs. All the populations coding for the stored items follow the same periodic dynamics, but they deliver PBs at evenly shifted phases, similarly to the splay states observed for globally coupled excitatory neuronal networks [50]. The inter-burst interval of two successive PBs approaches a value $T_{\rm b} \simeq 65$ ms when more than 2 items are loaded. This clearly induces the emergence of a peak in the power spectra of the mean membrane potentials in the β band for $f_{\rm b} \simeq 15$ Hz. However, the most prominent peak in the spectra is around 30 Hz, due to the resonance of the second harmonic of $f_{\rm b}$ with the oscillations of the inhibitory population in the $\beta - \gamma$ range, which are associated to memory loading and sustained by a PING mechanism involving, besides the inhibitory pool, the excitatory population active in that moment. This mechanism allows for the maintenance of multi-items in the WM without destructive or interference effects often reported for models based on persistent spiking [18]. Furthermore, the memory of different items is associated to preferential phases with respect to the collective limit cycle behaviour of the whole system, characterized by a period $T_{\rm c} = N_{\rm I} T_{\rm b}$, where $N_{\rm I}$ is the number of loaded item. Experimental evidences of phase-dependent neural coding of objects in the PFC of monkeys have been reported in [19].

Furthermore, we observed that the power in the γ-band (25-100 Hz) increases with the number $N_{\rm I}$ of loaded items, in agreement with several experimental studies related to the WM [54–56]. Instead the power in the $\alpha - \beta$ band reveals a non monotonic behaviour with $N_{\rm I}$, characterized by a rapid increase passing from 1 to 2 items, a rapid drop from 2 to 3 items and remaining essentially constant for $N_{\rm I} \geq 3$. This at variance

with other models revealing a decrease of power in the $\alpha - \beta$ band with $N_{\rm I}$ [16].

As a final aspect, we have considered the maximal capacity of our model and, by following [26], we have been able to derive an analytic expression for the maximal capacity for our neural mass model with STP. As shown in [26], the maximal capacity is essentially controlled by the ratio between the recovery time of the available synaptic resources $\tau_{\rm d}$ and the membrane time constant of the excitatory populations. Furthermore, in our case, the maximal capacity grows with the excitatory drive and decreases with the inhibitory one.

A fundamental aspect of WM, not included in our model, is the volitional control, which is a cognitive function that allows for the control of behaviour from the environment and to turn it towards our internal goals [2]. A flexible frequency control of cortical oscillations has been recently proposed as an unified mechanism for the rapid and controlled transitions of the WM between different computational operations [57–59]. In particular, the authors consider as a neural correlate of WM a bistable network with coexisting persistent and resting states that an external periodic modulation can drive from an operating mode to another one depending on the frequency of the forcing. However, in [59] it has been suggested that the forcing term, in order to be considered more realistic, should self-emerge by the network dynamics in form of trains of periodic PBs and not be imposed from the exterior. In our model for synaptic-based WM we have shown that some WM operations are associated with PBs delivered at different frequencies: namely, item loading and recall with transient oscillations in the δ band, as in [59], joined to burst oscillations in $\beta - \gamma$ band as in [57]; multi-item maintenance to harmonics in the $\beta - \gamma$ band. Therefore, we believe that our neural mass model with STP can represent a first building block for the development of an unified control mechanism for WM, relying on the frequencies of deliverance of the self-emerging trains of PBs. However, a development towards realistic neural architectures would require to design a multi-layer network topology to reproduce the interactions among superficial and deep cortical layers [18].

Methods

Spiking neuronal network model

The membrane potential dynamics of the i-th QIF neuron in a network of size N can be written as

$$\tau_{\rm m}\dot{V}_i = V_i^2(t) + \eta_i + I_{\rm B} + I_{\rm S}(t) + \tau_{\rm m}\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_{ij}(t)S_j(t) \quad , \qquad i = 1, \ldots, N \tag{5}$$

where $\tau_{\rm m} = 15$ ms is the membrane time constant and $\tilde{J}_{ij}(t)$ the strength of the direct synapse from neuron j to i that, in absence of plasticity, we assume to be constant and all identical, i.e. $\tilde{J}_{ij}(t) = J$. The sign of J determines if the neuron is excitatory ($J > 0$) or inhibitory ($J < 0$). Moreover, η_i represents the neuronal excitability, $I_{\rm B}$ a constant background DC current, $I_{\rm S}(t)$ an external stimulus and the last term on the right hand side the synaptic current due to the recurrent connections with the pre-synaptic neurons. For instantaneous post-synaptic potentials (corresponding to δ-spikes) the neural activity $S_j(t)$ of neuron j reads as

$$S_j(t) = \sum_{t_j(k)<t} \delta(t - t_j(k)), \tag{6}$$

where $S_j(t)$ is the spike train produced by the j-th neuron and $t_j(k)$ denotes the k-th spike time in such sequence. We have considered a fully coupled network without

autapses, therefore the post-synaptic current will be the same for each neuron apart corrections of order $\mathcal{O}(1/N)$.

In the absence of synaptic input, external stimuli and $I_{\rm B} = 0$, the QIF neuron exhibits two possible dynamics, depending on the sign of η_i. For negative η_i, the neuron is excitable and for any initial condition $V_i(0) < \sqrt{-\eta_i}$, it reaches asymptotically the resting value $-\sqrt{-\eta_i}$. On the other hand, for initial values larger than the excitability threshold, $V_i(0) > \sqrt{-\eta_i}$, the membrane potential grows unbounded and a reset mechanism has to be introduced to describe the spiking behaviour of a neuron. Whenever $V_i(t)$ reaches a threshold value $V_{\rm p}$, the neuron i delivers a spike and its membrane voltage is reset to $V_{\rm r}$, for the QIF neuron $V_{\rm p} = -V_{\rm r} = \infty$. For positive η_i the neuron is supra-threshold and it delivers a regular train of spikes with frequency $\nu_0 = \sqrt{\eta_i}/\pi$.

As already mentioned, the analytic derivation of the neural mass model can be performed when the excitabilities $\{\eta_i\}$ follows a Lorentzian distribution

$$g(\eta) = \frac{1}{\pi}\frac{\Delta}{(\eta - H)^2 + \Delta^2}. \tag{7}$$

centred at H and with HWHM Δ. In particular, the excitability values η_i have been generated deterministically for a network of N neurons by employing the following expression

$$\eta_i = H + \Delta \tan\left[\frac{2}{\pi}\frac{2i - N - 1}{N + 1}\right] \quad . \tag{8}$$

Numerical simulations of the QIF networks have been performed by employing an Euler scheme with a timestep $\Delta t = 10^{-4}\tau_{\rm m}$. Moreover, when performing the numerical experiments, analogous to the approach in [32], the threshold and reset values have been approximated to $V_p = -V_r = 100$ and a refractory period is introduced to deal with this approximation. In more details, whenever the neuron i reaches $V_i \geq 100$, its voltage is reset to $V_i = -100$ and the voltage remains unchanged at the reset value for a time interval $\frac{2\tau_{\rm m}}{100}$, accounting for the time needed to reach $V_i = \infty$ from $V_i = 100$ and to pass from $V_i = -\infty$ to $V_i = -100$. The spike emission of neuron i is registered at half of the refractory period.

