Synergistic short-term synaptic plasticity mechanisms for working memory

Abstract

Working memory (WM) is essential for almost every cognitive task and behavior. The neural and synaptic mechanisms supporting the rapid encoding and maintenance of memories in diverse tasks are the subject of an ongoing debate. The traditional view of WM as stationary persistent firing of selective neuronal populations has given room to newer ideas regarding mechanisms that support a more dynamic maintenance of multiple items, which may also tolerate more activity disruption. Various computational WM models based on different biologically plausible synaptic and neural plasticity mechanisms have been proposed. We show that these proposed short-term plasticity mechanisms may not necessarily be competing explanations, but instead yield interesting functional interactions on a wide set of WM tasks and enhance the biological plausibility of spiking neural network models, in particular of the underlying synaptic plasticity. While monolithic models (WM function explained by one particular mechanism) are theoretically appealing and have increased our understanding of specific mechanisms, they are narrow explanations. WM models need to become more capable, robust and flexible to account for new experimental evidence of bursty and activity-silent multi-item maintenance in more challenging WM tasks, and generally solve more than one particular task. More detailed models also allow for electrophysiological constraints from recordings.

In this study we evaluate the interactions between three commonly used classes of plasticity, namely intrinsic excitability, synaptic facilitation/augmentation and Hebbian plasticity. Combinations of these are systematically tested in a spiking neural network model on a broad suite of tasks or functional motifs deemed principally important for WM operation, such as one-shot encoding, free and cued recall, delay maintenance and updating. In our evaluation we focus on the operational task performance and biological plausibility. Our results indicate that a composite model, combining several commonly proposed plasticity mechanisms for WM function, is superior to more reductionist variants. Importantly, we attribute the observable differences to the principle nature of specific types of plasticity.

Introduction

Working memory (WM) has attracted a lot of attention in brain science largely due to its central role in cognition. Consequently, it has been the subject of numerous theoretical and experimental investigations as well as scientific debates. Recently, a multi-faceted dispute over the neural bases of WM, and in particular about the involvement of so-called persistent activity in WM delay maintenance, has come alive [1,2]. As a result, this has rekindled interest in synaptic and neural plasticity mechanisms for short-term memory that can buffer information in neural and synaptic state changes without any neuronal spiking manifestations (activity silent) [3–7]. It is tempting from a theoretical perspective to identify one particular biophysically motivated mechanism for implementing such a short-term memory

process. History has seen many notable attempts [8,9], yet biology is not reductionist and often relies on several mechanisms acting in concert to enable robust function. Synaptic plasticity is diverse across different types of neurons and connections [10]. Most of the proposed short-term plasticity mechanisms of relevance to WM can be broadly grouped into three categories: presynaptic, postsynaptic and pre-post-synaptic.

Presynaptic

The most cited candidate mechanism for presynaptic plasticity-based WM is synaptic facilitation or augmentation [11], which transiently enhances synaptic efficacy via a calcium-mediated increase in synaptic release probability with dominating time constants up to 5s [12,13]. In the short term, increased vesicle release from synaptic facilitation counteracts depression from the depletion of readily releasable transmitter, but longer lasting synaptic augmentation may store excitability over a few seconds. Clear evidence for powerful synaptic augmentation in the medial prefrontal cortex (mPFC) was obtained using extracellular stimulation [14] and patch-clamp recordings [12]. Consequently, there have been several models examining this synaptic WM hypothesis to explain WM delay maintenance [15–20] using a Tsodyks-Markram formalism [21], which - despite the complexity of the involved biological processes - is perhaps the most efficient and widely accepted phenomenological model of short-term plasticity (STP).

Postsynaptic

Neuronal (i.e. intrinsic) excitability is a neuron's tendency to generate action potentials (APs) upon synaptic integration. Rapid intrinsic excitability changes from one-shot learning, such as adaptation, plateau potentials, or graded increases in excitability [22] have also been proposed as a cellular analog of WM [23–25] and address specific WM phenomena, such as serial order effects in computational models [26–29]. This type of plasticity has been leveraged to achieve specific learning dynamics in computational models, such as Bayesian computation [30] and may even match the known experimental dynamics of graded increases in excitability quantitatively [31].

Pre-Post-synaptic (associative)

Associative (i.e. Hebbian) synaptic plasticity, dependent on pre- and post-synaptic co-activity has long been held as the most important neural mechanism behind different forms of semantic and episodic long-term memories. Computational models often use variants of spike timing dependent plasticity (STDP) to implement Hebbian learning [32]. Proposals that short-term Hebbian potentiation, such as NMDAR-STP may also be of importance to WM are less common [33–35], but increasingly part of the discussion [36,37]. This type of fast synaptic plasticity has also been difficult to study in detail with current experimental technology, in part because LTP induction protocols and baseline monitoring may disrupt it [38]. Theoretical neuroscience [39] and recent reviews [37] have suggested fast Hebbian plasticity as a WM mechanism. Neural network models capable of fast Hebbian learning generally support activity-silent WM with promising results and robust effects in attractor memory networks, compatible with many electrophysiological and cognitive WM findings [29,40–42]. The Bayesian Confidence Propagation Neural Network (BCPNN) learning apparatus [29,42–44] deserves special attention in this regard as it inherently encapsulates associative Hebbian plasticity

with intrinsic neuronal plasticity within an elegant phenomenological framework of Bayesian inference.

Computational synergies of WM mechanisms

Synergies between different plasticity mechanisms have been explored to equip models with capabilities enhancing robustness, stability or functionality and performance. Fast Hebbian plasticity mechanisms enabling the formation of new neuronal assemblies through strong recurrent excitatory connections are generally unstable and produce implausibly high firing rates (and excitatory postsynaptic potentials in spiking models). They thus require secondary mechanisms to stabilize runaway dynamics [45]. Synergies of Hebbian and non-Hebbian mechanisms have been shown not only to stabilize network dynamics but also to confer a degree of robustness and reliability to newly encoded memories [46–48].

Historically, early attractor WM models demonstrated a capability to maintain only a single memory item in the delay period. The stability of individual fixed-point attractors either in discrete [49–51] or continuous attractor space [52] supports memory maintenance in an active state. Later it was shown that, a combination of presynaptic facilitation/augmentation with synaptic depression or neural adaptation, making attractors quasi-stable (with short lifetime), gives rise to multi-item synaptic WM [15,17,53,54]. A distinctive prediction of synaptic WM models is that recently active items can be reactivated by a nonselective excitatory signal [55]. Multi-item delay maintenance then relies on multiplexing a few memory items through related short-lived attractor reactivations in time, thereby periodically refreshing the short-term memory traces of a small number of items. The same can be achieved by combining fast Hebbian learning with neural adaptation and synaptic depression instead of synaptic augmentation [29,41]. In each case single-item persistent activity is replaced by intermittent attractor dynamics of multi-item WM delay maintenance. Synergies between Hebbian plasticity and different neural and synaptic plasticity mechanisms also enhance the functional repertoire of WM models in other ways. For example, r-STDP, incorporating a selective extracellular dopamine signal with STDP solves the distal reward problem in the rapid encoding of WM items [56]. Prominent serial order effects in WM have also been explained by an additional component of intrinsic plasticity. So, learning rules that synergistically exploit the two, such as BCPNN [43], facilitate both one-shot encoding and well-known serial order effects such as primacy and recency [29,57].

While many computational WM models rely on combinations of two or more mechanisms, there are hardly any studies offering deeper insights into individual roles of these mechanistic components and how they substantiate the emerging synergies, especially in a wide spectrum of WM functionality. There are some notable isolated exceptions reflecting such intention: for example, Zenke et al. [46] tested the ablation of individual mechanisms and Orhan et al. [58] subjected their plastic short-term memory to a battery of WM testing scenarios to investigate its multi-task capability. Here we propose a more systematic approach that emphasises the contributions of the three aforementioned categories of WM mechanisms. To this end, we select a representative plasticity rule in each category (pre-, post-, pre-post-synaptic) underlying short-term memory effects and combine them in different constellations, together with other biologically plausible activity-limiting mechanisms such as synaptic depression and neural adaptation, in the framework of an attractor spiking neural network model with a cortical columnar architecture. In particular, we heavily exploit the

BCPNN plasticity formalism that can intrinsically account for both associative and postsynaptic effects [31]. The resulting WM models are thoroughly studied with biologically motivated constraints in mind, such as cell-type specific local connectivity, firing rates and the amplitudes of postsynaptic potentials (PSPs). We compare functional capabilities and performance across a variety of WM task demands.

Results

Our modeling approach relies on a spiking attractor memory neural network with a modular architecture and distributed representations, which has previously been used to model cortical memory [29,59] with a microcircuit model enabling various types of plasticity (see Methods, **Fig.M1, Fig.M2**). We use this model to simulate a range of WM tasks that provide insights into its various functional capabilities of key relevance for WM. By selecting one representative implementation for each of the 3 aforementioned categories of short-term plasticity mechanisms, we arrive at 7 meaningful plasticity configurations (**Fig.1**) to be examined across generic WM computational motifs (**Fig.2**), from one-shot novel item encoding, and multi-item delay maintenance with and without task disruption, to cued- and free (uncued, spontaneous) recall. The breadth of functional capabilities of WM models including the flexibility in dynamic adaptation to changing task demands, is often overlooked in computational studies primarily concerned with WM maintenance and recall, but may contribute explanatory power across different task scenarios. To address this gap we also investigate the updating of the task-relevant WM item set. The biophysical detail of the spiking model meanwhile enables us to constrain the model and assess biological plausibility on the basis of emergent electrophysiological dynamics, such as postsynaptic potentials and firing rates among others.

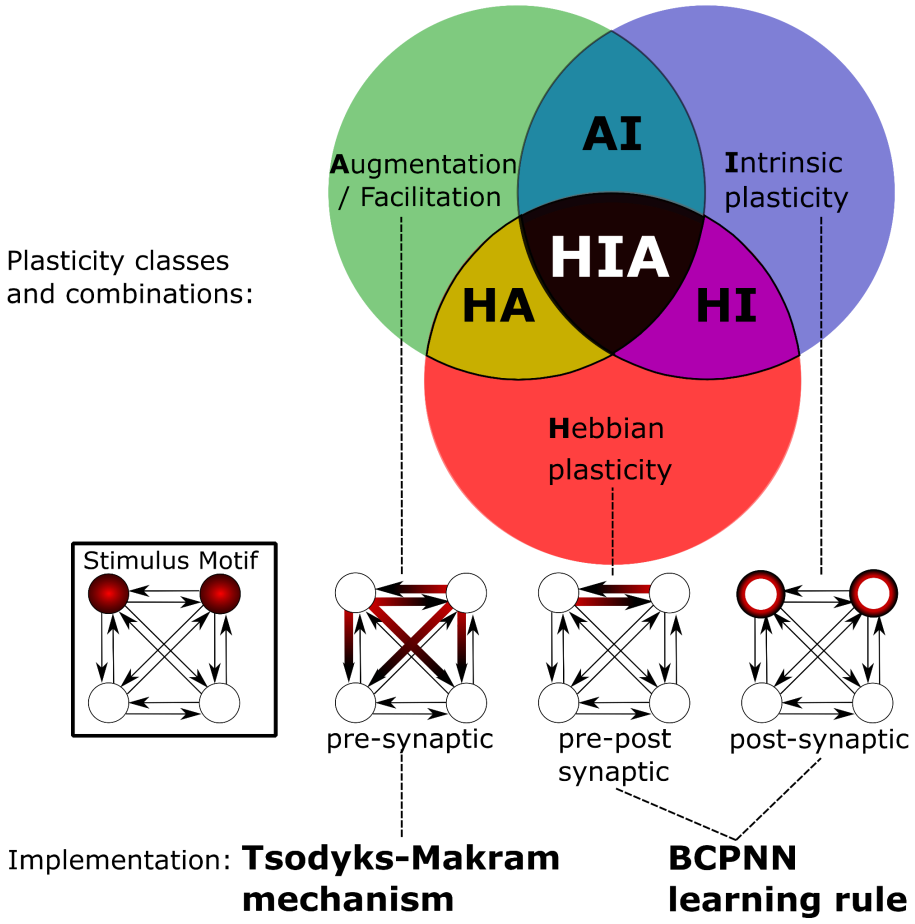

Figure 1 - Plasticity Combinations. The Augmentation plasticity model is implemented using the well-known Tsodyks-Makram mechanism [21]. The Bayesian Confidence Propagation Neural Network (BCPNN) learning rule implements intrinsic plasticity, as well as Hebbian plasticity. These 3 components can be simulated separately or together, yielding 7 scenarios to simulate and study (letters correspond to mechanism initials, e.g., **H** indicates Hebbian plasticity). Below, a simplified network schematic illustrates where different mechanisms may induce plastic changes at neurons and synapses in response to a spiking stimulus motif: **A**ugmentation changes with presynaptic activity, potentiating outgoing synapses of spiking neurons. **I**ntrinsic excitability changes with postsynaptic activity, increasing the excitability of spiking neurons. **H**ebbian learning at recurrent connections is driven by correlations in pre-post spiking.

