An Indexing Theory for Working Memory based on Fast Hebbian Plasticity

Florian Fiebig[1], Pawel Herman[1], and Anders Lansner[1,2]

[1] Lansner Laboratory, Department of Computational Science and Technology, Royal Institute of Technology, 10044 Stockholm, Sweden,
[2] Department of Mathematics, Stockholm University, 10691 Stockholm, Sweden

Abstract

Working memory (WM) is a key component of human memory and cognitive function. Computational models have been used to uncover the underlying neural mechanisms. However, these studies have mostly focused on the short-term memory aspects of WM and neglected the equally important role of interactions between short- and long-term memory (STM, LTM). Here, we concentrate on these interactions within the framework of our new computational model of WM, which accounts for three cortical patches in macaque brain, corresponding to networks in prefrontal cortex (PFC) together with parieto-temporal cortical areas. In particular, we propose a cortical indexing theory that explains how PFC could associate, maintain and update multi-modal LTM representations.

Our simulation results demonstrate how simultaneous, brief multi-modal memory cues could build a temporary joint memory representation linked via an "index" in the prefrontal cortex by means of fast Hebbian synaptic plasticity. The latter can then activate spontaneously and thereby reactivate the associated long-term representations. Cueing one long-term memory item rapidly pattern-completes the associated un-cued item via prefrontal cortex. The STM network updates flexibly as new stimuli arrive thereby gradually over-writing older representations. In a wider context, this WM model suggests a novel explanation for "variable binding", a long-standing and fundamental phenomenon in cognitive neuroscience, which is still poorly understood in terms of detailed neural mechanisms.

Introduction

By working memory (WM), we typically understand a flexible but volatile kind of memory capable of holding a small number of items over short time spans, allowing us to act beyond the immediate here and now. WM is thus a key component in cognition and is often affected early on in neurological and psychiatric conditions, e.g. Alzheimer's disease and schizophrenia[1]. Prefrontal cortex (PFC) has repeatedly been implicated as a key neural substrate for WM in humans and non-human primates[2,3].

Computational models of WM have so far focused mainly on its short-term memory aspects, explained either by means of persistent activity[4–7] or more recently fast synaptic plasticity[8,9] as the underlying neural mechanism. However, an equally important aspect of WM is the dynamic interaction between short- and long-term memory (STM, LTM), i.e. its ability to activate or "bring

online" a small set of task relevant LTM representations. This enables very prominent and complex cognitive phenomena, which have been characterized extensively in experiments on humans as well as animals. Nevertheless, the underlying neural mechanisms still remain elusive.

Here we present a large-scale spiking neural network model of WM and focus on investigating the neural mechanisms behind these critical STM-LTM interactions. In this context, we introduce a cortical indexing theory, inspired by the predecessor hippocampal memory indexing theory[10] originally proposed to account for hippocampus's role in storing episodic memories[11]. The core idea of our theory rests on the concept of cell assemblies formed in the PFC as "indices" that link LTM representations. Our model comprises a subsampled PFC network model of STM that is reciprocally connected with two LTM component networks representing different sensory modalities (e.g. visual and auditory) in temporal cortical areas. This new model builds on and extends our recent PFC-dependent STM model of human word-list learning[9] and it employs the same fast Hebbian plasticity as a key neural mechanism, intrinsically within PFC and in addition in PFC backprojections that target temporal LTM stores. To function in this context, plasticity needs to be Hebbian, i.e. associative, and has to be induced and expressed on a time-scale of a few hundred milliseconds. Recent experiments have demonstrated the existence of fast forms of Hebbian synaptic plasticity, e.g. short-term potentiation (STP)[12,13], which lends credibility to this type of WM mechanism.