Short-term synaptic plasticity

By following [15] in order to reproduce WM mechanisms in the PFC we assume that excitatory-to-excitatory synapses display depressing and facilitating transmission, described by a phenomenological model of STP developed by Markram, Tsodyks and collaborators [20–22]. In this model, each pre-synaptic spike depletes the neurotransmitter supply and, at the same time, the concentration of intracellular calcium. The accumulation of calcium ions is responsible for an increase of the release probability of neurotransmitters at the next spike emission. Each synapse displays both depression on a time-scale $\tau_{\rm d}$, due to the neurotransmitter depletion, and facilitation on a time-scale $\tau_{\rm f}$, linked to the increased calcium concentration and release probability. Experimental measurements for the facilitating excitatory synapses in the PFC indicates that $\tau_{\rm f} \simeq 1$ s and that $\tau_{\rm f} \gg \tau_{\rm d}$ [60].

As done in a previous analysis [44], we also assume for the sake of simplicity, that all the efferent synapses of a given neuron follow the same dynamical evolution, therefore they can be characterized by the index of the pre-synaptic neuron. In particular, the depression mechanism is mimicked by assuming that, at certain time t, each synapse has at disposal $X_i(t)$ synaptic resources and that each pre-synaptic spike uses a fraction $U_i(t)$ of these resources. In absence of emitted spikes the synapses will recover the

complete availability of their resources on a time scale $\tau_{\rm d}$, i.e. $X_i(t \gg \tau_{\rm d}) \to 1$. The facilitation mechanism instead leads to an increase of the fraction $U_i(t)$ of employed resources at each spike by a quantity $U_0(1-U_i)$. In between spike emissions U_i will recover to its baseline value U_0 on a time scale $\tau_{\rm f}$. The dynamics of the synaptic variables $X_i(t)$ and $U_i(t)$ is ruled by the following ordinary differential equations:

$$\dot{X}_i = \frac{1-X_i(t)}{\tau_{\rm d}} - X_i(t)U_i(t)S_i(t) \tag{9a}$$

$$\dot{U}_i = \frac{U_0-U_i(t)}{\tau_{\rm f}} + U_0(1-U_i(t))S_i(t) \ . \tag{9b}$$

In this framework U_iX_i represents the amount of resources employed to produce a post-synaptic potential, therefore the synaptic coupling $\tilde{J}_{ij}(t)$ entering in Eq. (5) in presence of STP will be modified as follows:

$$\tilde{J}_{ij}(t) = JU_j(t)X_j(t) \quad \forall i \quad . \tag{10}$$

The spiking network model with microscopic STP (μ-STP) (Eqs. (5), (9) and (10)) consists of $3N$ differential equations: one for the membrane voltage and two for the synaptic variables of each neuron. In order to reduce the system size effects and to obtain accurate simulations reproducing the neural mass dynamics, extremely large network sizes are required, namely $N > 100,000$.

For these network simulations massive numerical resources are required, however the complexity of the synaptic dynamics can be noticeably reduced, by treating the STP at a mesoscopic level under the assumption that the spike trains emitted by each neuron are Poissonian [21,45]. In this framework, the mesoscopic description of the synapse can be written as

$$\dot{x} = \frac{1-x(t)}{\tau_{\rm d}} - u(t)x(t)A(t) \tag{11a}$$

$$\dot{u} = \frac{U_0-u(t)}{\tau_{\rm f}} + U_0(1-u(t))A(t) \quad , \tag{11b}$$

where $x = \langle X_i \rangle$ and $u = \langle U_i \rangle$ represent the average of the microscopic variables $\{X_i\}$ and $\{U_i\}$ over the whole population[3] and $A(t) = \langle S_j(t) \rangle$ is the mean neural activity.

The synaptic coupling entering in Eq. (5) now becomes:

$$\tilde{J}_{ij}(t) = Ju(t)x(t) \quad \forall i,j \quad . \tag{12}$$

The dynamics of the QIF network with the mesoscopic STP (m-STP) is given by $N+2$ ODEs, namely Eqs. (5) and (11) with $\tilde{J}_{ij}(t)$ given by (12), therefore we can reach larger system sizes than with the μ-STP network model.

It should be noticed that the synaptic variables $X_i(t)$ and $U_i(t)$ for the same synapse are correlated, since they are both driven by the same spike train $S_i(t)$ delivered by neuron i. These correlations are neglected in the derivation of the m-STP model, therefore one can write $\langle U_i(t)X_i(t) \rangle = \langle U_i(t) \rangle \langle X_i(t) \rangle = u(t)x(t)$. This approximation is justified in [21] by the fact that the coefficients of variation of the two variables U_i and X_i are particularly small for facilitating synapses. Indeed it is possible to write a stochastic mesoscopic model for the STP that includes the second order moments for U_i and X_i, i.e. their correlations and the fluctuations due to the network's finite size [45]. The finite size effects are introduced in the m-STP deterministic model (11a) via the fluctuation of the mean population activity $A(t)$, which is generated by the firing of N neurons.

[3] The population average $\langle \cdot \rangle = \frac{1}{N_l} \sum_{i=1}^{N_l} \cdot$ is performed over all the neurons N_l in population l

In our simulations we fixed $\tau_{\rm d} = 200$ ms, $\tau_{\rm f} = 1500$ ms and $U = 0.2$ and we integrated the network equations by employing a standard Euler scheme with time step 0.0015 ms.

Exact neural mass model with STP

For the heterogeneous QIF network with instantaneous synapses (Eqs. (5)-(6)), an exact neural mass model, in absence of STP, has been derived in [32]. The analytic derivation is possible for QIF spiking networks thanks to the Ott-Antonsen Ansatz [29] applicable to phase-oscillators networks whenever the natural frequencies (here corresponding to the excitabilities $\{\eta_i\}$) are distributed accordingly to a Lorentzian distribution with median H and HWHM Δ. In particular, this neural mass model allows for an exact macroscopic description of the population dynamics, in the thermodynamic limit $N \to \infty$, in terms of only two collective variables, namely the mean membrane voltage $v(t)$ and the instantaneous population rate $r(t)$, as follows

$$\tau_{\rm m}\dot{r}(t) = \frac{\Delta}{\tau_{\rm m}\pi} + 2r(t)v(t) \tag{13a}$$

$$\tau_{\rm m}\dot{v}(t) = v^2(t) + H + I_{\rm B} + I_{\rm S}(t) - [\pi\tau_{\rm m}r(t)]^2 + \tau_{\rm m}\tilde{J}(t)r(t) \quad ; \tag{13b}$$

where the synaptic strength is assumed to be identical for all neurons and for instantaneous synapses in absence of plasticity $\tilde{J}(t) = J$. However, by including a dynamical evolution for the synapses and therefore additional collective variables, this neural mass model can be extended to any generic post-synaptic potentials, see e.g. [33] for exponential synapses or [35] for conductance based synapses with α-function profile.