An overview of a general task structure and performance evaluation scenarios

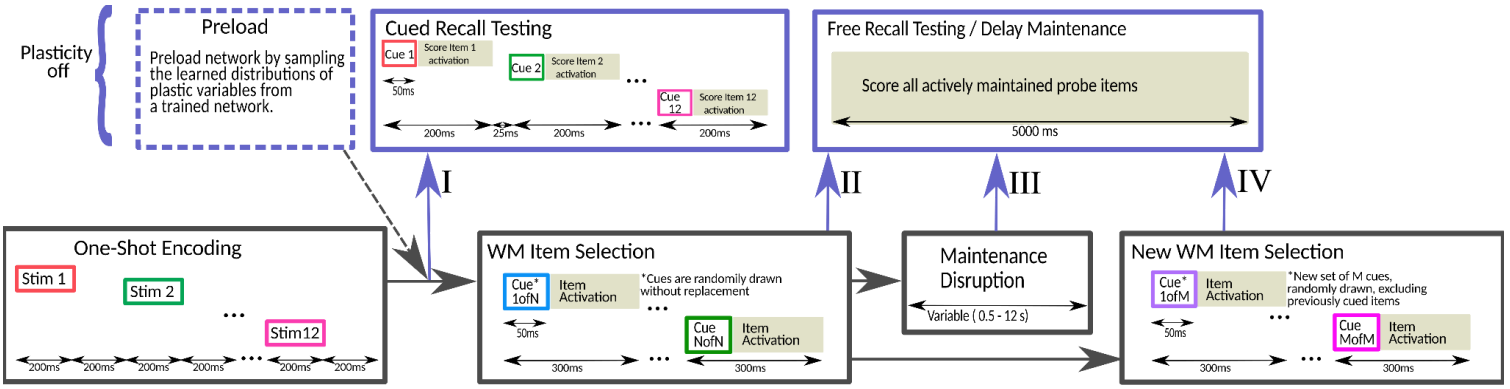

Figure 2 - Overview of a general task structure simulated in different configuration scenarios. All simulated task configurations are sequential flows of memory operations (WM functional motifs) with ongoing plasticity (bottom row) , and memory probes I-IV (top row), during which plasticity is typically disabled to assess the state of memory at that point without subsequent bias from rapid plastic changes induced by the testing protocol (*Methods - Stimulation Protocol, Memory Performance Tracking*). Stimulation is shown in item-colored boxes. Gray background denotes a free running network driven only by unspecific background noise. **One-shot Encoding** sequentially activates 12 neuronal assemblies to encode them as separate memory items into the network (acting as a reservoir of attractor memories). **Cued Recall Testing** activates all encoded items in turn by brief partial cues, checking for the stable recurrent activation of the full memory item thereafter. Probe I: probe all memory items. **WM Item Selection** uses cues to sequentially activate a subset of N unique items; cues are partial and brief to let the network activate the full item representation (entire neuronal assembly) on its own. Probe II: probe the selected WM items. **Maintenance Disruption** transiently lowers the background noise to disrupt Delay maintenance for a variable amount of time. Probe III: re-probe the WM items. **New WM item selection** repeats the previous item selection with M unique and novel items, drawn from the reservoir of 12 encoded items. Probe IV: probe new WM items.

To enable the study of WM operations, memory items have to first be encoded in the recurrent connections of the spiking network. Whilst many published models build on pre-encoded neuronal assemblies, we use the BCPNN learning rule to embed representations in an initially empty attractor memory network. So, prior to any WM operation, One-shot encoding (Fig.2) trains the model through the sequential stimulation of 12 non-overlapping stimuli (distributed sets of MCs encoding memory items, see Methods - Network Architecture and WM Neural Representations and Stimulation Protocol) for later retrieval or other WM operations. In order to save simulation resources and to enable some specific hypothesis testing later, we can also replace this fast online learning with a Preload, where all the network parameters affected by this initial learning are instead sampled from an already trained network, i.e. with encoded items (neural bias, synaptic AMPA and NMDA receptor conductances and the synaptic facilitation state; see Methods). Cued recall testing is used to examine the trained model. In particular, all the encoded items are probed in turn by brief partial cues (see Methods - Stimulation Protocol) in order to recall them (see Methods - Memory Performance Tracking and typical item activity in Fig.3). This way we evaluate a Cued Recall Score, the proportion of successfully recalled WM items.

To show how the network maintains several WM items at the same time, we select a small subset of memory items and track their activations over time. *WM item selection* (Fig.2) uses brief cues to sequentially activate (typically 4) memory items, similar to a set of sample cues

in experimental WM tasks. We refer to this subset of activated items as the *WM set*. We then analyse the ongoing maintenance activity of the network during Delay maintenance / Free recall testing (Fig.2) to detect any memory item reactivations. Importantly, in this period, unlike during Cued recall testing, we do not provide any cues, but merely raise the unspecific background noise (see *Methods - Stimulation Protocol*). The observed memory item activations can be either interpreted as WM delay maintenance through WM replay [60,61] or as a so-called free WM recall [57], depending on the memory task paradigm. In both cases, we expect that only WM items belonging to the active WM set reactivate freely (uncued, spontaneously). To quantify this WM performance, we calculate a *Delay Maintenance Score*, the ratio of all successfully maintained (freely recalled) WM set items during Delay maintenance period over the total WM set size, corresponding to WM load (N in Fig.2). For further details on the identification of memory item activations and scoring, see *Methods - Memory Performance Tracking*.

We also test how robust the active maintenance of multi-item WM is by disrupting the delay period (Fig.2 - Maintenance disruption). We count all memory items as successfully retained after disruption if we then detect repeated free activations of the WM set items once we raise the background excitation again for a period of active Delay maintenance. The *Disruption Score* is the ratio of all successfully reactivated items over the total number of items in the current WM set (see *Methods - Memory Performance Tracking*).

To test the flexibility of the active maintenance we activate an entirely new WM set during *New WM item selection* (Fig.2). We expect the model to update its currently maintained memory content and switch to actively maintain new WM set items at the cost of the previous set items. This WM flexibility is evaluated by probing the active maintenance of the new WM set items with emphasis on the maintenance specificity wrt. the new updated WM set. The *Update Score* is the ratio of all successfully reactivated items from the new set over the total number of items in the new WM set (M in Fig.2). *Specificity* is the ratio of reactivations from the new WM set over the total number of reactivations during the delay period after the WM update (see Methods - Memory Performance Tracking), so unlike the *Update Score*, it accounts not only for the maintenance of the new WM set but also for how effectively the old WM set has been replaced relative the previous Delay maintenance.

Cued recall reveals the need for Hebbian plasticity to encode novel items

Following One-shot encoding, we performed Cued recall testing in each plasticity model (cf. Fig.2). The full plasticity model featuring all three plasticity components (HIA) reliably retrieves all the encoded items. However, once the Hebbian plasticity component is removed (i.e resulting in the AI model) no WM items can be recalled. The spike raster in Fig.3. illustrates a simplified 5-item trial of encoding and subsequent Cued recall testing in both models. The stimulation drives the pyramidal cells in target minicolumns to fire in brief bursts. A similar pattern of activity reoccurs when we test for cued recall thereafter. Brief 50 ms cues to the previously stimulated populations quickly complete into the full spiking cell assembly, now driven by learned, recurrent connections rather than external stimulation. This activation persists until either neural adaptation and synaptic depression wears out the

neuronal assembly, or a new stimulus or cue disrupts the ongoing reverberating activity through mutual inhibition (see Methods, Fig.M2). Notably, the model without Hebbian plasticity shown in Fig.3 (bottom), does not achieve cued recall of any of the 5 items, even though the initial stimuli evoke a similar number of spikes during One-shot encoding.

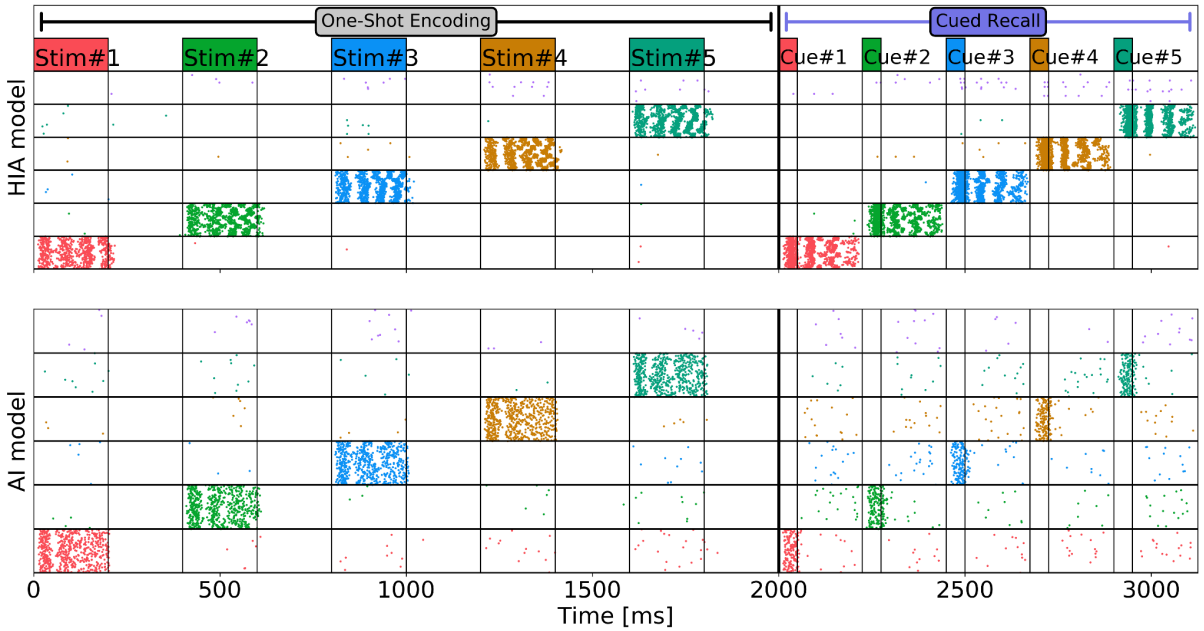

Figure 3 - Subsampled, sorted spike rasters during One-Short Encoding and Cued recall testing epochs. Spikes are sorted and colored by their item-selectivity (only a subset corresponding to the same items during both simulated task epochs is shown). Stimuli and Cues are annotated in the same color on top. TOP: HIA model. BOTTOM: AI model (i.e. the same model, except without Hebbian plasticity).

In fact, the non-Hebbian models (I,A,AI) cannot perform cued recall following the prior attempt at One-shot encoding (see *Methods, Fig.M3*). This is not a matter of tuning the models correctly, but simply because any recall of memory items as dynamic attractors in associative memory networks requires sufficiently strong associative connections to activate. This increase in excitation needs to be specific to the connections between neurons selective for the same item (Hebbian), rather than spread out (e.g. augmentation strengthens all presynaptic inputs). If the relevant synaptic connections are not preloaded (hard-coded) upon initialization of the network they instead need to be learned, yet non-Hebbian learning rules lack this ability [37]. Previously published models of non-Hebbian synaptic WM typically rely on hard-coded associative connections establishing attractors for memory items prior to any WM test [15,17]. If we are to compare non-Hebbian models with Hebbian models on such tasks, e.g. evaluating active delay maintenance, then we can first preload (frozen) associative connections from an otherwise identical network trained with Hebbian plasticity (see *Methods - Preload and Discussion - Preloading WM*).

Compensating the unspecific excitability effects of synergy

The synergistic effect of different plasticity mechanisms naturally results in the enhanced excitability (or spiking activity) of the full HIA model, featuring all forms of plasticity examined in this work, when compared to reduced models. Removing entire plasticity mechanisms can shift the general excitability of the resulting models and in an attractor WM network model, this affects baseline recall performance even prior to any WM task. As we are not primarily

interested in analyzing specific effects of plasticity on general excitability but rather in studying their implications for WM performance in different WM task paradigms, there is a need to establish a general baseline memory excitability across all the models so that a fair and insightful functional comparison can be made.

We have thus functionally compensated all incomplete models by replacing the missing activity-dependent plasticity components with fixed (non-plastic) initializations of the corresponding synaptic or neural parameters tuned to the minimum necessary level enabling cued recall of all the 12 baseline memory items. As explained in the previous section, non-Hebbian models (A,I,AI) are preloaded with fixed weights acquired through Hebbian learning in the same model with Hebbian plasticity. Furthermore, our implementation of intrinsic plasticity affects the excitability of neurons via a plastic neuronal current so the models without intrinsic plasticity (H,A,HA) are compensated via an extra constant neuronal current tuned to be sufficiently strong to enable cued recall (rather than setting this current to zero or some other unmotivated value). Finally, synaptic augmentation - as implemented by the Tsodyks-Makram mechanism - acts as a dynamic multiplier of synaptic efficacy. Accordingly, in all the model configurations without augmentation (H,I,HI) we replace this kind of plasticity with a fixed synaptic efficacy multiplier, adequate to enable cued recall. For further details see Methods - Preload. Having all the models compensated to perform cued recall of all memory items (Fig.2, Probe I), we can now focus on much more interesting comparisons in diverse WM task paradigms.