We hypothesize that activity in parieto-temporal LTM stores targeting PFC via fixed patchy synaptic connections triggers an activity pattern in PFC, which is rapidly connected by means of fast Hebbian plasticity to form a cell assembly displaying attractor dynamics. The connections in backprojections from PFC to the same LTM stores are also enhanced and connects specifically with the triggering/indexing cell assemblies there. Our simulations demonstrate that such a composite WM model can function as a robust and flexible multi-item and cross-modal WM that maintains a small set of activated task relevant LTM representations and associations. Transiently formed cell assemblies in PFC serve the role of indexing and temporary binding of these LTM representations, hence giving rise to the name of the proposed indexing theory. The PFC cell assemblies can activate spontaneously and thereby reactivate the associated long-term representations. Cueing one LTM item rapidly activates the associated un-cued item via PFC by means of pattern completion. The STM network flexibly updates WM content as new stimuli arrive whereby older representations gradually fade away. Interestingly, this model implementing the cortical indexing theory can also explain the so far poorly understood cognitive phenomenon of variable binding or object – name association, which is one key ingredient in human reasoning and planning[14–16].

Results

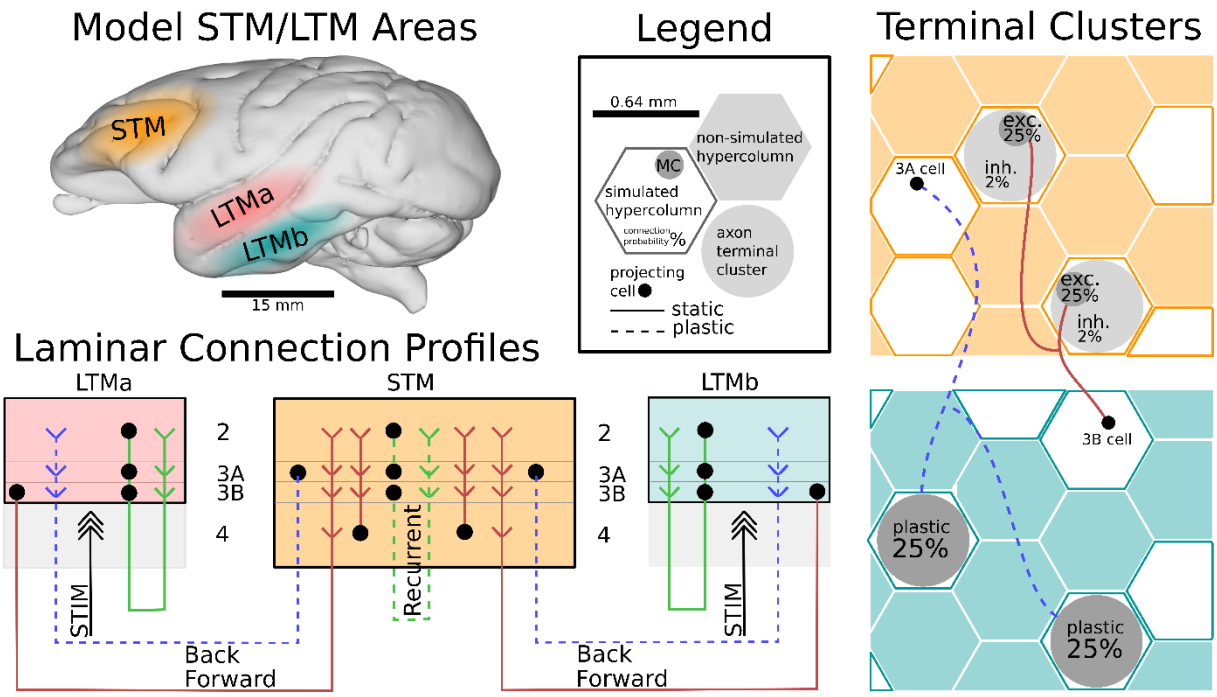


Figure 1. Schematic of modeled connectivity within and across representative STM and LTM areas in macaque. The model organizes cells into grids of nested hypercolumns (HCs) and minicolumns (MCs), sometimes referred to as macro columns, and "functional columns" respectively. STM features 25 HCs, whereas LTMa and LTMb both contain 16 simulated HCs. Each network spans several hundred mm² and the simulated columns constitute a spatially distributed subsample of biological cortex, defined by conduction delays. Pyramidal cells in the simulated supragranular layers form connections both within and across columns. STM features an input layer 4 that shapes the input response of cortical columns, whereas LTM is instead stimulated directly to cue the activation of previously learned long-term memories. Additional corticocortical connections (feedforward in brown, feedback in dashed blue) are sparse (<1% connection probability) and implemented with terminal clusters (rightmost panels) and specific laminar connection profiles (bottom left). The