In our case the synaptic evolution at a mesoscopic level is given by the m-STP model described above, therefore it will be sufficient to include the dynamical evolution of the m-STP in the model (13) to obtain an exact neural mass model for the QIF spiking network with STP. In more details, we consider the equations (11),(12) for the m-STP dynamics and we substitute the population activity $A(t)$ with the instantaneous firing rate $r(t)$, that corresponds to its coarse grained estimation in the limit $N \to \infty$. Furthermore, the synaptic coupling entering in Eq. (13) will become $\tilde{J}(t) = Ju(t)x(t)$ and the complete neural mass model reads as

$$\tau_{\rm m}\dot{r}(t) = \frac{\Delta}{\tau_{\rm m}\pi} + 2r(t)v(t) \tag{14a}$$

$$\tau_{\rm m}\dot{v}(t) = v^2(t) + H + I_{\rm B} + I_{\rm S}(t) - [\pi\tau_{\rm m}r(t)]^2 + J\tau_{\rm m}u(t)x(t)r(t) \tag{14b}$$

$$\dot{x}(t) = \frac{1 - x(t)}{\tau_{\rm d}} - u(t)x(t)r(t) \tag{14c}$$

$$\dot{u}(t) = \frac{U_0 - u(t)}{\tau_{\rm f}} + U_0(1 - u(t))r(t) \quad . \tag{14d}$$

The macroscopic dynamics generated by the neural mass model (14) and by the QIF network with μ-STP and m-STP are compared in the sub-section Network dynamics versus neural mass evolution of the Section Results. In particular in order to compare the value obtained from the simulation of the neural mass model (14) with those obtained by the integration of the microscopic networks, we evaluated the following instantaneous population average from the microscopic variables $v(t) = \langle V_i(t)\rangle$, $x(t) = \langle x_i\rangle$ and $u(t) = \langle u_i\rangle$. To estimate the instantaneous population firing activity $r(t)$, appearing in (14), from the microscopic simulations is more complicated. This estimation is based on the spike-count, namely one counts all the spikes $N_s(W)$ emitted in the network within a time window W. Then the firing rate is estimated as $r(t) = N_s(W)/W$ with $W = 0.01\tau_{\rm m}$. The neural mass model has been numerically

integrated by employing the adaptive Dormand-Prince method with a absolute and relative tolerances of 10^{-12}.

Multi-populations models

The discussed models can be easily extended to account for multiple interconnected neuronal populations. In the following we restrict this extension to the QIF network with m-STP and the corresponding neural mass model, since the simulations we performed for the QIF network with μ-STP are limited to a single population.

We consider a network composed of one inhibitory and $N_{\rm pop}$ excitatory interacting neural populations, each composed by N_k neurons. Therefore the dynamics of the membrane potential $V_{i,k}(t)$ of the i-th QIF neuron of the k-th population ($k = 0, \ldots, N_{\rm pop}$) and of the mesoscopic synaptic variables $(u_k(t), x_k(t))$ for the excitatory populations ($k > 0$) can be written as follows

$$\tau_{\rm m}^n \dot{V}_{i,k} = V_{i,k}^2(t) + \eta_{i,k} + I_{\rm B} + I_{\rm S}^{(k)}(t) + \tau_{\rm m}^n \sum_{l=0}^{N_{\rm pop}} \frac{1}{N_l} \sum_{j(l)=1}^{N_l} \tilde{J}_{ij,kl}(t) S_{j,l}(t) \tag{15a}$$

$$i = 1, \ldots, N_k$$

$$\dot{x}_k = \frac{1 - x_k(t)}{\tau_{\rm d}} - u_k(t) x_k(t) A_k(t) \tag{15b}$$

$$\dot{u}_k = \frac{U_0 - u_k(t)}{\tau_{\rm f}} + U_0 (1 - u_k) A_k(t) \ ; \tag{15c}$$

where $I_{\rm S}^{(k)}(t)$ is the stimulation current applied to the population k and $A_k(t)$ is the population activity of the k-th population. The synaptic couplings for excitatory-excitatory connections are plastic, therefore they can be written as

$$\tilde{J}_{ij,kl}(t) = J_{kl} u_l(t) x_l(t) \quad ; \tag{16}$$

while if one of the populations k or l is inhibitory we will simply set $\tilde{J}_{ij,kl}(t) = J_{kl}$, where neuron i (j) belongs to population k (l). Moreover, we have assumed that the neurons are globally coupled at both the intra- and inter-population level and characterized by the same directed synaptic couplings strength J_{kl}, whose value depends only on the belonging populations. The sign is determined by the pre-synaptic population l, if it is excitatory (inhibitory) $J_{kl} > 0$ ($J_{kl} < 0$). In this paper we have always considered a single inhibitory population indexed as $k = 0$ and $N_{\rm pop}$ excitatory populations with positive indices. Furthermore, we have usually assumed $\tau_{\rm m}^n = 15$ ms for excitatory and inhibitory populations, apart in sub-section Multi-Item Memory Loading where we set $\tau_{\rm m}^{\rm e} = 15$ ms and $\tau_{\rm m}^{\rm i} = 10$ ms. For each population, we always considered $N_k = 200,000$ neurons.

The corresponding multi-population neural mass model can be straightforwardly written as

$$\tau_{\rm m}^n \dot{r}_k = \frac{\Delta_k}{\tau_{\rm m}^n \pi} + 2 r_k(t) v_k(t) \quad k = 0, 1, \ldots, N_{\rm pop} \tag{17a}$$

$$\tau_{\rm m}^n \dot{v}_k = v_k^2(t) + H_k + I_{\rm B} + I_{\rm S}^{(k)}(t) - (\pi \tau_{\rm m}^n r_k(t))^2 + \tau_{\rm m}^n \sum_{l=0}^{N_{\rm pop}} \tilde{J}_{kl}(t) r_l(t) \tag{17b}$$

$$\dot{x}_l = \frac{1 - x_l(t)}{\tau_{\rm d}} - u_l(t) x_l(t) r_l(t) \tag{17c}$$

$$\dot{u}_l = \frac{U_0 - u_l(t)}{\tau_{\rm f}} + U_0 (1 - u_l(t)) r_l(t) \qquad l = 1, \ldots, N_{\rm pop} \quad ; \tag{17d}$$

where for excitatory-excitatory population interactions

$$\tilde{J}_{kl}(t) = J_{kl} u_l(t) x_l(t) \quad ; \tag{18}$$

and whenever population k or l is inhibitory $\tilde{J}_{kl}(t) = J_{kl}$.

Bifurcation analysis

In order to exemplify the dynamical mechanisms underlying the maintenance of different items in presence of STP, we analyse the bifurcation diagram for the model (17) for three populations. In particular we study the emergence of the different dynamical states occurring in a network made of $N_{\rm pop} = 2$ excitatory and one inhibitory population corresponding to the network architecture introduced in sub-section Multi-Item Architecture.