WM item selection and delay maintenance (Probe II)

The ability to concurrently hold multiple task-specific WM items in an elevated active state is of general importance to WM function. We thus compare our models on their Delay maintenance performance (Fig.2, Probe II).

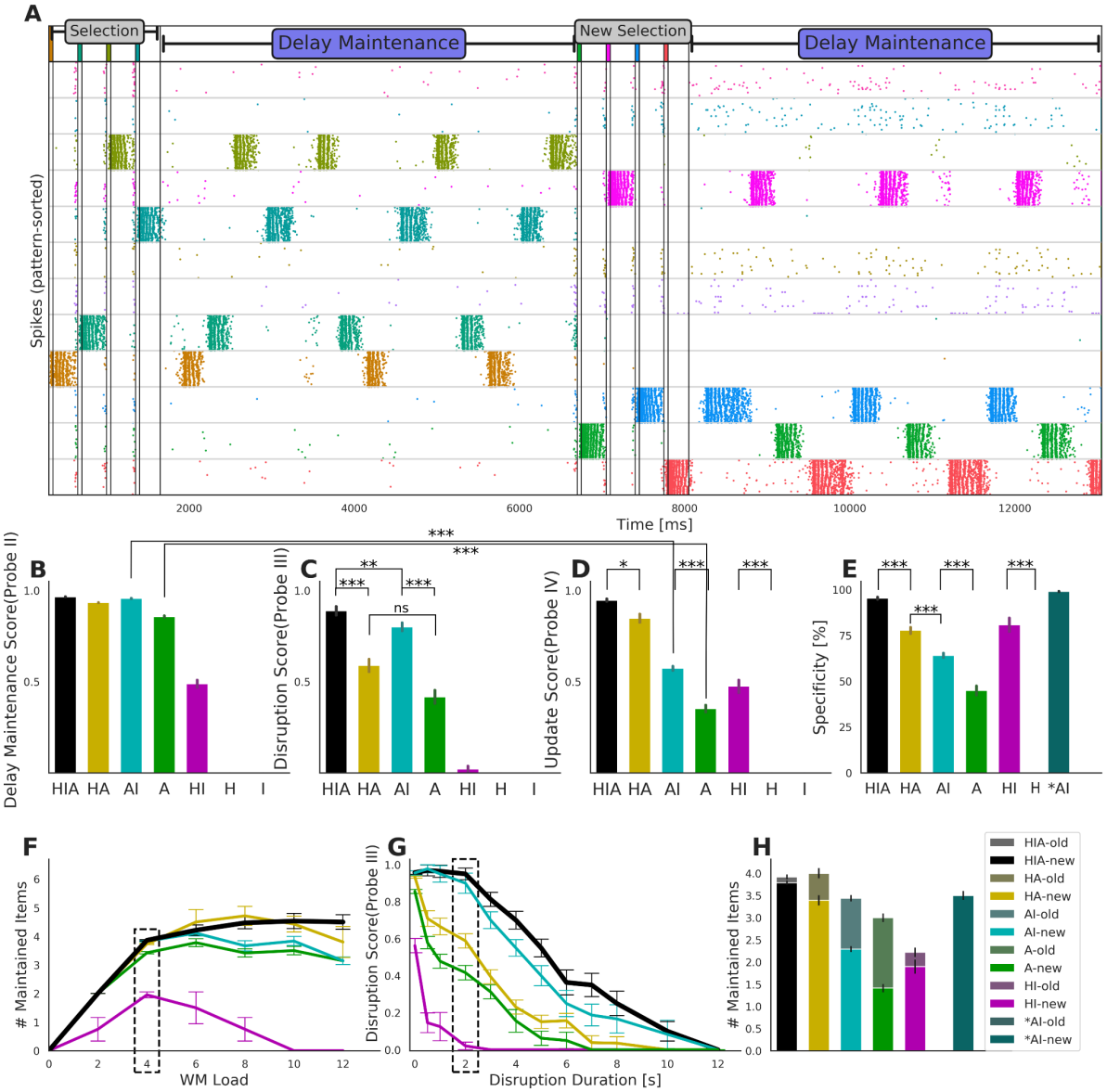

Figure 4. WM maintenance performance following WM item selection, disruption, and updating.

A: Single trial of HIA model, sorted spike raster of WM item selection (4 items), subsequent 5 s Delay maintenance, New WM item selection (updating the WM set), followed by another period of delay maintenance.

B: Delay Maintenance Score (Probe II) following WM item selection of 4 items for each of the 7 plasticity scenarios in the multi-trial average.**

C: Disruption Score (Probe III) following WM item selection and 2 s of Maintenance disruption for each of the 7 models across trials.**

D: Update Score (Probe IV) after New WM item selection for each of the 7 plasticity scenarios in the multi-trial average.**

E: Maintenance Specificity after New WM item selection in the multi-trial average.**

F: Load Saturation of the Delay maintenance. Datapoints mark the mean number of actively maintained items across trials at a given load in the multi-trial average.** The dashed marking annotates the 4-item-load default scored in B (relative to the load).

G: Disruption Score over increasing duration of task disruption in all functioning models in the multi-trial average.** Disruption of 0 s denotes the control condition of immediate free recall (same data as **B**). The dashed marking annotates the 2 s disruption data (same data as **C**).

H: Total number of items maintained during Delay maintenance after an update of the WM Set via New WM item selection, showing the multi-trial average**. 2-tone shading indicates items from the new WM set (colored according to the scenario), and items still maintained from the previous (old) WM set (grayed colors). *AI denotes an artificial hypothesis test based on the AI model (see next section)

***All error bars indicate sem.*

As we can see in the single trial example of Fig.4A, the HIA model can maintain all 4 initially selected task items through a short series of bursts over a delay interval of 5 s. The duration of these free WM item reactivations is limited by neural adaptation and synaptic depression. The multi-trial averaged *Delay Maintenance Score* at Probe II (Fig.4B) reveals that all models with synaptic augmentation (HIA,HA,AI,A) are able to reliably maintain most of the 4 selected items over the delay period, whilst the HI model often just maintains 2 of the 4 items, leading to a lower performance score. The H, and I models fail to maintain any WM set items (with the given model parameters), and disrupting Delay maintenance for a mere 2 s already incurs a dramatic loss of that limited maintenance performance (Fig.4C). We explore this in more detail in the section *Maintenance disruption*. Different models also show different maximum memory capacity during Delay maintenance, as we can see in Fig.4E. Overloading active maintenance by an excessive WM load may even decrease memory capacity.

Updating and maintenance (Probe IV)

Beyond the general ability to maintain a limited number of rapidly selected task relevant memory items over some delay period, WM also needs to flexibly update its contents [62,63]. We simulated this aspect of WM functionality with the New WM item selection (Fig.4D). The subsequent maintenance performance, quantified with the Update Score, reveals considerable differences between the models. While the HIA and HA models score high again, both the A (0.86 vs 0.35, $p<1e-30$, Mann-Whitney test) and AI models (0.96 vs 0.57, $p<1e-48$, Mann-Whitney test) drop their performance compared to the previous Delay Maintenance Score (cp. Fig.4D vs. Fig.4B). What do these performance differences tell us about the role of different plasticity mechanisms for flexible WM? It is helpful to break down the maintained items into 2 subgroups: those from the old WM set and those that should be maintained after the update with a new WM set (Fig.4H). We can see that the AI and A models (non-Hebbian) actively maintain 3 to 4 items, but often many of these maintained items actually belong to the old WM set. We measure this effect with the Specificity of maintenance (Fig.4E), the fraction of all reactivations during the delay period that belong to the new WM set. Apparently, maintenance may fail to fully shift to the new WM set, particularly in models without Hebbian plasticity. To show the network origins of this low Specificity of maintenance, particularly in non-Hebbian models, we examine the effects of plasticity after the update.

Since Hebbian plasticity is potentially strengthening the connections encoding individual items during New WM item selection, a first tangible hypothesis may be that items from the new set are more strongly encoded in the better performing Hebbian models (i.e. HA and HIA). Are new items simply more strongly encoded through increased recurrent excitation than old items in Hebbian models? To assess this, we can compute an effective “strength of the encoding” by averaging the contributions of all AMPA and NMDA synapses recurrently connecting pyramidal cells encoding the same item, filtering for items in the new WM set, and then performing the same analysis for items in the old WM set. It turns out, the ratio of relative recurrent excitation strength is generally very close to 1 in the multi-trial average (Fig.5-left), meaning new set items are not much stronger encoded than those in the old set. In particular, there is no obvious correlation between this strength ratio, and increased maintenance Specificity across trials and plasticity scenarios.

Another aspect to consider is the beneficial role of intrinsic plasticity for the memory performance - models with intrinsic plasticity score higher (Fig.4D, compare H vs HI: 0.00 vs 0.47, p<1e-6; A vs AI: 0.35 vs 0.57, p<1e-10; and HA vs HIA: 0.85 vs 0.95, p=0.0194, Mann-Whitney test) and are more specific in their maintenance (Fig. 4E, compare H vs HI: 0.00 vs 0.81, p<1e-8; A vs AI: 0.45 vs 0.64, p<1e-6; and HA vs HIA: 0.78 vs 0.95, p<1e-6, Mann-Whitney test). The mechanism can learn a cellular bias towards the most recently active items by boosting the intrinsic excitability of cells coding for new (i.e. recent) items. This is evident in the multi-trial average (Fig.5-middle), but still fails to explain why the HA model (which has fixed intrinsic excitability currents that are thus equal over the two sets) is significantly more specific in its maintenance than the AI model (0.78 vs. 0.64, p<1e-7, Mann-Whitney test). There is no obviously strong correlation between the stronger intrinsic excitability of new set items and higher Specificity during maintenance.

Yet one more possibility is that items in the new set actually suppress items from the old set in Hebbian scenarios, as the associative rule can learn disynaptic inhibition (see Methods). We can quantify this by averaging the learned connection strength from all neurons selective for items in the new set to all neurons selective for items in the old set (actually the learned peak conductances, see Methods). Indeed there is a large and significant difference between non-Hebbian and Hebbian models (average across models and samples: -1.41 vs -4.10, p<1e-49, Mann-Whitney test) and this targeted inhibition is negatively correlated with Specificity of the Delay maintenance (r=-0.60 (p<1e-30) across all models and simulations, see Fig.5-right). If we accept this as a hypothesis for why Hebbian models have higher Specificity, it still needs to be explained how this strong inhibition comes about and to demonstrate the causality of this effect on Specificity (and the Update Score).

The reason why the Hebbian model can learn such strong specific inhibition has to do with the temporal task structure. New WM item selection is a massed (rather than spaced) learning event, meaning the new items are activated in close temporal proximity (rapidly cued right after each other), leading to some amount of synaptic trace overlap (particularly with respect to the slow NMDA receptors), whilst spiking activity of the old WM set is temporally distant, and in fact anti-correlated from new item activity. The BCPNN learning rule learns excitation from spiking activity correlations and inhibition from anti-correlations (see *Methods - Spike-based BCPNN Learning Rule*). In our Hebbian models, this means that synaptic plasticity strengthens excitatory cell assemblies binding coactive cells (sharing the same item selectivity) and that neurons selective for items in the same set are also weakly exciting each other. However BCPNN also learns inhibitory connections between neurons with anti-correlated activity, and the neurons with the most anti-correlated activity are selective for the other WM set items. One way we can see the spiking activity outcome of this fast Hebbian learning process is exemplified in Fig 3A after the updating the WM Set: Cells belonging to items in the old set have much lower background spiking activity than the cells selective for the new set items, or even the cells selective for the 4 completely unused (i.e. uncorrelated) items.

If this additional specific learned inhibition is causal (and sufficient) to the high *Specificity* in Hebbian models then synthetically adding only this targeted inhibition to an AI model should raise its *Specificity* to that of the complete model. To test this we identify the specific connections between cells selective for new set items and cells selective for old set items.

We then sample the inhibitory weight values from such connections in the HIA model to set these particular connections in an AI model, without enabling Hebbian plasticity, and leaving everything else the same.

The result of this “transplantation” of structured inhibition dramatically boosts model *Specificity* to 97%, shown as the dark cyan **AI* model in Fig. 4 and 5. The targeted inhibition effectively eliminates old set items from Delay maintenance (see Fig. 4H). As new set items are not competing for reactivation with old items now, this results in an improved *Update Score* of 92%. This demonstrates the value of fast Hebbian plasticity in WM models, not only for rapid novel item encoding (Fig.3), but also for item set switching, as it occurs naturally in experimental WM tasks that include multi-item maintenance and repeated trials with different items.