connection schematic illustrates laminar connections realizing a direct supragranular forward-projection, as well as a common supragranular backprojection. Layer 2/3 recurrent connections in STM (dashed green) and corticocortical backprojections (dashed blue) feature fast Hebbian plasticity. For an in-depth model description, including the columnar microcircuits, please refer to *Online Methods* and **Supplementary Figure 1**.

Our model implements WM function arising from the interaction of STM and LTM networks, which manifests itself in multi-modal memory binding phenomena. To this end, we simulate three cortical patches with significant biophysical detail: an STM and two LTM networks (LTMa, LTMb), representing PFC and parieto-temporal areas, respectively (**Figure 1**). The computational network model used here represents a detailed modular cortical microcircuit architecture in line with previous models[17,18]. In particular, the current model is built upon a recent STM model[9]. The associative cortical layer 2/3 network of that model was sub-divided into layers 2, 3A, and 3B and extended with an input layer 4 and corticocortical connectivity to LTM stores in temporal cortical regions. This large, composite model synthesizes many different anatomical and electrophysiological cortical data and produces complex output dynamics. We specifically focus on the dynamics of memory specific subpopulations in the interaction of STM and LTM networks.

We introduce the operation of the WM model in several steps. First, we take a brief look at background activity and active memory states in isolated cortical networks of this kind to familiarize the reader with some of its dynamical properties. Second, we describe the effect of memory activation on STM with and without plasticity. Third, we add the plastic backprojections from STM to LTM and monitor the encoding and maintenance of several memories in the resulting STM-LTM loop. We track the evolution of acquired cell assemblies with shared pattern-selectivity in STM and show

their important role in WM maintenance (a.k.a. delay activity). We then demonstrate that the emerging WM network system is capable of updating the set of maintained memories. Finally, we simulate multi-modal association and analyze its dynamical correlates. We explore temporal characteristics of network activations and cross-cortical delays during WM encoding, maintenance, and cue-driven associative recall of multi-modal memories (LTMa-LTMb pairs of associated memories).