Due to the symmetries in the synaptic couplings and in the structure of the dynamical equations (17) the macroscopic dynamics observable for the two excitatory networks will be equivalent. Therefore, we will display the bifurcation diagram in terms of the instantaneous firing rate of one of the two excitatory populations $r_k(t)$ with $k > 0$ as a function of the common background current $I_{\rm B}$ by fixing all the other parameters of the model (17) as in Fig. 3. The phase diagram, shown in Fig. 9, reveals that at low values of the background current ($I_{\rm B} \leq I_{\rm sn}^{(1)} \simeq 1.2532$) there is a single stable fixed point with an asynchronous low firing dynamics. This looses stability at $I_{\rm bp}^{(1)} \simeq 1.25647$ giving rise to two coexisting stable fixed points with asynchronous dynamics: one at low firing rate due to spontaneous activity in the network and one at high firing rate corresponding to a persistent state. As shown in the inset in Fig. 9, there is a small region $I_{\rm B} \in [I_{\rm sn}^{(1)}, I_{\rm bp}^{(1)}]$ where we can have the coexistence of these three stable asynchronous states.

At $I_{\rm B} \equiv I_{\rm hb}^{(1)} \simeq 1.34998$ we observe the emergence of coexisting collective oscillations (periodic PBs) at low and high firing rates via supercritical Hopf bifurcations. These oscillations exist in a quite limited range of parameters, namely $I_{\rm B} \in [I_{\rm hb}^{(1)}, I_{\rm hb}^{(2)}]$, and they disappear at $I_{\rm hb}^{(2)} \simeq 1.5363$. Beyond this parameter value we again have a persistent state coexisting with a low firing activity regime, these states finally annihilate with two unstable fixed point branches at $I_{\rm sn}^{(2)} \equiv 4.13715$.

The knowledge of the bifurcation diagram shown in Fig. 9 allows to interpret the numerical experiments discussed in sub-section Memory maintenance with synaptic facilitation and displayed in Fig. 3.

Let us consider the first experiment, reported in column (A) of Fig. 9 and showing the selective reactivation of the WM item via a nonspecific read-out signal. The item is firstly loaded into population one via a specific step current of amplitude $\Delta I^{(1)}(t) = 0.2$ for a time interval $\Delta T_1 = 350$ ms. The selective reactivation of the target is obtained by applying a non-specific readout signal of amplitude $\Delta I^{(1)}(t) = \Delta I^{(2)}(t) = 0.1$ to all excitatory neurons in both populations for a shorter time interval, namely $\Delta T_1 = \Delta T_2 = 250$ ms. For this numerical experiment we fixed $I_{\rm B} = 1.2 < I_{\rm sn}^{(1)}$, thus the only possible stable regime is a low firing asynchronous activity. The item is loaded into population one by increasing, for a limited time window, the background current only for this population to the level $I_{\rm B} + \Delta I^{(1)} > I_{\rm hb}^{(1)}$, thus leading to the emission of a series of PBs, whose final effect is to strongly facilitate the efferent synapses of population one. The sub-sequent application of a non-specific read-out signal amounts to an effective increase of their common background current to a value beyond $I_{\rm bp}^{(1)}$, where the persistent state coexists with the low firing activity. Indeed during the read-out stimulation, population one displays a burst of high activity, due to the facilitated state of its synapses, while population two is essentially unaffected by the

read-out signal. As soon as the stimulus is removed, the system moves back to the spontaneous activity regime.

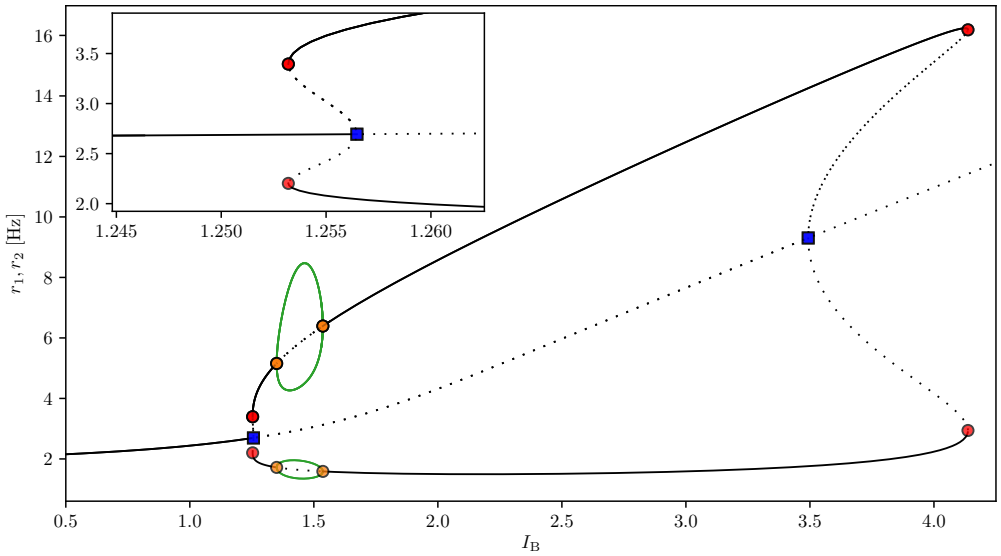


Fig 9. Bifurcation diagram for two excitatory populations. Bifurcation diagram displaying the instantaneous firing rate $r_k(t)$ for the excitatory populations as a function of the background current I_B. Solid (dashed) black lines refer to stable (unstable) asynchronous states, while green solid lines denote the maxima and the minima of stable collective oscillations. Symbol refer to bifurcation points: branch points (blue squares), Hopf bifurcations (orange circles) and saddle-node bifurcations (red circles). The inset displays an enlargement of the bifurcation diagram. All the remaining parameters are as in Fig. 3. Continuation performed with the software AUTO-07P [61].

The second experiment, shown in the column (B) of Fig. 3, concerns spontaneous reactivation of the WM item via collective oscillations (periodic PBs). In this case the background current is set to $I_B = 1.532$, in order to have the coexistence of two stable limit cycles corresponding to periodic PBs with low and high firing rate. The whole system is initialized in the asynchronous regime with spontaneous activity (which is unstable for this value of I_B); upon the presentation of the stimulus to population one, this jumps to the upper limit cycle after loading the item into the memory. Once the stimulation is removed, due to the periodic synaptic refreshment, population one reactivates spontaneously by emitting a periodic sequence of PBs that is terminated by reducing I_B to a value smaller than $I_{sn}^{(1)}$.

The last experiment shown in column (C) of Fig. 3 refers to the spontaneous reactivation of the memory associated to a persistent state activity. In this case we set $I_B = 2$, thus the system is in an asynchronous regime beyond $I_{hb}^{(2)}$, where there is coexistence of persistent and low firing activity beyond $I_{hb}^{(2)}$. As in the previous experiment, the system is initialized in the asynchronous unstable regime. Upon a brief stimulation, population one is led in the high activity persistent regime. Reducing the background current to $I_B = 1.2$ stops the persistent activity.

Spectrogram Estimation

In order to generate the spectrograms shown in Figs. 3, 4, 6, 7, Fig. S2 and Fig. S3, the *signal* package, which is part of the SciPy library [62], is used. The subroutine *stft*

(short time Fourier transform, STFT) generates Fourier transforms $\mathcal{F}[s(t)](t,f)$ of a signal $s(t)$ within a running time window of length $\Delta T_{\rm win}$ at time t. The STFT is performed using overlapping windows (95% overlap) throughout this work. For Figs. 3 and 4 the window length is set to $\Delta T_{\rm win} = 0.2$ s, leading to a sufficiently fine resolution in time and frequency. For Figs. 6, 7, Fig. S2 and Fig. S3 it is set to $\Delta T_{\rm win} = 1$ s, resulting in a better frequency resolution and decrease of time-resolution. The colours in the spectrograms code the normalized power spectral density $|\mathcal{F}[v_k(t)](t,f)|^2/(\max|\mathcal{F}[v_k(t)](t,f)|^2)$ obtained from voltage signals v_k of different populations. For better visibility a log10 scale is used and values $< 10^{-2}$ set to 10^{-2}. We employed the standard Jet colourmap for the colour coding.