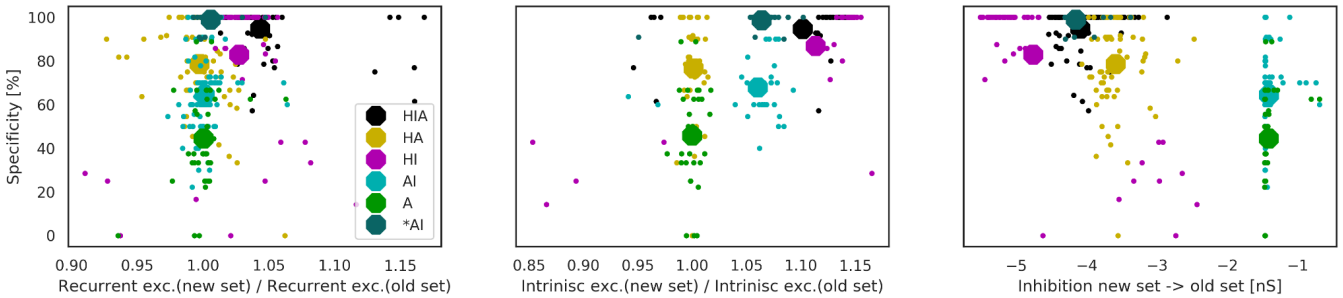

Figure 5. Trial-wise scatter plots of maintenance Specificity over three different metrics of learned network connectivity after WM updating. Each dot corresponds to an independent trial simulation. Its color marks the plasticity model used: HIA black, HA yellow, HI purple, AI cyan, A green, dark cyan denotes a special AI model with additional targeted inhibition. Larger octagons mark the multi-trial mean of trials with the same plasticity model. Left: Scatter over the ratio of recurrent excitation strength for cells selective for new and the old WM set items. Strength is based on the peak conductance transient between pyramidal cells sharing selectivity to the same item; the average is calculated over all recurrent connections of pyramidal cells selective for the same item and part of the same set). Middle: Scatter over the relative ratio of the average intrinsic excitability between items in the new WM set vs. items in the old WM set (averaged across all pyramidal cells selective for the same set). Right: Scatter over the average inhibition of cells selective for new set items on cells selective for old set items (measured as averaged across all pairwise connections).

Maintenance disruption (Probe III)

The level of WM disruption has been shown to depend on the type of external interference [64]. Due to the simplicity of our model, we confined our study to the case where interruption of the active maintenance of task relevant memory items in the delay period (replay of WM items) is operationalized as a transient reduction of background excitation to the spiking network to disable attractor activity (i.e. free, uncued recall; Fig.2 - Maintenance disruption) for some variable amount of time (.5 to 12 s). This enforces an activity-silent maintenance until background excitation is restored to normal levels.

We expect that a prolonged disruption should degrade the Delay maintenance performance of the model even after the end of the disruption, whereas short disruption should not have much of a lasting effect at all. This is because synaptic WM models generally do not require strictly persistent activity, and memory items are intermittently silent during Delay maintenance due to competition for foreground activity. A representative simulation trial can be seen in Fig. 6. After activating a WM set of 4 items, we can see active maintenance of all

items (*Delay Maintenance Score* of 1.0). Lowering the background excitation for 4 s disrupts this ongoing maintenance (see *Methods - Stimulation Protocol, Table 2*). When we raise the background excitation to its original value, only 2 memory items remain active during Delay maintenance leading to a *Disruption Score* of 0.5.

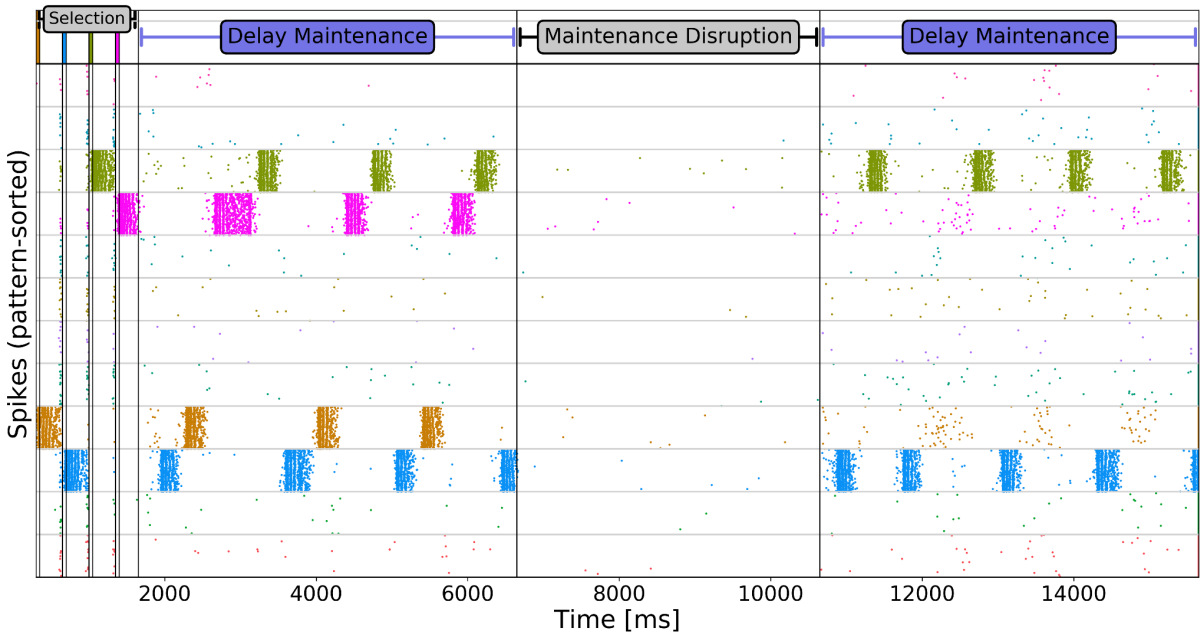

Figure 6 - HIA model, Sorted spike raster of Delay maintenance with intermittent Maintenance disruption. Following the selection of 4 items and their successful maintenance, we disrupt the task by 4 s of intermittent silence in the network caused by considerably lower level of background excitation (uncorrelated noise). When we raise the background excitation back to the previous level, 2 of the 4 items are no longer being maintained in the subsequent delay period.

All plasticity mechanisms are based on neural and synaptic traces which generally decay with some time constant over the quiet period. The overall time constants for each mechanism is the same (by design), so we expect that the model can maintain WM for about that time constant, 5 s, but we can articulate some further expectations:
Firstly, free reactivations depend on many factors but are a somewhat thresholded process, so the most highly performing models (HIA and AI) can probably tolerate the longest periods of disruption and low-performing model, such as HI, may collapse after much shorter disruption. So overall, we do not expect the performance rank order of models to change much with disruption length.
Secondly, there is also an interesting aspect of Hebbian plasticity that may cause us to expect it to play a lesser role than other mechanisms here: Hebbian plasticity (as BCPNN) learns from spike correlations, which are not as strongly disrupted by silence, as they would be by uncorrelated noise. Neurons that are silent together, are in-fact correlated in their (in)activity, meaning correlation-based associative connection weights (H-type models) are probably not affected by silence as much as the more broadly activity-depended (A- and I-type models) that will react to the reduced pre- and post-synaptic activity, regardless of correlations. Because of the shared time-constant, we can expect a similar rate of performance loss from trace decay, in A- and I-type models then.

Fig.4G shows how much disruption each plasticity model can actually tolerate. The multi-trial averaged performance of each plasticity model degrades with the increasing duration of the

disruption in the WM delay period. The combined model reliably preserves half of the task-relevant items even across 5 s long periods of delay Maintenance disruption. Confirming our expectations, the rank-order of models does not really change much. Models with higher Delay Maintenance Score can also tolerate the longest disruption. The presence of Hebbian plasticity does not strongly alter the Disruption Score (cp. A vs. HA: 0.62 vs. 0.69, p=0.07; AI vs. HIA: 0.76 vs 0.78, p=0.003). The effect of intrinsic plasticity is much more significant(cp. A vs. AI: 0.69 vs 0.76, p<1e-4; HA cs. HIA: 0.62 vs 0.78, p<1e-7). Two-way ANOVA reveals that the length of the disruption and the presence of intrinsic plasticity are both highly significant, while there is no interaction between the two.

Matching Experimental EPSP Dynamics

We expect plausible WM models to be able to address a variety of WM tasks (as shown above). Because our model is primarily phenomenological we do not expect to make biophysical predictions, yet the model is detailed enough to invite attempts to link its electrophysiological properties to experimentally observable quantities, such as PSPs, that may help us identify further constraints on plausible plasticity models.

The BCPNN learning rule is a particular combination of Hebbian and Intrinsic plasticity and learns excitatory and inhibitory postsynaptic potentials (EPSP/IPSP) between co-active and competing item-selective neurons. There is a stark difference between the performance of the HI and the HIA model across most task configurations however, such as Delay maintenance (Fig.4B). While the difference in baseline synaptic weights of those two models is small (see first spike in Fig.7-right, and also Methods - Fig.M3), there is a big difference in the PSP dynamics during short spike bursts (often referred to as STP). The most common activity motif during maintenance is that of brief bursts of item-selective neurons with 3-7 spikes. Wang et al. [12] measured EPSP and their dynamics in prefrontal pyramidal cells with strong synaptic facilitation. We compare the dynamics of burst EPSPs (6 spikes at 20Hz) across different plasticity models with the biological record (Fig.7).

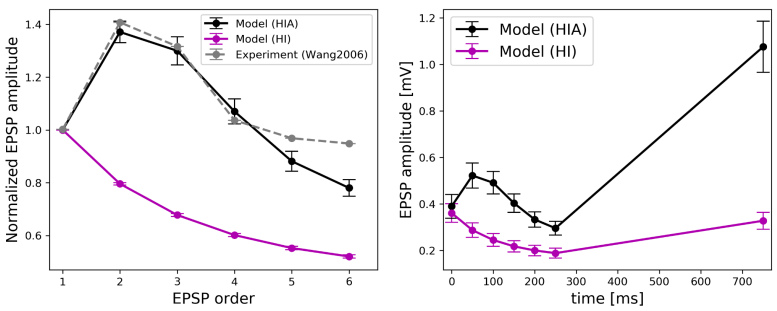

Figure 7 - Excitatory Postsynaptic Potentials (EPSPs). Left: Normalized mean EPSP amplitudes in a 6-spike burst at 20Hz in two of the models and the relevant recordings by Wang et al. 2006. **Right:** EPSP mean amplitudes during the same stimulation protocol and subsequent and recovery test amplitude after 500ms. Error bars denote the 95% confidence interval.

As we can see in Fig.7 (left), the HIA model can reproduce the known short-term dynamics of facilitating prefrontal pyramidal cell EPSPs in ways the HI model cannot. The shape of the STP dynamic is influenced by synaptic facilitation and depression. The Tsodyks-Markram model implementing the decay of transmitter is governed by the ratio of remaining transmitter depleted with each transmitted spike (U=0.2 in our model, with Wang et al. reporting similar estimates), and the depression time constant (τ_{rec}=280ms, as reported by Wang et al. (2006) for this class of synapses). Similar to the experimental record, augmenting synapses in the HIA model experience a roughly 40% increase in synaptic efficacy after the first spike, which then decays with the rapid depletion of transmitter across

the burst (**Fig.7 left**). Because the synaptic depression decays away after some time, a recovery test spike 500 ms later will evoke a much larger EPSP if the time constant of facilitation/augmentation is longer than the depression time constant. Accordingly, the HI model without facilitation/augmentation recovers about 90% of its EPSP amplitude over 500ms, while the efficacy of HIA synapses increases to 176% above baseline amplitude (.4 mV to 1.1 mV). This is the reason why all our models with augmentation (A, HA, AI, HIA) perform much better at free recall (including Delay maintenance, **see Fig.4B,C,D**) than non-augmenting models (the average performance score for models with augmentation vs models without augmentation: 0.93 vs 0.31, $p<1e-50$, Mann-Whitney test). Long lasting augmentation (5s), after the end of a burst leaves attractors more "explosive" for a future item reactivation without raising the baseline EPSP. The latter would make HI models much more functional, as shown by a published HI-type model with 4-times larger baseline EPSPs closer to 1.5mV [29].

Summary of key findings

In light of our results, **Hebbian plasticity** is a necessary component for novel encoding into an attractor WM model [29,37]. Even when the novelty is just an association of long-term memory items, Hebbian plasticity is necessary to bind these, and has been used by Hebbian models to account for fast WM [40,42]. Yet Hebbian plasticity alone is hardly sufficient. Under plausible biological constraints our own Hebbian-only (H) model achieves cued recall, but fails the majority of our Delay maintenance tasks in a straightforward manner from a general lack of excitability. Unpublished observations in the development of this model showed that the Hebbian only model can be restored to broader function by retuning the model, including specific violations of biological constraints, such as implausibly high background excitation, pyramidal firing rates larger than 30Hz, connection density (>20%) or strength (baseline pyramidal EPSPs>>1mV) outside the experimentally known distribution, stimulation durations exceeding typical WM tasks (larger than 1s), or more encoding repetitions (e.g. triple-shot encoding) enable active delay maintenance in the H model. A highly related HI-type model was previously shown to perform active delay maintenance with much longer stimuli which also resulted in larger EPSPs [29]). We decided to forgo any such specific tuning of models in this study, because the fragility of reduced models is itself a valuable observation.