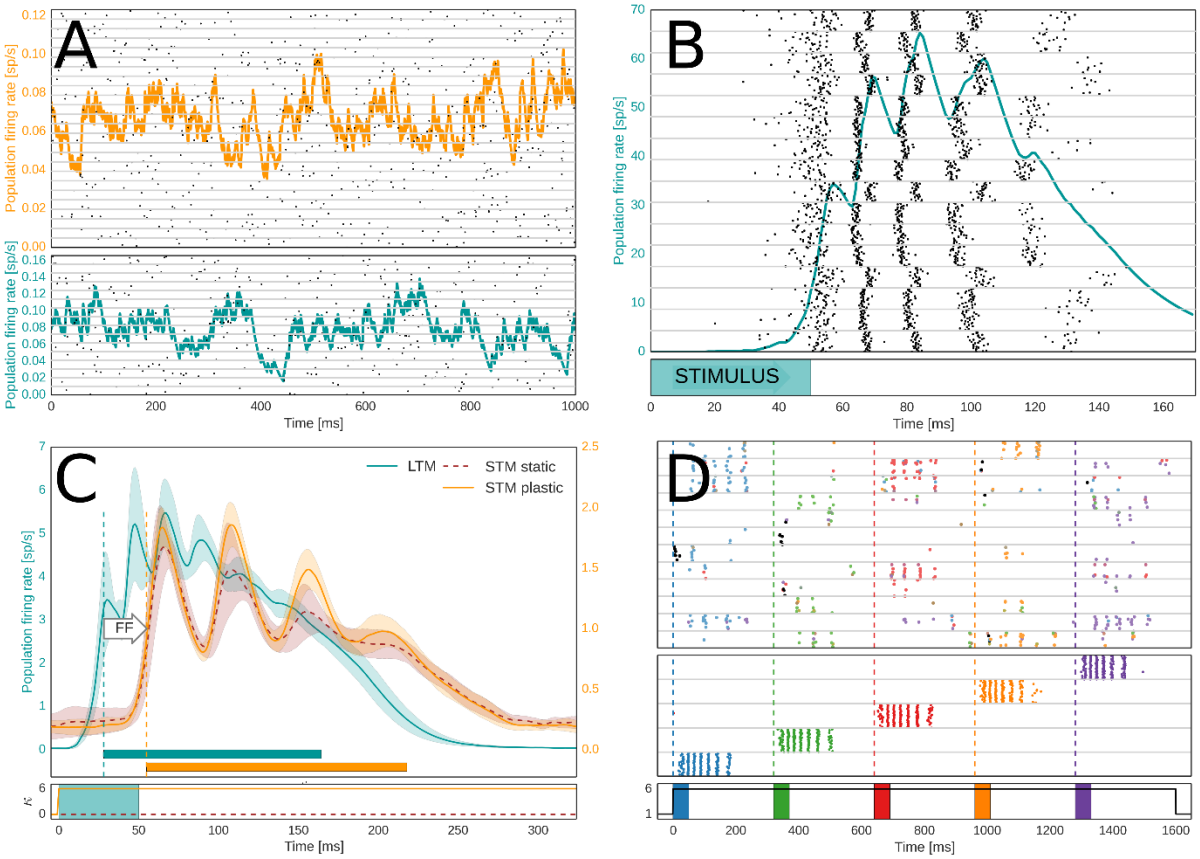


Figure 2. Basic Network behavior in spike rasters and population firing rates. A: The untrained networks STM (top) and LTM (bottom) feature low rate, asynchronous activity (CV2 = 0.7±0.2). The underlying spike raster shows layer 2/3 activity in each HC (separated by grey horizontal lines) in the simulated network. **B:** Cued LTM memory activation expressed as fast oscillation bursts (40-50 Hz), organized into a theta-like envelope (3 Hz). The underlying spike rasters shows layer 2/3 activity of the activated MC in each HC, revealing spatial synchronization. The brief stimulus is a memory specific cue. **C:** LTM-to-STM forward dynamics as shown in population firing rates of STM and LTM activity following LTM-activation induced by a 50 ms targeted stimulus at time 0. LTM-driven activations of STM are characterized by a feedforward delay (FF). Shadows indicate the standard deviation of 100 peri-stimulus activations in LTM (blue) and STM with plasticity (orange) and without intrinsic plasticity (dashed, dark orange). Horizontal bars indicate the activation half-width (*Online Methods*). Onset is denoted by vertical dashed lines. The stimulation of LTM and activation of plasticity is denoted underneath. **D:** Subsampled spike raster of STM (top) and LTM (middle) during forward activation of the untrained STM by five different LTM memory patterns, triggered via specific memory cues in LTM at times marked by the vertical dashed lines. Bottom spike raster shows LTM layer 2/3 activity of one selective MC per activated pattern (colors indicate different patterns). Top spike raster shows layer 2/3 activity of one HC in STM. STM spikes are colored according to each cells dominant pattern-selectivity (based on the memory pattern correlation of individual STM cell spiking during initial pattern activation, see Online Methods, *Spike Train Analysis and Memory Activity Tracking*). Bottom: The five stimuli to LTM (colored boxes) and modulation of STM plasticity (black line).