Maximal Working Memory Capacity

By following [26] we can give an estimate of the maximal memory capacity for our neural mass model (1) with m-STP (2). The maximal capacity can be estimated as the ratio of two time intervals

$$N_{\rm c}^{\rm max} \simeq \frac{T_{\rm c}^{\rm max}}{T_{\rm b}} \tag{19}$$

where $T_{\rm c}^{\rm max}$ is the maximal period of the network limit cycle and $T_{\rm b}$ the inter-burst interval among two successive PBs. In [26], $T_{\rm c}^{\rm max}$ has been estimated as the time needed to the synaptic efficacy $u_k(t)x_k(t)$ of a generic population k to recover to the maximum value after a PB emission. Since all the excitatory populations are identically connected among them and with the inhibitory population, this time does not depend on the considered population. The approximate expression reported in [26] is the following

$$T_{\rm c}^{\rm max} \simeq \tau_{\rm d} \ln \frac{\tau_{\rm f}/\tau_{\rm d}}{1-U_0} \tag{20}$$

As expected the recovery time is essentially ruled by the depression time scale $\tau_{\rm d}$.

It can be shown that $T_{\rm b}$ has three components, i.e. the time width of the previous excitatory PB, the delay of the inhibitory burst triggered by the excitatory PB, plus its time width, and the time needed for the next active excitatory population to recover from inhibition and to elicit a PB. In our model framework, we can neglect the first two time intervals and limit ourselves to estimate the latter time.

Let us denote the next firing population as the m-th one, we can assume that during $T_{\rm b}$ the connection strength does not vary much and that the firing rates are essentially constants. Therefore we can rewrite the time evolution of the mean membrane potential appearing in (1) as follows :

$$\tau_{\rm m}^{\rm e}\dot{v}_k = v_k^2 + \left[H^{(\rm e)} - (\pi\tau_{\rm m}^{\rm e}\bar{r}^{(\rm e)})^2 + I_{\rm B} + \tau_{\rm m}^{\rm e}\left(-|J_{\rm ei}|\bar{r}^{(\rm text)} + \bar{J}\bar{r}^{(\rm e)}\right)\right] = v_k^2 + C \tag{21}$$

where $\bar{J} = [J_{\rm ee}^{(\rm s)} + (N_{\rm pop}-2)J_{\rm ee}^{(\rm c)}]\bar{x}\bar{u}$ takes in account the synaptic efficacy of all excitatory synapses in an effective manner, $\bar{r}^{(\rm i)}$ and $\bar{r}^{(\rm e)}$ are the inhibitory and excitatory population rates and C is the constant quantity within square brackets on the right-hand side. The expression of C can be further simplified by noticing that the quadratic term is negligible and by assuming that the excitatory and inhibitory firing rates are similar. Moreover, by assuming that the excitatory neurons are almost uncoupled during $T_{\rm b}$, one gets:

$$C = \left[H^{(\rm e)} + I_{\rm B} + \tau_{\rm m}^{\rm e}\left(-|J_{\rm ei}| + \bar{J}\right)\frac{\sqrt{H^{(\rm e)} + I_{\rm B}}}{\pi}\right] \tag{22}$$

where $\bar{r}^{(\rm i)} = \bar{r}^{(\rm e)}$ and $\bar{r}^{(\rm e)} = \frac{\sqrt{H^{(\rm e)}+I_{\rm B}}}{\pi}$ as for an isolated QIF neuron driven by the mean excitability and by the background current.

Therefore the time needed to the mean membrane potential to go from an initial negative value $v_m(0) = V_0$, determined by the discharge of the inhibitory neurons, to a *threshold* value V_{th}, where the PB starts to be delivered, is given by

$$T_{\mathrm{b}} = \frac{\tau_{\mathrm{m}}^{\mathrm{e}}}{\sqrt{C}} \left[\arctan \frac{V_{\mathrm{th}}}{\sqrt{C}} - \arctan \frac{V_0}{\sqrt{C}} \right] \simeq \frac{\tau_{\mathrm{m}}^{\mathrm{e}} \pi}{\sqrt{C}} \tag{23}$$

where on the right hand side of the equation we have finally assumed that $V_{th} >> 1$ and $V_0 << -1$.

Thus the following expression for the maximal capacity is obtained

$$N_{\mathrm{c}}^{\max} \simeq \frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}^{\mathrm{e}}} \ln \left[\frac{\tau_{\mathrm{f}}/\tau_{\mathrm{d}}}{1 - U_0} \right] \frac{\sqrt{C}}{\pi} \quad . \tag{24}$$

For the parameters employed in sub-section Multi-Item Memory Loading we obtained the following theoretical values $T_{\mathrm{c}}^{\max} \simeq 447$ ms, $T_{\mathrm{b}} \simeq 93 - 126$ ms depending on the value of $0.5 \leq \bar{x}\bar{u} \leq 1.0$, thus $3.6 \leq N_{\mathrm{c}}^{\max} \leq 4.8$ not far from the measured value that was $N_{\mathrm{c}}^{\max} = 5$.

Funding

AT received financial support by the Excellence Initiative I-Site Paris Seine (Grant No ANR-16-IDEX-008), by the Labex MME-DII (Grant No ANR-11-LBX-0023-01) and by the ANR Project ERMUNDY (Grant No ANR-18-CE37-0014), all part of the French programme "Investissements d'Avenir".

Acknowledgements

The authors acknowledge preliminary discussions with D. Avitabile, M. Desroches, and E. Montbrió.