Synaptic Augmentation is key to effective, and biologically plausible active (bursty) maintenance in the attractor WM model. The Tsodyks-Makram implementation acts as a dynamic multiplier of the effective (learned) recurrent connectivity. This enables one-shot learning from brief stimuli, as it boosts the efficacy of synapses. Free reactivations enable active delay maintenance in our model. Without synaptic facilitation/augmentation, baseline EPSPs would need to be 3-4 times larger to achieve similar recovery spike amplitudes after short bursts (Fig.7, recovery spike EPSP), which are important to trigger attractor activations from noisy background activity. This is the main reason why models without synaptic augmentation do not perform well. Augmentation marries known EPSP constraints (i.e. sub 1mV baseline EPSPs [12,65]) with attractor dynamics on a timescale relevant to WM, thus enabling a plausible attractor theory of WM in the first place. However, facilitation/augmentation by itself cannot encode novel items (Fig.3), is less robust against

task disruption (Fig.4G), lacks capacity (Fig.4F), and is imprecise with the updating of actively maintained WM sets (Fig.4E,H).

Intrinsic excitability generally helps with stability, *Specificity* of maintenance (Fig.4E), and significantly improves performance across all our tasks (Fig.4B,C,D cp. A vs AI, H vs. HI, HA vs HIA), including robustness against longer task disruptions (Fig.4G). Our implementation of intrinsic excitability as part of the BCPNN learning rule was originally designed to match experimentally observed graded increases in excitability with repeated stimulation [22,31], but as with all other forms of plasticity there are many possible implementations of intrinsic excitability changes beyond simple spike-triggered adaptation that are outside the scope of this study, but could be be explored in future work.

Discussion

We set out to explore the free combination of three different kinds of plasticity rules that have previously been proposed as WM mechanisms in their own right in a spiking neural network under loosely defined biological constraints, to evaluate whether there are specific synergies between them when applied to a broad battery of WM task requirements. After quickly finding that each additional mechanism contributes to the overall excitability of the network we normalized the general excitability such that all models would achieve cued recall of the encoded items (which also required the preloading of item encodings for non-Hebbian models). Even on that shared baseline we found and explored important task-specific differences in the contributions of each of these plasticity mechanisms, which all contribute to the superiority of the fully combined model across the breadth of multiple WM task motifs.

Modeling choices

Our model follows the attractor theory of memory – a powerful paradigm for neural network WM models with reverberatory dynamics [8]. Unlike popular bump attractor memory networks used to simulate visuospatial WM [52], our model operates in the space of discrete WM items, embodied by sets of cell assemblies wired by synaptic plasticity processes and activated upon WM recall or short-term maintenance. This modelling framework helps explain, though not exclusively, a large body of experimental findings at the crossroads of neuropsychological and cognitive neuroscience domains. To reflect this, our generic model task flow accounts for functional operations needed to execute many real-world working tasks in the spirit of goal-directed behaviour. The synergies we find in this regard are meaningful, because the combination of mechanisms allows us to address more than one task through a unified model, enabling an integrative rather than reductionist approach to addressing WM as a system with rich functionality. At the same time, it should be emphasised here that we do not account for the role of attention, executive control, short-term consolidation potentially supported by hippocampus or, from a broader perspective, the brainwide distributed nature of WM related phenomena [66]. Instead, our focus is on core short-term memory network computations tentatively located in the prefrontal cortex, without explicit support of other WM related circuits or subsystems.

Choosing mechanism implementations

We acknowledge that there are many possible ways to classify and combine plasticity mechanisms. Our division of mechanisms by their activity factors (pre-, post-, pre-post-synaptic) conforms with well established nomenclature. In contrast to many other WM models that combine plasticity mechanisms (see Introduction), we also decided to focus on combinations of mechanisms that have previously been proposed as short-term memory mechanisms in their own right. We picked one representative from each class of short-term memory plasticity mechanisms to limit the scope, but one might choose others (i.e. STDP for Hebbian learning), or a more detailed calcium model for the presynaptic plasticity. We chose Tsodyks-Markram as our mechanisms for implementing the latter because it is the most commonly used model for synaptic short-term memory [15]. For Hebbian plasticity, we opted for BCPNN over STDP because it already integrates Hebbian plasticity with intrinsic excitability in a principled way [31] and for its superior associative memory storage capacity [44]. Any mechanism of the full model can be removed without weight or activity instability, enabling us to focus on the performance contributions of the learning rules to different WM tasks instead.

Alternative Hebbian learning rules (such as STDP) may require additional plasticity mechanisms to achieve stable weight and activity dynamics and avoid significant top-down interventions (like weight cropping, renormalization, network resetting etc.), and would likely involve some form of inhibitory learning corresponding to the di-synaptic inhibition learning inherent to BCPNN. BCPNN resolves this by being stable on its own. Whether for stability, robustness, biological plausibility or performance, the obvious solution is to combine Hebbian plasticity with other forms of fast plasticity [46]. As our experiment with repeated WM item selection demonstrated, local Hebbian plasticity is also a candidate mechanism supporting some WM functions that are often hypothetically addressed at the systems level (see Rapid WM updating below). While Behavioral Time Scale Synaptic Plasticity (BTSP) is another associative learning rule that supports one-shot encoding and has significant experimental support, it is not strictly Hebbian, and has also been closely linked to hippocampal function, putting it outside the scope of this paper [67,68].

Preloading WM

Working memory is multimodal, and leverages existing long-term memory [69,70] and sensory representations [71]. Behavioural studies generally involve highly trained subjects/animals, which can be presumed to be highly familiar with the relevant stimuli and have stable representations. Many synaptic WM models [53] have used similar arguments to justify the preloading of the associative connectivity required for stable WM representations (cell assemblies/attractors). Hebbian WM models [37] provide an alternative to this preloading step by enabling the rapid formation of novel assemblies. Our Hebbian-type models explored here did not require this kind of preloading, as they are capable of one-shot learning novel representations. We still preloaded all non-Hebbian models though, as we believe that these views do not exclude each other. WM can predominantly operate on existing representations and also feature fast Hebbian plasticity, resulting in powerful models that may help address dynamic coding in WM. While we did not explore this aspect in depth here, previously published models have done so. By flexibly binding preexisting long-term memories, or selective populations through a dynamically recruited short-term memory (STM) population with flexible conjunctive cells [40,72], working memory implements an indexing theory [42].

Associated memories can be maintained or retrieved from the indexes bidirectional connections. The memories themselves are presumed to be encoded in pre-existing, distributed representations, that may well span many brain areas and multiple modalities. Because an index to existing representations does not need to encode a copy of the information, it acts more like a pointer, freeing up representational space for relevant task information, an aspect future models could explore in more detail.

Multiplexing WM in time

Multiplexing the maintenance of multiple items in time provides a natural explanation for the capacity limits of WM, which saturates at about 5 items in our model (see Fig.4F). There have been attempts to implement multi-item maintenance based on synaptic augmentation without multiplexing in time, but those results rely on dramatically implausible pyramidal firing rates around 100Hz to support the simultaneous persistent activation of multiple attractors in the same network [16]. The idea of multiplexing multi-item maintenance in time is supported by macaque studies showing load-dependent increases in gamma burst rates during the delay period [73], and predicted by related models [29]. Attractor dwell times emerge from the time-constants of synaptic depression and spike-frequency adaptation, and because there is a limited number of items that can be refreshed before memory traces decay too much, the WM overloading effects seen in our model (Fig.4F) are somewhat related to computational models based on the Time-Based Resource-Sharing theory [74].

Rapid WM updating

Historically, many WM tasks, such as a delayed-match-to-sample, have been simplified for replicability in the research setting, but thus also no longer require complex WM; a simple palimpsest short-term memory model will often suffice. Our study demonstrates the importance of rapidly learned (i.e. structured) di-synaptic inhibition for WM tasks that require the rapid displacement of memory items. More generally, rapid task switching and memory operations in complex WM tasks require more control over maintenance, such as the active suppression of old content. Experimental evidence supports an integrated, simultaneous removal and encoding process, separate from passive forgetting that limits working memory load and thus facilitates the maintenance of goal-relevant information by improving access to the remaining goal-relevant content [62,63]. Our model captures this dynamic. While rapid updating is often addressed at the systems level as executive control (and we do not claim to address this here), our results indicate that known plasticity mechanisms can support such memory operations. Some Hebbian plasticity mechanisms (such as BCPNN) simulate direct effects of inhibitory plasticity that can help address the need for control without resorting to systems-level processes.

Limitations

Our model is first and foremost a phenomenological model of memory dynamics and follows an attractor theory of memory with distributed neuronal assemblies for memory representations. Importantly, it lacks experimental verification down to the biophysical reality at the synaptic level. For a more realistic model, less constrained statistics of stimuli (overlapping or correlated random stimuli, spanning different coding levels, and complex protocols of presentation) should be considered.

Hebbian learning may be symmetric or not, which affects sequentiality in the activity [31], and may have important implications for WM function, particularly on tasks that involve sequential information. We did not address this here, but the issue was rigorously studied by [75]. We also did not look into the rich oscillatory dynamics of this model, which have been strongly associated with WM function in monkey experiments [73]. They largely emerge from our modular cortical architecture and related models studied this in more detail [17].

Conclusion

Unlike what past reductionist models suggest, biology is likely to utilize many WM mechanisms in parallel. Our study suggests that commonly used forms of plasticity proposed for the buffering of WM information besides persistent activity are eminently compatible, and yield synergies that improve function and biological plausibility in a combined modular spiking neural network model of cortical architecture, tasked with WM. Combinations enable a more holistic model of WM responsive to broader task demands. Conversely, the targeted ablation of specific plasticity components reveals that different mechanisms are differentially important to specific aspects of WM function, advancing the search for more capable, robust and flexible models accounting for new experimental evidence of bursty and activity-silent multi-item maintenance. Importantly, we attribute the observable differences to the principle nature of specific types of plasticity. For example, we find a previously undescribed synergistic function of Hebbian plasticity and synaptic facilitation/augmentation that supports the rapid updating of multi-item WM sets through rapidly learned inhibition, that shifts active multi-item delay maintenance more effectively towards task-relevant representations than previously published spiking models. Our computational model simultaneously captures relevant electrophysiological, cognitive and behavioural effects, even though the synaptic plasticity rule is largely phenomenological. Further progress in understanding would benefit greatly from more quantitative experimental data on the dynamics of synaptic plasticity.

Methods

Neuron Model

We use an integrate-and-fire point neuron model with spike-frequency adaptation [76]which was modified by [31] for compatibility with a custom-made BCPNN synapse model in NEST (see *Simulation Environment*) through the addition of the intrinsic excitability current I . The model was simplified by excluding the subthreshold adaptation dynamics. Membrane potential V_m and adaptation current are described by the following equations:

$$C_m \frac{dV_m}{dt} = - g_L\left(V_m - E_L\right) + g_L \Delta_T \frac{e^{\frac{V_m - V_t}{\Delta_T}}}{} - I_w(t) + I_\beta(t) + I_{syn}(t) + I_{ext} \qquad (1)$$

$$\frac{dI_w(t)}{dt} = \frac{-I_w(t)}{\tau_w} + b\delta(t - t_{sp}) \quad (2)$$

The membrane voltage changes through incoming currents over the membrane capacitance C_m. A leak reversal potential E_L drives a leak current through the conductance g_L, and an

upstroke slope factor Δ_T determines the sharpness of the spike threshold V_t. Spikes are followed by a reset of membrane potential to V_r. Each spike increments the adaptation current I_w by b, which decays with time constant τ_w . I_β is another intrinsic excitability current shaped by the Spike-based BCPNN Learning rule (see below). Simulated basket cells feature neither the intrinsic excitability currents I_β nor the the spike-triggered adaptation current I_w.