Background activity and Activated memory

The untrained network (see *Online Methods*) features fluctuations in membrane voltages and low-rate, asynchronous spiking activity (**Figure 2-A**). At higher background input levels, the empty network transitions into a state characterized by global oscillations in the alpha/beta range (**Supplementary Figure 2**). This is largely an effect of fast feedback inhibition from local basket cells (**Supplementary Figure 1**), high connection density within MCs, and low latency local spike transmission.[19] If the network has been trained with structured input so as to encode memory (i.e.

attractor states), a specific cue (*Online Methods*) can trigger memory item reactivations accompanied by fast oscillations modulated by an underlying slow oscillation in the lower theta range (~3 Hz) [20,21] (**Figure 2-B**). The spiking activity of memory activations (a.k.a. attractors) is short-lived due to neural adaptation and synaptic depression. When unspecific background excitation is very strong, this can result in a random walk across stored memories[9,20].

LTM-to-STM Forward Dynamics

We now consider cued activation of several memories embedded in LTM. Each HC in LTM features selectively coding MCs for given memory patterns that activate synchronously in theta-like cycles each containing several fast oscillation bursts (**Figure 2-B**). Five different LTM memory patterns are triggered by brief cues, accompanied by an upregulation of STM plasticity, see **Figure 2-D (bottom)**. To indicate the spatio-temporal structure of evoked activations in STM, we also show a simultaneous subsampled STM spike raster (**Figure 2-D top**). STM activations are sparse (ca 5%), but despite this nearby cells (in the same MC) often fire together. The distributed, patchy character of the STM response to memory activations (**Figure 2-D top**) is shaped by branching forward-projections from LTM layer 3B cells, which tend to activate close-by cells. STM input layer 4 receives half of these corticocortical connections and features very high fine-scale specificity in its projections to layer 2/3 pyramidal neurons, which furthers recruitment of local clusters with shared selectivity. STM cells initially fire less than those in LTM because the latter received a brief, but strong activation cue and have strong recurrent connections if they code for the same embedded memory pattern. STM spikes in **Figure 2-D** are colored according to the cells' dominant memory pattern selectivity (Online Methods, *Spike Train Analysis and Memory Activity Tracking*), which reveals that STM activations are mostly non-overlapping as well. Unlike the organization of LTM with strictly non-overlapping memory patterns, MC activity in STM is not exclusive to any given input pattern, but nearby cells often still have similar pattern selectivity. This is not only an effect of competition via basket cell feedback inhibition, but also a result of short-term dynamics, such as neural adaptation and synaptic depression. Neurons that have recently been activated by a strong, bursting input from LTM are refractory and thus less prone to spike again for some time thereafter (τ_{rec} and τ_{I_w}, **Supplementary Table 1**), further reducing the likelihood of activating overlapping STM activation patterns. **Figure 2-C** shows a peri-stimulus population firing rate of both STM and LTM networks (mean across 100 trials with five triggered memories each). There is a bottom-up response delay between stimulus onset at t=0 and LTM activation, as well as a substantial forward delay (scrutinized in more detail later on). Oscillatory activity in STM is lower than in LTM mostly because the untrained STM lacks strong recurrent connections. It is thus less excitable, and therefore does not trigger its basket cells (the main drivers of fast oscillations in our model) as quickly as in LTM. Fast oscillations in STM and the amplitude of their theta-like envelope build up within a few seconds as new cell assemblies become stronger (e.g. **Figure 3-A** and **Supplementary Figure 3**). As seen in **Figure 2-B**, bursts of co-activated MCs in LTM can become asynchronous during activation. Dispersed forward axonal conduction delays further decorrelate this gamma-like input to STM. Activating strong plasticity in STM ($\kappa = \kappa_p$, *Online Methods* and **Supplementary Table 1**) has a noticeable effect on the amplitude of stimulus-locked oscillatory STM activity after as little as 100 ms (cf. **Figure 2-C, STM**).