Author Contributions

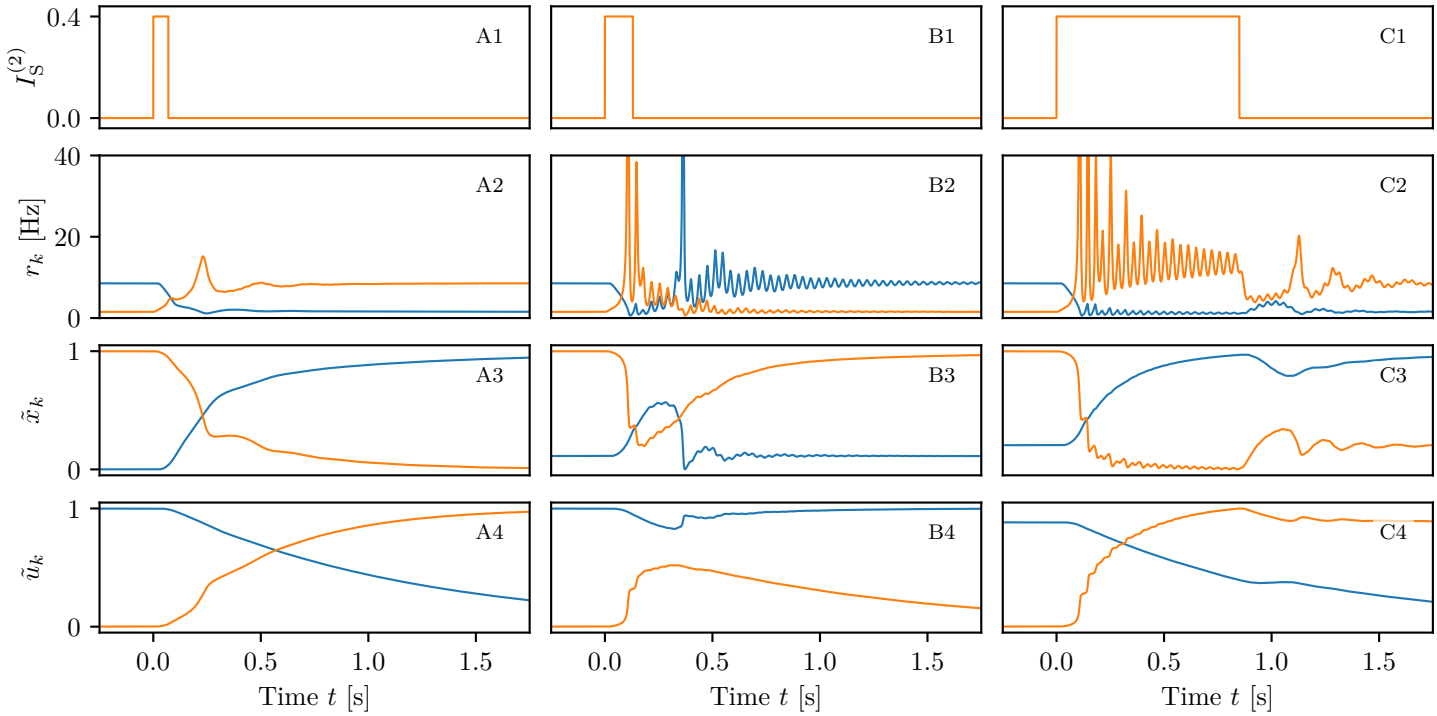
HT performed the simulations and data analysis. SO and AT were responsible for the state-of-the-art review and the paper write-up. All the authors conceived and planned the research.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