Besides external input I_{ext} (*Stimulation Protocol*) neurons receive a number of different synaptic currents from their presynaptic neurons in the network (AMPA, NMDA and GABA), which are summed at the membrane accordingly:

$$I\ (t) = \sum_{syn}\sum_{i} g_{ij}^{syn}(t)\left(V\ - E_{ij}^{syn}\right) = I_j^{AMPA}(t)\ + I_j^{NMDA}(t)\ + I_j^{GABA}(t) \qquad (3)$$

Synapse Model and Augmentation Mechanism

Excitatory AMPA and NMDA synapses have a reversal potential $E^{AMPA} = E^{NMDA}$, while inhibitory synapses drive the membrane potential toward E^{GABA}. Every presynaptic input spike (at t_{sp}^i with transmission delay t_{ij}) evokes a transient synaptic current through a change in synaptic conductance that follows an exponential decay with time constants τ^{syn} depending on the synapse type ($\tau^{AMPA} \ll \tau^{NMDA}$).

$$g_{ij}^{syn}(t) = u_{ij}^{fac}(t)x_{ij}^{dep}(t)w_{ij}^{syn}e^{\frac{-t-t^i-t_{ij}}{\tau^{syn}}}H(t\ - t_{sp}^i\ - t_{ij}) \qquad (4)$$

$H(\cdot)$ is the Heaviside step function. w_{ij}^{syn} is the peak amplitude of the conductance transient, learned by the *Spike-based BCPNN Learning Rule* (next Section). Plastic synapses are also subject to STP according to the Tsodyks-Markram formalism [77], modeling the transmission-dependent synaptic facilitation/adaptation u_{ij}^{fac} and the depletion of available synaptic resources x_{ij}^{dep} with a depression/reuptake time constant $\tau_{rec}$and a facilitation time constant τ_{fac}:

$$\frac{dx_{ij}^{dep}}{dt} = \frac{1-x_{ij}^{dep}}{\tau_{rec}} - u_{ij}^{fac}x_{ij}^{dep}\sum\delta(t\ - t_{sp}^i\ - t_{ij}) \qquad (5)$$

$$\frac{du_{ij}^{fac}}{dt} = \frac{U-u_{ij}^{fac}}{\tau_{fac}} - U(1\ - u_{ij}^{fac})\sum\delta(t\ - t_{sp}^i\ - t_{ij}) \qquad (6)$$

Spike-based BCPNN Learning Rule (Hebbian and intrinsic plasticity)

Plastic AMPA and NMDA synapses are modeled to mimic NMDA-dependent Hebbian Short-term potentiation [33,35] with a spike-based version of the Bayesian Confidence Propagation Neural Network (BCPNN) learning rule [31,78]. For a full derivation from Bayes

rule, deeper biological motivation, and proof of concept, see Tully et al. (2014) and an earlier memory model implementation [29].

Briefly, the BCPNN learning rule makes use of biophysically plausible local traces to estimate normalized pre- and postsynaptic firing rates, as well as co-activation, which can be combined to implement Bayesian inference because connection strengths and neural unit activations have a statistical interpretation [31,43]. Crucial parameters include the synaptic activation trace Z, which is computed from spike trains via pre- and postsynaptic time constants τ^{syn}, which are the same here, but differ between AMPA and NMDA synapses: $\tau^{AMPA} = 5ms, \ \tau^{NMDA} = 100ms$

The larger NMDA time constant reflects the slower closing dynamics of NMDA-receptor gated channels. All excitatory connections are drawn as AMPA and NMDA pairs, such that they feature both components. Further filtering of the z-traces leads to rapidly expressing memory traces (referred to as p-traces) that estimate activation and co-activation:

$$\tau_p \frac{dp_i}{dt} = \kappa\left(z_i - p_i\right), \qquad \tau_p \frac{dp_j}{dt} = \kappa\left(z_j - p_j\right), \qquad \tau_p \frac{dp_{ij}}{dt} = \kappa\left(z_i z_j - p_{ij}\right) \quad (7)$$

These traces maintain moving averages that support a palimpsest memory function. Short-term potentiation decay is known to take place on timescales that are highly variable and activity dependent [79].

The learning rule parameter κ (**Eq.7**), may reflect the action of endogenous neuromodulators, e.g. dopamine acting on D1 receptors that signal relevance and thus modulates learning efficacy. It can be used to switch off learning to fixate the network, or temporarily increase plasticity. In general, we trigger BCPNN plasticity concurrent with external stimulation, but switch it off during memory assessment, so that ongoing plasticity does not interfere with estimation of memory performance [80].

Tully et al. [31] showed that Bayesian inference can be recast and implemented in a network using the spike-based BCPNN learning rule. Prior activation levels are realized as an intrinsic excitability of each postsynaptic neuron, which constitutes a form of intrinsic plasticity, as these are derived from the postsynaptic firing rate estimate p_j and implemented in the NEST neural simulator [81] as an individual neural current I_β with scaling constant β_{gain}

$$I_\beta = \beta_{gain} log(p_j) \quad (8)$$

I_β is thus an activity-dependent intrinsic membrane current to the neurons, similar to the A-type potassium channel [82] or TRP channel [83]. Intrinsic plasticity can be deactivated by inhibiting learning of p_j at the neuronal level, which freezes the intrinsic membrane current at its current value.

Synaptic weights are modeled as peak amplitudes of the conductance transient (**Equation 4**) and determined from the logarithmic BCPNN weight, as derived from the p-traces with a synaptic scaling constant w^{syn}_{gain}.

$$w^{syn}_{ij} = w^{syn}_{gain} log \frac{p_{ij}}{p_i p_j} \qquad (9)$$

In this model, AMPA and NMDA synapses make use of w^{AMPA}_{gain} and w^{NMDA}_{gain} respectively. The logarithm in Equations 8,9 is motivated by the Bayesian underpinnings of the learning rule,

and means that synaptic weights w_{ij}^{syn} multiplex both the learning of excitatory and di-synaptic inhibitory interaction. The positive weight component is here interpreted as the conductance of a monosynaptic excitatory pyramidal to pyramidal synapse (Fig.M2, plastic connection to the co-activated MC), while the negative component (Fig.M2, plastic connection to the competing MC) is interpreted as di-synaptic via a dendritic targeting and vertically projecting inhibitory interneuron like a double bouquet and/or bipolar cell [84–86]. Accordingly, BCPNN connections with a negative weight use a GABAergic reversal potential instead, as in previously published models of this kind [29,31,87]. Model networks with negative synaptic weights have been shown to be functionally equivalent to those with both excitatory and inhibitory neurons with only positive weights [88]. In the context of this particular model microcircuit and learning rule, this was explicitly and conclusively demonstrated by the addition of double bouquet cells [89].

The most relevant parameters are found in *Tables 1-3*. Code for the NEST implementation of the BCPNN synapse is openly available (see *Code Accessibility*).

Network Architecture and WM Neural Representations

The model organizes cells into a grid of nested hypercolumns (HCs) and minicolumns (MCs), sometimes referred to as macrocolumns, and "functional columns" respectively. The Network is simulated with 16 HCs spread out on a rectangular grid. This spatially distributed network of columns has small conduction delays due to the distance between columns. Each of the non-overlapping HCs has a diameter of about 500 μm, comparable to estimates of cortical column size [90], contains 24 basket cells, and its pyramidal cell population has been divided into 12 MCs. This constitutes a sub-sampling from the roughly 100 MC per HC when mapping the model to biological cortex. We simulate n_{MC}^{PYR} pyramidal neurons per MC to represent roughly the layer 2/3 population of an MC.

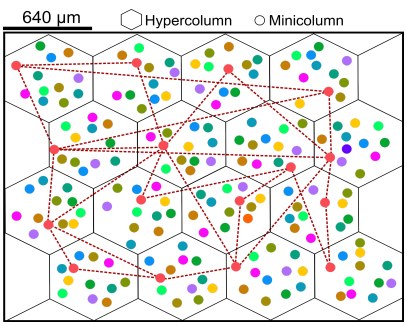

Figure M1 - Basic modular network layout. The network is comprised of 16 hypercolumns (HCs), representing a small patch of neocortex. Each HC contains 12 minicolumns (MCs), which are composed of 30 pyramidal cells each, and preferentially active for 1 of the 12 orthogonal neuronal assemblies, as indicated by the 12 colors. For one neuronal assembly we also indicate the learning induced sparse excitatory long-range connections between subsampled pyramidal cells of similar selectivity

The distributed population code for each memory items is defined by a set of activated minicolumns across the networks hypercolumns (sparse due to WTA within each hypercolumn). This code is orthogonal, as we choose one active minicolumn in each hypercolumn to be selective for any given memory item without overlap between items (see **Fig.M1**). The embedding of this code is achieved either via Hebbian learning during one shot item encoding or associative weights are preloaded at the start of the trial (see **Fig.2,** and **Methods-Preload**).

Network Connectivity

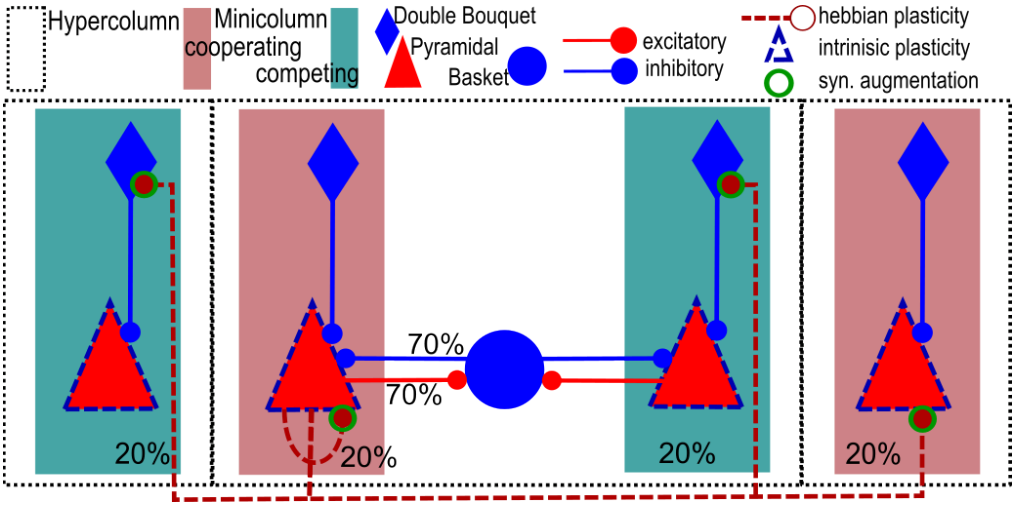

Figure M2 - Columnar microcircuit and plasticity. Pyramidal cells connect recurrently within each MC, and share a pool of inhibitory basket cells within each HC, providing local feedback-inhibition to all pyramidal cells within the HC. Pyramidals also connect across HCs to other coactive or competing MCs. The strength of plastic connection (dashed in red) develops according to the Bayesian-Hebbian plasticity rule (BCPNN). Synaptic efficacy is modulated by facilitation/augmentation (dashed green), and the excitability of each pyramidal further influenced by the presence of intrinsic plasticity at each pyramidal cell (dashed blue). Connection probabilities are shown as percentages (see Methods-Network Connectivity). Note that di-synaptic inhibition via DBC cells is not modeled directly, but instead expressed through the duplexing of excitatory and inhibitory learning in the BCPNN learning rule.

Pyramidal cells project laterally to basket cells within their own HC via AMPA-mediated excitatory projections with a connection probability of p_{P-B}, i.e. connections are randomly drawn without duplicates until the target fraction of all possible pre-post connections exist. In turn, they receive GABAergic feedback inhibition from basket cells (p_{B-P}) that connect via static inhibitory synapses rather than plastic BCPNN synapses. This strong loop implements a competitive soft-WTA subnetwork within each HC [91]. Local basket cells fire in rapid bursts, and induce alpha/beta oscillations in the absence of attractor activity and gamma, when attractors are present and active [92]. Pyramidal cells form connections both within and across HCs at connection probability p_{pyr}. These projections are implemented with plastic synapses and contain both AMPA and NMDA components, as explained in subsection *Spike-based BCPNN Learning Rule.* Because BCPNN duplexes excitatory and inhibitory learning, inhibitory weights are expressed through a GABA receptor onto pyramidal cells, which indirectly models the existence of local double-bouquet cells, providing disynaptic inhibition onto pyramidal cells in their local columns. The existence of such local specificity of double-bouquet cell mediated inhibition was shown by DeFelipe et al. [93] and the equivalency of that indirect modeling approach confirmed by Chrysanthidis et al. [89].

In summary, the model thus features a total of 6.64 million plastic connections between its 5,760 simulated pyramidal cells, as well as 193,000 static AMPA- and GABA-mediated connections to and from 384 simulated basket cells and 192 double bouquet cells (virtual, one per MC), targeting all the pyramidal cells within their respective MC.

Stimulation Protocol

The term I_{ext} in Equation 1 subsumes specific and unspecific external inputs. To simulate unspecific input from non-simulated columns, and other areas, pyramidal cells are

continually stimulated with a zero mean noise background throughout the simulation. In each layer, 2 independent Poisson sources generate spikes at rate r_{bg} , and connect onto all pyramidal neurons in that layer, via non-depressing conductances $\pm g_{bg}$. Before each simulation, we distribute the initial values of all plastic weights by a process of learning from 1000 ms of this unstructured background activity. A lower background rate (r_{bg-low}) is used to disrupt active maintenance (Fig.4). During One-shot encoding we stimulate all 12 items in turn, such that Hebbian plasticity may encode the correlated firing structure of each neuronal assembly. To do this, we excite pyramidal cells belonging to component MCs of each distributed representation with an additional targeted excitatory Poisson spike train (rate r_{stim} , length t_{stim}, conductance g_{stim}). This is followed by an inter stimulus period of the same length with lowered noise background rate r_{bg-low}, followed by the stimulation of the next item, until all memory items have been stimulated once. To later cue the reactivation of any specific memory item (i.e. neuronal assembly/attractor) during Cued recall testing or WM item selection (Fig.4), we excite pyramidal cells belonging to a component MC with a brief excitatory Poisson spike train (rate r_{cue}, length t_{cue}, conductance g_{stim}). We use partial cues defined by leaving out one of the coding minicolumns, which allows us to check for the associative completion of each activation as well. A brief 50 ms stimulus is usually sufficient to activate any previously encoded memory.