Multi-item Working Memory

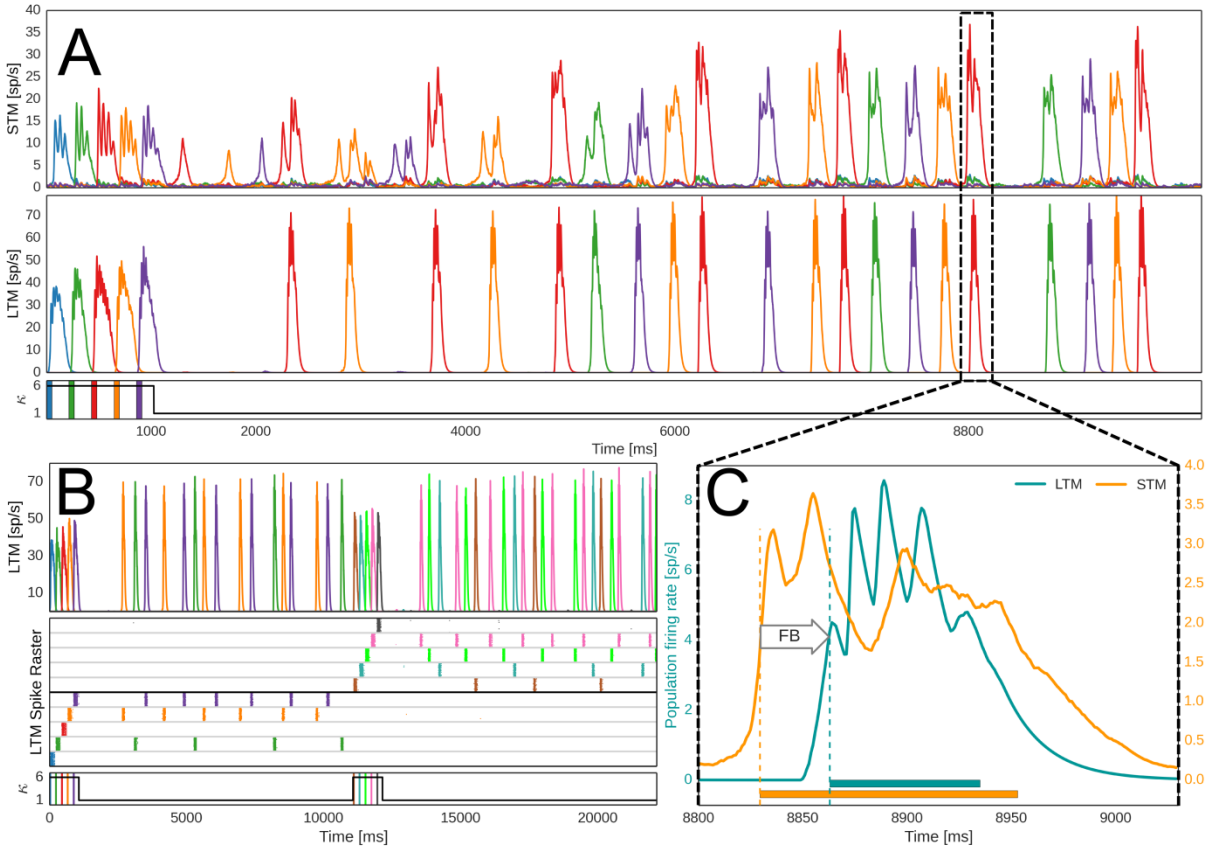


Figure 3. Encoding and feedback-driven reactivation of LTM. A: Firing rates of pattern-specific subpopulations in STM and LTM during encoding and subsequent maintenance of five memories. Just as in the plasticity-modulated stimulation phase shown in Figure 2D, five LTM memories are cued via targeted 50 ms stimuli (shown underneath). Plasticity of STM and its backprojections is again elevated six-fold during the initial memory activation. Thereafter, a strong noise drive to STM causes spontaneous activations and plasticity induced consolidation of pattern-specific subpopulations in STM (lower plasticity, $\kappa = 1$). Backprojections from STM cell assemblies help reactivate associated LTM memories. **B:** Updating of WM. Rapid encoding and subsequent maintenance of a second group of memories following an earlier set. The LTM Spike raster shows layer 2/3 activity of one LTM HC (MCs separated by grey horizontal lines), the population firing rate of pattern-specific subpopulations across the whole LTM network is seen above. Underneath we denote stimuli to LTM and the modulation of plasticity, κ, in STM and its backprojections. **C:** STM-to-LTM loop dynamics during a spontaneous reactivation event. STM-triggered activations of LTM memories are characterized by a feedback delay and a second peak in STM after LTM activations. Horizontal bars at the bottom indicate activation half-width (*Online Methods*). Onset is denoted by vertical dashed lines.

In **Figure 2-D,** we have shown pattern-specific subpopulations in STM emerging from feedforward input. Modulated STM plasticity allows for the quick formation of rather weak STM cell assemblies from one-shot learning. When we include plastic STM backprojections, these assemblies can serve as an index for specific memories.

Their recruitment is temporary, but they can act as top-down control signals for memory maintenance and retrieval. STM backprojections with fast Hebbian plasticity can index multiple activated memories in the closed STM-LTM loop**.** In **Figure 3-A**, we show network activity following targeted activation of five LTM memories (Spike raster in **Supplementary Figure 3**). Under an increased unspecific noise-drive ($r_{bg-high}^{L23}$, **Supplementary Table 2**), STM cell assemblies, formed during the brief plasticity-modulated stimulus phase (cf. **Figure 2D**) may activate spontaneously. These brief bursts of activity are initially weak and different from the theta-like cycles of repeated fast bursting seen in LTM attractor activity.