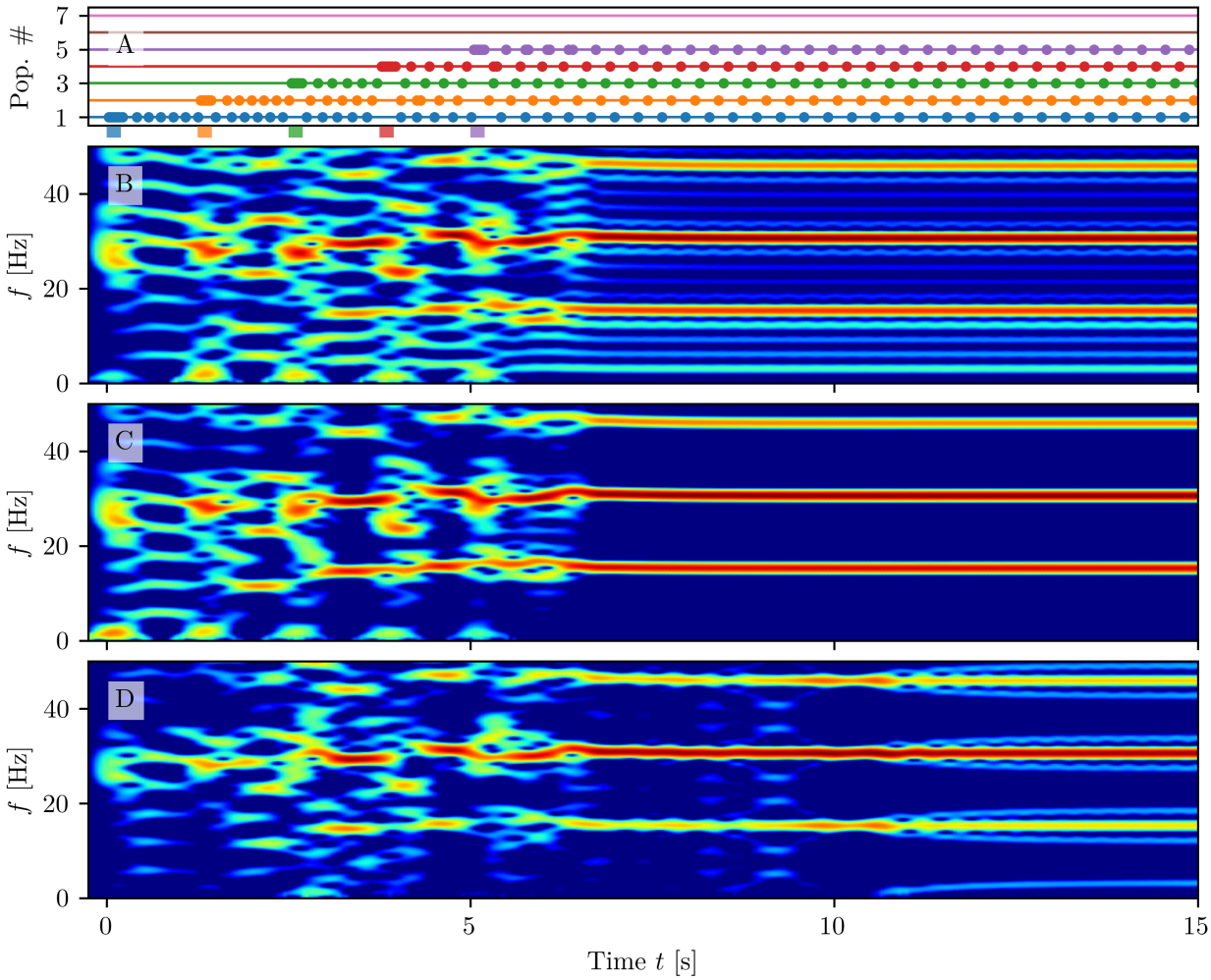
Supporting information

Fig. S1



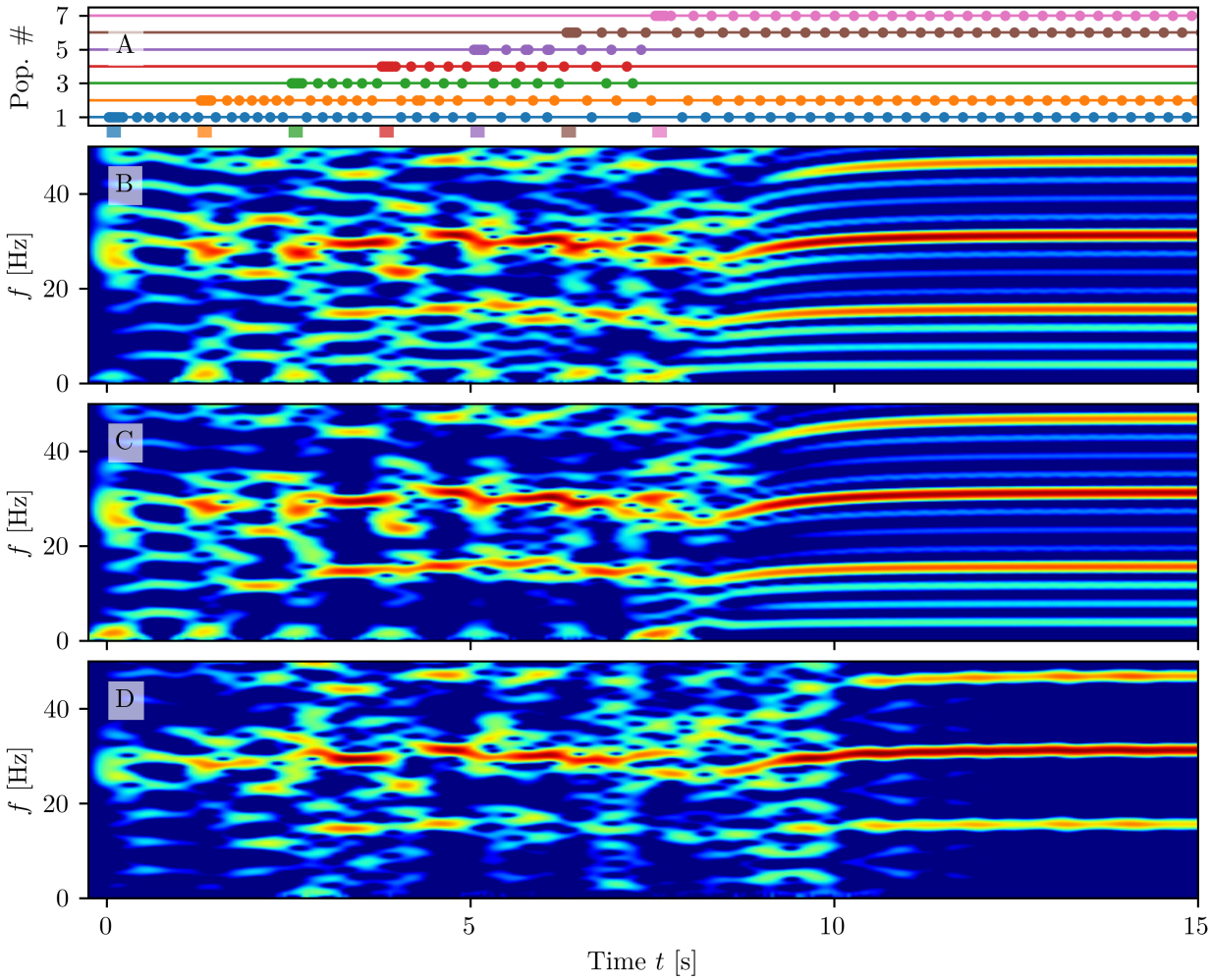
Memory item switching with persistent activity. The three columns (A-C) refer to three values of ΔT_2 for which we have a memory switching from one item to the other for a background current $I_B = 2$ supporting persistent state activity. (A) $\Delta T_2 = 70$ ms; (B) $\Delta T_2 = 130$ ms and (C) $\Delta T_2 = 850$ ms. Profiles of the stimulation current $I_S^{(2)}(t)$ for the second excitatory population (**A1-C1**). Population firing rates r_l(t) (**A2-C2**), normalized available resources $\tilde{x}_l(t)$ (**A3-C3**) and normalised utilization factors $\tilde{u}_l(t)$ (**A4-C4**) of the excitatory populations calculated from the simulations of the neural mass model. Other parameters as in Fig. 3.

Fig. S2



Multi-item memory loading: 5 items (**A**) Population spikes of excitatory populations: horizontal lines in absence of dots indicate quiescence phases at low firing rates r_k for populations $k = 1, ...7$. Dots mark population spikes of the corresponding population. The coloured bars on the time axis mark the starting and ending time of stimulating pulses, targeting each population. Spectrograms of the mean membrane potential $v_1(t)$ (**B**), $v_0(t)$ (**C**) and of the mean membrane potential averaged over all the excitatory populations (**D**). Parameters as in Fig. 6.

Fig. S3



Multi-item memory loading: 7 items Population spikes of excitatory populations: horizontal lines in absence of dots indicate quiescence phases at low firing rates r_k for populations $k = 1, ...7$. Dots mark population spikes of the corresponding population. The coloured bars on the time axis mark the starting and ending time of stimulating pulses, targeting each population. Spectrograms of the mean membrane potential $v_1(t)$ (**B**), $v_0(t)$ (**C**) and of the mean membrane potential averaged over all the excitatory populations (**D**). Parameters as in Fig. 6.

References

1. Just MA, Carpenter PA. A capacity theory of comprehension: individual differences in working memory. Psychological review. 1992;99(1):122.

2. Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995;14(3):477 – 485. doi:https://doi.org/10.1016/0896-6273(95)90304-6.

3. Engle RW, Tuholski SW, Laughlin JE, Conway AR. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of experimental psychology: General. 1999;128(3):309.

4. Fuster JM. Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate. MIT press; 1999.

5. Vogel EK, Machizawa MG. Neural activity predicts individual differences in visual working memory capacity. Nature. 2004;428(6984):748–751.

6. Chatham CH, Badre D. Multiple gates on working memory. Current opinion in behavioral sciences. 2015;1:23–31.

7. Fuster JM, Alexander GE. Neuron activity related to short-term memory. Science. 1971;173(3997):652–654.

8. Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annual review of neuroscience. 2001;24(1):167–202.

9. Lara AH, Wallis JD. The role of prefrontal cortex in working memory: a mini review. Frontiers in systems neuroscience. 2015;9:173.

10. Shafi M, Zhou Y, Quintana J, Chow C, Fuster J, Bodner M. Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience. 2007;146(3):1082–1108.

11. Lundqvist M, Rose J, Herman P, Brincat SL, Buschman TJ, Miller EK. Gamma and beta bursts underlie working memory. Neuron. 2016;90(1):152–164.

12. Hussar CR, Pasternak T. Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. Journal of Neuroscience. 2012;32(8):2747–2761.

13. Fujisawa S, Amarasingham A, Harrison MT, Buzsáki G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nature neuroscience. 2008;11(7):823.

14. Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nature neuroscience. 2006;9(4):534–542.

15. Mongillo G, Barak O, Tsodyks M. Synaptic Theory of Working Memory. Science. 2008;319(5869):1543–1546. doi:10.1126/science.1150769.

16. Lundqvist M, Herman P, Lansner A. Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. Journal of cognitive neuroscience. 2011;23(10):3008–3020.

17. Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, et al. Reactivation of latent working memories with transcranial magnetic stimulation. Science. 2016;354(6316):1136–1139.

18. Miller EK, Lundqvist M, Bastos AM. Working Memory 2.0. Neuron. 2018;100(2):463–475.

19. Siegel M, Warden MR, Miller EK. Phase-dependent neuronal coding of objects in short-term memory. Proceedings of the National Academy of Sciences. 2009;106(50):21341–21346.

20. Tsodyks MV, Markram H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proceedings of the national academy of sciences. 1997;94(2):719–723.