Axonal Conduction Delays

We compute axonal delays t_{ij} between presynaptic neuron i and postsynaptic neuron j, based on a constant conduction velocity V and the Euclidean distance between respective columns. Conduction delays were randomly drawn from a normal distribution with mean according to the connection distance divided by conduction speed and with a relative standard deviation of 30% of the mean in order to account for individual arborization differences, and varying conduction speed as a result of axonal thickness/myelination. Further, we add a minimal conduction delay t_{min}^{syn} of 1.5 ms to reflect not directly modeled delays, such as diffusion of transmitter over the synaptic cleft, dendritic branching, thickness of the cortical sheet, and the spatial extent of columns:

$$\overline{t_{ij}} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{V} + t_{min}^{syn}\, ms \qquad t_{ij} \sim N(\overline{t_{ij}},\ .3\overline{t_{ij}}) \qquad (10)$$

Memory Performance Tracking

We track memory activity in time by analyzing the population firing rate of item-specific and network-wide spiking activity usually using an exponential moving average filter time-constant of 25 ms. Because the population code is orthogonal, and reactivations characterized by sizable gamma-like bursts, a simple threshold detector can extract candidate reactivation events and decode the activated memory. We further check for the complete activation of each assembly. Whenever targeted stimuli are used, we analyze peri-stimulus activation traces. When reactivation onsets are less predictable, such as during

Delay maintenance, we extract reactivation candidates via a threshold detector trained at the 50th percentile of the cumulative distribution of the population firing rate signal.

During Cued recall testing (Fig.4, Probe I), memory items are considered retrieved, if they fully activate within 175 ms following the cue for that item. The *Cued Recall Score* is then calculated as the ratio between the number of successfully retrieved items, and the total number of items cued (typically 12). During WM Item Selection and New WM Item Selection we cue the activation of a select working memory set (typically 4 randomly selected items). New WM Item Selection excludes preciously selected patterns. During Delay maintenance (Fig.4, Probe II) there are no cues provided as items reactivate freely. Items are scored as successfully maintained if they activate at least twice. The *Delay Maintenance Score* is then calculated as the ratio between the number of successfully retrieved items, and the working memory load (typically 4). The *Disruption Score* is computed the same way, except for the fact that the delay period only starts after a intermittent silence of variable length (.5 - 12 s). The *Update Score* is computed by tracking the activity of items in the updated working memory set (typically 4) and then scored the same way. Notably, this excludes old working memory items from the first set. Notably, we deactivate plasticity during the assessment of memory by setting $\tau_{fac} = 0$ (**Eq.6**) and κ=0 (**Eq.7**). This is not because the model is unstable (BCPNN is stable, even with additional plasticity mechanisms), but because ongoing plasticity introduces a bias towards the memories, tested earlier, making it impossible to assess the state of the memory network without influencing subsequent task motifs. For example, if we do a cued recall test on any memory item with ongoing plasticity, we are more likely to see subsequent reactivations of that item if we then also test delay maintenance. By forgoing plasticity during memory assessment we can thus score multiple task motifs in a single continuous simulation run (Fig.4).

Preload

Instead of training the network before each memory task, we may save time by preloading the network, which means sampling the values of plastic state variables from an already trained network. Each plasticity rules affects some of the network state variables (such as u, x, w, p-traces, etc.). As we discovered early on, incomplete models may fail to achieve basic cued recall (see Results Fig.3 and Methods Fig.M3-D), because the inactivation of a mechanism freezes inappropriate initialization values (Fig.M3 A-C), lowering overall excitability. Rather than re-tuning background excitation for each model, preload offers a means to compensate for this by loading appropriate values. In order to capture the variability of the learned distribution of such network state variables in dynamic spiking neural networks, we can even preload state variables from learned and stored distributions of these parameters. If a model does not have the specific mechanism, the affected variables will obviously be frozen after the preload.

Intrinsic plasticity for example affects neuronal excitability through a change in a neuron-specific membrane current I_β(Eq.1), which is itself dependent on the learned neuronal p-trace (Eq.8). If we disable intrinsic plasticity in any model, this means freezing the strength of this current, such that it no longer changes with activity but what value to freeze the underlying p-trace at, is not self-evident. So in the example of the HA model without intrinsic plasticity we preload the underlying distribution of neural p-traces from the full HIA

model, which then determine a distribution of bias currents I_{β} (**Eq.8**). We then set a bias gain $\beta_{gain} = 35pA$, that enables cued recall of all items. Similarly, non-Hebbian models are preloaded with frozen attractor weights (no synaptic gain chain is necessary to achieve cued recall here), and models without augmentation have their synaptic gains (w_{gain}^{AMPA} and w_{gain}^{NMDA}) boosted by 50%, which allows all non-augmentation models to achieve cued recall (see **Tab.4**). This compensates the reduced models for the loss of overall excitability (**Fig.M3-E**), but not more than necessary for cued recall of all encoded items and thus enables a more interesting comparison of the 7 different plasticity combinations (**Fig.1**) across the more dynamic WM tasks thereafter (**Fig.2**).

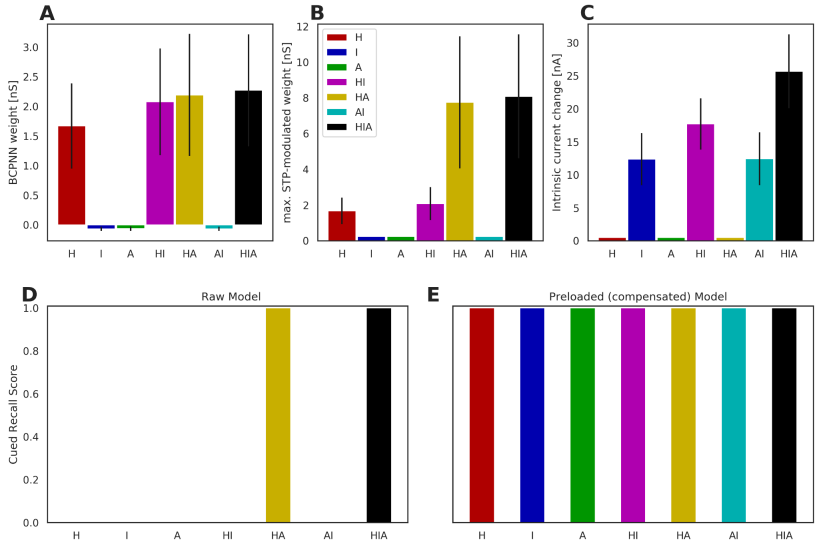

Figure M3. A: BCPNN weight of excitatory recurrent attractor connections (combined across AMPA and NMDA receptors) following 1-shot-encoding, averaged over all excitatory item assembly connections and trials. Error bars denote standard deviation over all excitatory item assembly connections and trials. **B:** STD-modulated maximal weight (max. amplitude of the conductance transient of excitatory recurrent attractor connections) including the multiplicative effect of augmentation) following 1-shot encoding of excitatory recurrent attractor connections, averaged over all excitatory item assembly connections and trials. Error bars denote standard deviation over the same. **C:** Absolute increase in intrinsic current (nA) of item-selective neurons following 1-shot encoding, averaged over all neurons in item assemblies and trials. Error bars denote standard deviation over all neurons in item assemblies and trials. **D:** Cued Recall Score (Probe I) prior to compensation through preload. No variability across at least 20 trials per scenario. **E:** Cued Recall Score (Probe I) after compensation through preload, constituting the baseline for all further model comparisons, averaged over at least 20 trials per scenario. No variability in *Cued Recall Score* detected, as all trials succeed in retrieving all tested items.

Code Accessibility

We use the NEST simulator [81] version 2.2 for our simulations, running on a Cray XC-40 Supercomputer of the PDC Centre for High Performance Computing. The custom-build spiking neural network implementation of the BCPNN learning rule for MPI-parallelized NEST is freely available on github: https://github.com/Florian-Fiebig/BCPNN-for-NEST222-MPI and the microcircuit model also available on ModelDB: modeldb.yale.edu/257610

Parameter Tables

Adaptation current	b	86 pA	Depression time constant	τ_{rec}	280 ms	BCPNN AMPA gain	w_{gain}^{AMPA}	0.766 nS
Adaptation time constant	τ_{w}	280 ms	AMPA synaptic time constant	τ^{AMPA}	5 ms	BCPNN NMDA gain	w_{gain}^{NMDA}	0.076 nS

Membrane Capacity	C_m	280 pF	NMDA synaptic time constant	τ^{NMDA}	100 ms	BCPNN bias current gain	β_{gain}	$40\ pA$
Leak Reversal Potential	E_L	-70 mV	GABA synaptic time constant	τ^{GABA}	5 ms	BCPNN lowest rate	f_{min}	0.2 Hz
Leak Conductance	g_L	28 pS	AMPA Reversal Potential	E^{AMPA}	0 mV	BCPNN highest rate	f_{max}	25 Hz
Upstroke slope factor	Δ_T	3 mV	NMDA Reversal Potential	E^{NMDA}	0 mV	BCPNN lowest probability	ε	0.008
Spike Threshold	V_t	-55 mV	GABA Reversal Potential	E^{GABA}	-75 mV	BCPNN Spike event duration	Δt	1 ms
Spike Reset Potential	V_r	-60 mV	Regular Plasticity	κ	1.0	P-Trace time constant	τ_p	5 s
Utilization factor	U	.2				Augmentation time constant	τ_{fac}	5 s

Table 1. Neurons, synapses, and plasticity

STM patch size	2 x 2 mm		Background activity rate	r_{bg}	1190Hz
Simulated HCs	n_{HC}	16	Low Background activity rate	r_{bg-low}	870Hz
Simulated MC per HC	n_{MC}	12	Background conductance	g_{bg}	$\pm$1.5 nS
Hypercolumn diameter	d_{HC}	0.5 mm	Targeted Stimulation rate	r_{stim}	350 Hz
Layer 23 pyramidal per MC	n_{MC}^{PYR}	30	Targeted Stimulus conductance	g_{stim}	+3.0 nS
Basket cells per MC	n_{MC}^{basket}	2	Targeted Stimulus duration	t_{stim}	200 ms
Axonal Conduction Speed	V	.2 $\frac{m}{s}$	Cue stimulus duration	t_{cue}	50 ms
Minimal conduction delay	t_{min}^{syn}	1.5 ms	Cue stimulation rate	r_{cue}	375 Hz

Table 2. Network size, Conduction delay, Stimulation

Source	**Target**	**Type**	**Symbol**	**Value**
Pyramidal	Basket	probability	p_{P-B}	0.7
Pyramidal	Basket	conductance (static)	g_{P-B}	+3. nS
Basket	Pyramidal	probability	p_{B-P}	0.7
Basket	Pyramidal	conductance (static)	g_{B-P}	-7 nS
Pyramidal	Pyramidal	probability	p_{pyr}	0.2
Pyramidal	Pyramidal	AMPA gain (BCPNN)	w_{gain}^{AMPA}	0.766nS
Pyramidal	Pyramidal	NMDA gain (BCPNN)	w_{gain}^{NMDA}	0.076nS

Table 3. Projections

1. Shafi, M. *et al.* Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* **146**, 1082–1108 (2007).
2. Lundqvist, M., Herman, P. & Miller, E. K. Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *J. Neurosci.* **38**, 7013–7019 (2018).
3. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
4. Kamiński, J. & Rutishauser, U. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Ann. N. Y. Acad. Sci.* **1464**, 64–75 (2020).

5. Oberauer, K. & Awh, E. Is There an Activity-silent Working Memory? *J. Cogn. Neurosci.* **34**, 2360–2374 (2022).
6. Manohar, S. Quiet Trajectories as Neural Building Blocks. *J. Cogn. Neurosci.* **35**, 14–16 (2022).
7. Beukers, A. O., Buschman, T. J., Cohen, J. D. & Norman, K. A. Is Activity Silent Working Memory Simply Episodic Memory? *Trends Cogn. Sci.* **25**, 284–293 (2021).
8. Barak, O. & Tsodyks, M. Working models of working memory. *Curr. Opin. Neurobiol.* **25**, 20–24 (2014).
9. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Neurocomputational models of working memory. *Nat. Neurosci.* **3 Suppl**, 1184–1191 (2000).
10. Abbott, L. F. & Nelson, S. B. Synaptic plasticity: taming the beast. *Nat. Neurosci.* **3 Suppl**, 1178–1183 (2000).
11. Thomson, A. M. Facilitation, augmentation and potentiation at central synapses. *Trends Neurosci.* **23**, 305–312 (2000).
12. Wang, Y. *et al.* Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
13. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
14. Hempel, C. M., Hartman, K. H., Wang, X. J., Turrigiano, G. G. & Nelson, S. B. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* **83**, 3031–3041 (2000).
15. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
16. Rolls, E. T., Dempere-Marco, L. & Deco, G. Holding multiple items in short term memory: a neural mechanism. *PLoS One* **8**, e61078–e61078 (2013).
17. Lundqvist, M., Herman, P. & Lansner, A. Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *J. Cogn. Neurosci.* **23**, 3008–3020 (2011).
18. Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput. Biol.* **15**, e1006928 (2019).
19. Hansel, D. & Mato, G. Short-term plasticity explains irregular persistent activity in working

memory tasks. *J. Neurosci.* **33**, 133–149 (2013).