STM recurrent connections remain plastic ($\kappa = 1$) throughout the simulation, so each reactivation event further strengthens memory-specific cell assemblies in STM. As a result, there is a noticeable ramp-up in the strength of STM pattern-specific activity over the course of the delay period (cf. increasing burst length and amplitude in **Figure 3-A,** or **Supplementary Figures 4, 6**). STM backprojections are also plastic and thus acquire memory specificity from STM-LTM co-activations, especially during the initial stimulation phase. Given enough STM cell assembly firing, their sparse but potentiated backprojections can trigger associated memories. Weakly active assemblies may fail to do so. In the example of **Figure 3-A,** we can see a few early STM reactivations that are not accompanied (or quickly followed) by a corresponding LTM pattern activation (of the same color) in the first two seconds after the plasticity-modulated stimulation phase. When LTM is triggered, there is a noticeable feedback delay (**Figure 3-C**), which we will scrutinize later on.

Cortical feedforward and feedback pathways between LTM and STM form a loop, so each LTM activation will again feed into STM, typically causing a second peak of activation in STM 40 ms after the first (**Figure 3-C**). The forward delay from LTM to STM, that we have seen earlier in the stimulus-driven input phase **(Figure 2-C),** is still evident here in this delayed secondary increase of the STM activation following LTM onset, which also extends/sustains the STM activation and helps stabilize memory-specific STM cell assemblies and their specificity. This effect may be called auto-consolidation and it is an emergent feature of the plastic STM-LTM loop in this model. It happens on a timescale governed by the unmodulated plasticity time constant ($\kappa = \kappa_{normal}$, $\tau_p = 5\ s$, **Supplementary Table 1**). After a few seconds, the network has effectively stabilized and typically maintains a small set of 3-4 activated long-term memories. The closed STM-LTM loop thus constitutes a functional multi-item WM.

A crucial feature of any WM system is its flexibility, and **Figure 3-B** highlights an example of rapid updating. The maintained set of activated memories can be weakened by stimulating yet another set of input memories. Generally speaking, earlier items are reliably displaced from active maintenance in our model if activation of the new items is accompanied by the same transient elevation of plasticity (κ_p/κ_{normal}, **Supplementary Table 1**) used during the original encoding of the first five memories (Corresponding spike rasters and population firing rates are shown in **Supplementary Figures 4**, and **5)**.