21. Tsodyks M, Pawelzik K, Markram H. Neural networks with dynamic synapses. Neural computation. 1998;10(4):821–835.

22. Markram H, Wang Y, Tsodyks M. Differential signaling via the same axon of neocortical pyramidal neurons. Proceedings of the National Academy of Sciences. 1998;95(9):5323–5328.

23. Tsodyks M, Uziel A, Markram H. Synchrony generation in recurrent networks with frequency-dependent synapses. J Neurosci. 2000;20(1):825–835.

24. Luccioli S, Ben-Jacob E, Barzilai A, Bonifazi P, Torcini A. Clique of functional hubs orchestrates population bursts in developmentally regulated neural networks. PLoS Computational Biology. 2014;10(9).

25. Wilson HR, Cowan JD. Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical journal. 1972;12(1):1–24.

26. Mi Y, Katkov M, Tsodyks M. Synaptic correlates of working memory capacity. Neuron. 2017;93(2):323–330.

27. Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang XJ. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proceedings of the National Academy of Sciences. 2017;114(2):394–399.

28. Ermentrout GB, Kopell N. Parabolic bursting in an excitable system coupled with a slow oscillation. SIAM Journal on Applied Mathematics. 1986;46(2):233–253.

29. Ott E, Antonsen TM. Low dimensional behavior of large systems of globally coupled oscillators. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2008;18(3):037113.

30. Luke TB, Barreto E, So P. Complete classification of the macroscopic behavior of a heterogeneous network of theta neurons. Neural computation. 2013;25(12):3207–3234.

31. Laing CR. Derivation of a neural field model from a network of theta neurons. Physical Review E. 2014;90(1):010901.

32. Montbrió E, Pazó D, Roxin A. Macroscopic Description for Networks of Spiking Neurons. Phys Rev X. 2015;5:021028. doi:10.1103/PhysRevX.5.021028.

33. Devalle F, Roxin A, Montbrió E. Firing rate equations require a spike synchrony mechanism to correctly describe fast oscillations in inhibitory networks. PLoS computational biology. 2017;13(12):e1005881.

34. Laing CR. Phase oscillator network models of brain dynamics. Computational models of brain and behavior. 2017; p. 505–517.

35. Coombes S, Byrne A. Next generation neural mass models. In: Corinto F, Torcini A, editors. Nonlinear Dynamics in Computational Neuroscience. Springer; 2019. p. 1–16.

36. Dumont G, Gutkin B. Macroscopic phase resetting-curves determine oscillatory coherence and signal transfer in inter-coupled neural circuits. PLoS computational biology. 2019;15(5):e1007019.

37. di Volo M, Torcini A. Transition from asynchronous to oscillatory dynamics in balanced spiking networks with instantaneous synapses. Physical review letters. 2018;121(12):128301.

38. Segneri M, Bi H, Olmi S, Torcini A. Theta-nested gamma oscillations in next generation neural mass models. Front Comput Neurosci. 2020;14(47).

39. Ceni A, Olmi S, Torcini A, Angulo-Garcia D. Cross frequency coupling in next generation inhibitory neural mass models. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2020;30(5):053121.

40. Bi H, Segneri M, di Volo M, Torcini A. Coexistence of fast and slow gamma oscillations in one population of inhibitory spiking neurons. Physical Review Research. 2020;2(1):013042.

41. Okun M, Naim A, Lampl I. The subthreshold relation between cortical local field potential and neuronal firing unveiled by intracellular recordings in awake rats. Journal of neuroscience. 2010;30(12):4440–4448.

42. Amit DJ, Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cerebral cortex (New York, NY: 1991). 1997;7(3):237–252.

43. Fino E, Yuste R. Dense inhibitory connectivity in neocortex. Neuron. 2011;69(6):1188–1203.

44. di Volo M, Livi R, Luccioli S, Politi A, Torcini A. Synchronous dynamics in the presence of short-term plasticity. Physical Review E. 2013;87(3):032801.

45. Schmutz V, Gerstner W, Schwalger T. Mesoscopic population equations for spiking neural networks with synaptic short-term plasticity. The Journal of Mathematical Neuroscience. 2020;10(1):1–32.

46. Tiesinga P, Sejnowski TJ. Cortical enlightenment: are attentional gamma oscillations driven by ING or PING? Neuron. 2009;63(6):727–732.

47. Lewis-Peacock JA, Norman KA. Competition between items in working memory leads to forgetting. Nature Communications. 2014;5(1):1–10.

48. Lai YC, Tél T. Transient chaos: complex dynamics on finite time scales. vol. 173. Springer Science & Business Media; 2011.

49. Cortes JM, Desroches M, Rodrigues S, Veltz R, Muñoz MA, Sejnowski TJ. Short-term synaptic plasticity in the deterministic Tsodyks–Markram model leads to unpredictable network dynamics. Proceedings of the National Academy of Sciences. 2013;110(41):16610–16615.

50. Olmi S, Politi A, Torcini A. Stability of the splay state in networks of pulse-coupled neurons. The Journal of Mathematical Neuroscience. 2012;2(1):12.

51. Postman L, Phillips LW. Short-term temporal changes in free recall. Quarterly journal of experimental psychology. 1965;17(2):132–138.

52. Morrison AB, Conway AR, Chein JM. Primacy and recency effects as indices of the focus of attention. Frontiers in human neuroscience. 2014;8:6.

53. Kornblith S, Buschman TJ, Miller EK. Stimulus load and oscillatory activity in higher cortex. Cerebral Cortex. 2016;26(9):3772–3784.

54. Howard MW, Rizzuto DS, Caplan JB, Madsen JR, Lisman J, Aschenbrenner-Scheibe R, et al. Gamma oscillations correlate with working memory load in humans. Cerebral cortex. 2003;13(12):1369–1374.

55. van Vugt MK, Schulze-Bonhage A, Litt B, Brandt A, Kahana MJ. Hippocampal gamma oscillations increase with memory load. Journal of Neuroscience. 2010;30(7):2694–2699.

56. Roux F, Wibral M, Mohr HM, Singer W, Uhlhaas PJ. Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. Journal of Neuroscience. 2012;32(36):12411–12420.

57. Dipoppa M, Gutkin BS. Flexible frequency control of cortical oscillations enables computations required for working memory. Proceedings of the National Academy of Sciences. 2013;110(31):12828–12833.

58. Dipoppa M, Szwed M, Gutkin BS. Controlling working memory operations by selective gating: the roles of oscillations and synchrony. Advances in cognitive psychology. 2016;12(4):209.

59. Schmidt H, Avitabile D, Montbrió E, Roxin A. Network mechanisms underlying the role of oscillations in cognitive tasks. PLoS computational biology. 2018;14(9):e1006430.

60. Wang Y, Markram H, Goodman PH, Berger TK, Ma J, Goldman-Rakic PS. Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nature neuroscience. 2006;9(4):534–542.

61. Doedel EJ, Fairgrieve TF, Sandstede B, Champneys AR, Kuznetsov YA, Wang X. AUTO-07P: Continuation and bifurcation software for ordinary differential equations; 2007.

62. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:https://doi.org/10.1038/s41592-019-0686-2.