20. Verduzco-Flores, S., Bodner, M., Ermentrout, B., Fuster, J. M. & Zhou, Y. Working memory cells' behavior may be explained by cross-regional networks with synaptic facilitation. *PLoS One* **4**, e6399–e6399 (2009).
21. Tsodyks, M., Pawelzik, K. & Markram, H. Neural Networks with Dynamic Synapses. *Neural Comput.* **10**, 821–835 (1998).
22. Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E. & Alonso, A. A. Graded persistent activity in entorhinal cortex neurons. *Nature* **420**, 173–178 (2002).
23. Debanne, D., Inglebert, Y. & Russier, M. Plasticity of intrinsic neuronal excitability. *Curr. Opin. Neurobiol.* **54**, 73–82 (2019).
24. Marder, E., Abbott, L. F., Turrigiano, G. G., Liu, Z. & Golowasch, J. Memory from the dynamics of intrinsic membrane currents. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13481–13486 (1996).
25. Chen, L. *et al.* The role of intrinsic excitability in the evolution of memory: Significance in memory allocation, consolidation, and updating. *Neurobiol. Learn. Mem.* **173**, 107266 (2020).
26. Conde-Sousa, E. & Aguiar, P. A working memory model for serial order that stores information in the intrinsic excitability properties of neurons. *J. Comput. Neurosci.* **35**, 187–199 (2013).
27. Fitz, H. *et al.* Neuronal spike-rate adaptation supports working memory in language processing. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20881–20889 (2020).
28. Salaj, D. *et al.* Spike frequency adaptation supports network computations on temporally dispersed information. *Elife* **10**, (2021).
29. Fiebig, F. & Lansner, A. A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *J. Neurosci.* **37**, 83–96 (2017).
30. Nessler, B., Pfeiffer, M., Buesing, L. & Maass, W. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Comput. Biol.* **9**, e1003037–e1003037 (2013).
31. Tully, P. J., Hennig, M. H. & Lansner, A. Synaptic and nonsynaptic plasticity approximating probabilistic inference. *Front. Synaptic Neurosci.* **6**, 8 (2014).
32. Markram, H., Gerstner, W. & Sjöström, P. J. A history of spike-timing-dependent plasticity. *Front. Synaptic Neurosci.* **3**, 4 (2011).
33. Ingram, R. & Volianskis, A. Promiscuous involvement of mGluRs in the storage of

NMDAR-dependent STP. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* (2024).

34. Lisman, J. Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes: long-term potentiation, long-term depression, short-term potentiation and scaling. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, (2017).
35. Erickson, M. A., Maramara, L. A. & Lisman, J. A Single Brief Burst Induces GluR1-dependent Associative Short-term Potentiation: A Potential Mechanism for Short-term Memory. *J. Cogn. Neurosci.* **22**, 2530–2540 (2010).
36. Lim, S. Hebbian learning revisited and its inference underlying cognitive function. *Curr. Opin. Behav. Sci.* **38**, 96–102 (2021).
37. Lansner, A., Fiebig, F. & Herman, P. Fast Hebbian plasticity and working memory. *Curr. Opin. Neurobiol.* **83**, 102809 (2023).
38. Volianskis, A. *et al.* Different NMDA receptor subtypes mediate induction of long-term potentiation and two forms of short-term potentiation at CA1 synapses in rat hippocampus in vitro. *J. Physiol.* **591**, 955–972 (2013).
39. Clark, D. G. & Abbott, L. F. Theory of Coupled Neuronal-Synaptic Dynamics. *Phys. Rev. X* **14**, 021001 (2024).
40. Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P. & Husain, M. Neural mechanisms of attending to items in working memory. *Neurosci. Biobehav. Rev.* **101**, 1–12 (2019).
41. Sandberg, A., Tegnér, J. & Lansner, A. A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems* vol. 14 789–802 Preprint at https://doi.org/10.1088/0954-898X/14/4/309 (2003).
42. Fiebig, F., Herman, P. & Lansner, A. An Indexing Theory for Working Memory Based on Fast Hebbian Plasticity. *eNeuro* **7**, (2020).
43. Sandberg, A., Lansner, A., Petersson, K. M. & Ekeberg, O. A Bayesian attractor network with incremental learning. *Network* **13**, 179–194 (2002).
44. Lansner, A., Ravichandran, N. B. & Herman, P. Benchmarking Hebbian learning rules for associative memory. *arXiv [cs.NE]* (2023).
45. Gerstner, W. & Kistler, W. M. *Spiking Neuron Models*. (Cambridge University Press, Cambridge, England, 2005).
46. Zenke, F., Agnes, E. J. & Gerstner, W. Diverse synaptic plasticity mechanisms orchestrated to

form and retrieve memories in spiking neural networks. *Nature Communications* vol. 6 Preprint at https://doi.org/10.1038/ncomms7922 (2015).

47. Li, X., Wang, W., Xue, F. & Song, Y. Computational modeling of spiking neural network with learning rules from STDP and intrinsic plasticity. *Physica A: Statistical Mechanics and its Applications* **491**, 716–728 (2018).
48. Savin, C. & Triesch, J. Emergence of task-dependent representations in working memory circuits. *Front. Comput. Neurosci.* **8**, 57 (2014).
49. Amit, D. J. & Brunel, N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* **7**, 237–252 (1997).
50. Fransén, E. & Lansner, A. A model of cortical associative memory based on a horizontal network of connected columns. *Network* **9**, 235–264 (1998).
51. Brunel, N. & Wang, X. J. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* **11**, 63–85 (2001).
52. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cereb. Cortex* **10**, 910–923 (2000).
53. Amit, D. J. & Mongillo, G. Spike-Driven Synaptic Dynamics Generating Working Memory States. *Neural Comput.* **15**, 565–596 (2003).
54. Tiddia, G., Golosio, B., Fanti, V. & Paolucci, P. S. Simulations of working memory spiking networks driven by short-term plasticity. *Front. Integr. Neurosci.* **16**, 972055 (2022).
55. Mongillo, G. Synaptic models of working memory. in *Reference Module in Neuroscience and Biobehavioral Psychology* (Elsevier, 2025).
56. Izhikevich, E. M. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* **17**, 2443–2452 (2007).
57. Lansner, A., Marklund, P., Sikström, S. & Nilsson, L. Reactivation in Working Memory: An Attractor Network Model of Free Recall. *PLoS One* **8**, e73776–e73776 (2013).
58. Orhan, A. E. & Ma, W. J. Publisher Correction: A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 505 (2019).
59. Chrysanthidis, N., Fiebig, F., Lansner, A. & Herman, P. Traces of semantization - from episodic to semantic memory in a spiking cortical network model. *eNeuro* **9**, ENEURO.0062–22.2022 (2022).

60. Chen, Z. S. & Wilson, M. A. How our understanding of memory replay evolves. *J. Neurophysiol.* **129**, 552–580 (2023).

61. Miller, E. K., Lundqvist, M. & Bastos, A. M. Working Memory 2.0. *Neuron* vol. 100 463–475 Preprint at https://doi.org/10.1016/j.neuron.2018.09.023 (2018).

62. Li, C., Frischkorn, G. T. & Oberauer, K. Updating of information in working memory: Time course and consequences. *Cogn. Psychol.* **156**, 101702 (2025).

63. Lewis-Peacock, J. A., Kessler, Y. & Oberauer, K. The removal of information from working memory. *Ann. N. Y. Acad. Sci.* **1424**, 33–44 (2018).

64. Clapp, W. C., Rubens, M. T. & Gazzaley, A. Mechanisms of working memory disruption by external interference. *Cereb. Cortex* **20**, 859–872 (2010).

65. Holmgren, C., Harkany, T., Svennenfors, B. & Zilberter, Y. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J. Physiol.* **551**, 139–153 (2003).

66. D'Esposito, M. & Postle, B. R. The Cognitive Neuroscience of Working Memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).

67. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).

68. Wu, Y. & Maass, W. A simple model for Behavioral Time Scale Synaptic Plasticity (BTSP) provides content addressable memory with binary synapses and one-shot learning. *Nat. Commun.* **16**, 342 (2025).

69. Fuster, J. M. Distributed memory for both short and long term. *Neurobiol. Learn. Mem.* **70**, 268–274 (1998).

70. Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F. & Nyberg, L. Neurocognitive Architecture of Working Memory. *Neuron* vol. 88 33–46 Preprint at https://doi.org/10.1016/j.neuron.2015.09.020 (2015).

71. Sreenivasan, K. K., Curtis, C. E. & D'Esposito, M. Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18**, 82–89 (2014).

72. Bocincova, A., Buschman, T. J., Stokes, M. G. & Manohar, S. G. Neural signature of flexible coding in prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2200400119 (2022).

73. Lundqvist, M. *et al.* Gamma and Beta Bursts Underlie Working Memory. *Neuron* **90**, 152–164 (2016).

74. Oberauer K. & Lewandowsky, S. Modeling working memory: a computational implementation of the Time-Based Resource-Sharing theory. *Psychon. Bull. Rev.* **18**, 10–45 (2010).

75. Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).

76. Brette, R. & Gerstner, W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.* **94**, 3637–3642 (2005).

77. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences* **94**, 719–723 (1997).

78. Wahlgren, N. & Lansner, A. Biological evaluation of a Hebbian-Bayesian learning rule. *Neurocomputing* **38-40**, 433–438 (2001).

79. Volianskis, A. *et al.* Long-term potentiation and the role of N-methyl-D-aspartate receptors. *Brain Res.* **1621**, 5–16 (2015).

80. Matzel, L. D. & Sauce, B. A multi-faceted role of dual-state dopamine signaling in working memory, attentional control, and intelligence. *Front. Behav. Neurosci.* **17**, 1060786 (2023).

81. Gewaltig, M.-O. & Diesmann, M. NEST (NEural Simulation Tool). *Scholarpedia J.* **2**, 1430–1430 (2007).

82. Hoffman, D. A., Magee, J. C., Colbert, C. M. & Johnston, D. K+ channel regulation of signal propagation in dendrites of hippocampal pyramidal neurons. **387**, 869–875 (1997).

83. Petersson, M. E., Yoshida, M. & Fransén, E. A. Low-frequency summation of synaptically activated transient receptor potential channel-mediated depolarizations. *Eur. J. Neurosci.* **34**, 578–593 (2011).

84. Kapfer, C., Glickfeld, L. L., Atallah, B. V. & Scanziani, M. Supralinear increase of recurrent inhibition during sparse activity in the somatosensory cortex. *Nat. Neurosci.* **10**, 743–753 (2007).

85. Ren, M., Yoshimura, Y., Takada, N., Horibe, S. & Komatsu, Y. Specialized inhibitory synaptic actions between nearby neocortical pyramidal neurons. *Science* **316**, 758–761 (2007).

86. Silberberg, G. & Markram, H. Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells. *Neuron* **53**, 735–746 (2007).

87. Tully, P. J., Lindén, H., Hennig, M. H. & Lansner, A. Spike-Based Bayesian-Hebbian Learning of Temporal Sequences. *PLoS Comput. Biol.* **12**, e1004954–e1004954 (2016).

88. Parisien, C., Anderson, C. H. & Eliasmith, C. Solving the Problem of Negative Synaptic Weights in Cortical Models. *Neural Comput.* **20**, 1473–1494 (2008).

89. Chrysanthidis, N., Fiebig, F. & Lansner, A. Introducing double bouquet cells into a modular cortical associative memory model. *bioRxiv* 462010–462010 (2018).

90. Mountcastle, V. B. The columnar organization of the cerebral cortex. *Brain* **120**, 701–722 (1997).

91. Douglas, R. J. & Martin, K. A. C. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27**, 419–451 (2004).

92. Herman, P. A., Lundqvist, M. & Lansner, A. Nested theta to gamma oscillations and precise spatiotemporal firing during memory retrieval in a simulated attractor network. in vol. 1536 68–87 (2013).

93. DeFelipe, J., Ballesteros-Yáñez, I., Inda, M. C. & Muñoz, A. Double-bouquet cells in the monkey and human cerebral cortex with special reference to areas 17 and 18. *Progress in Brain Research* vol. 154 15–32 Preprint at https://doi.org/10.1016/S0079-6123(06)54002-6 (2006).