In line with earlier results[9], cued activation can usually still retrieve previously maintained items. The rate of decay for memories outside the maintained set depends critically on the amount of noise in the system, which erodes the learned associations between STM and LTM neurons as well as STM cell assemblies. We note that such activity-dependent memory decay is substantially different from time-dependent decay, as in Mi et al.[22].

Multi-modal, Multi-item Working Memory

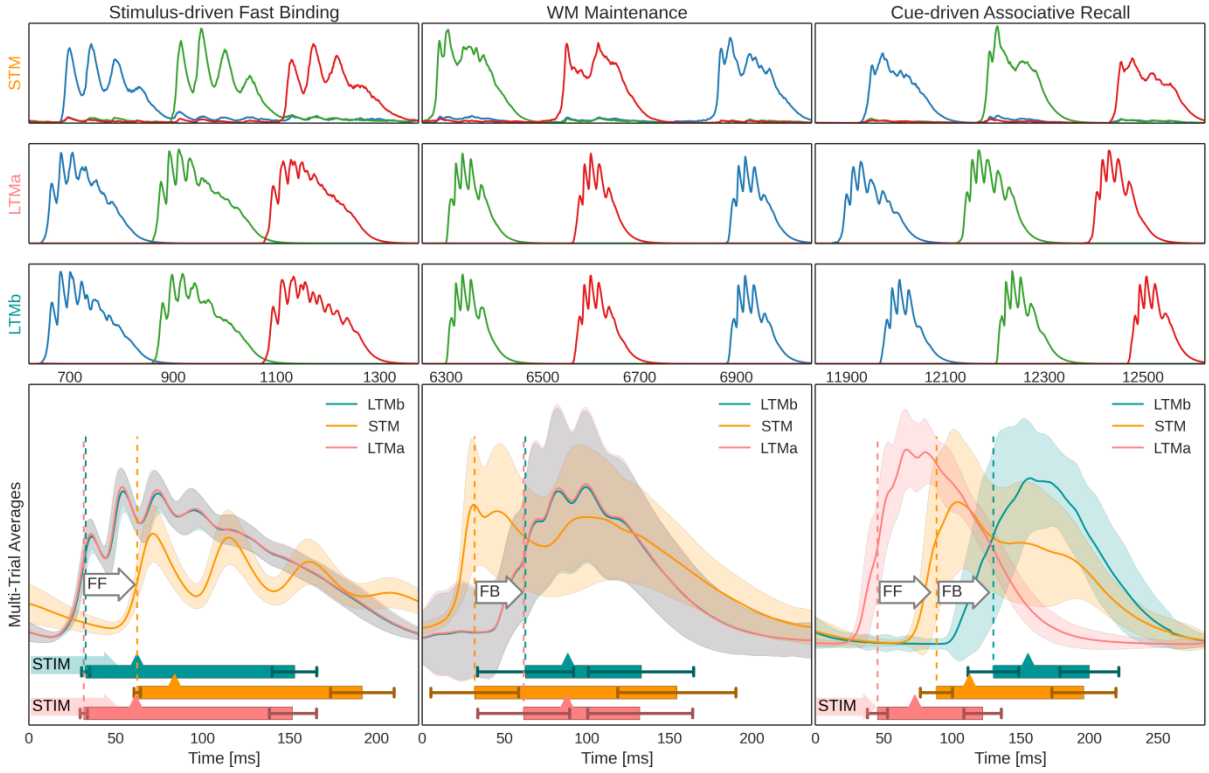


Figure 4. Population firing rates of networks and memory-specific subpopulations during three different modes of network activity : **Top-Half**: Exemplary activation of three memories (blue, green, red respectively) in STM (1st row), LTMa (2nd row), and LTMb (3rd row) during three different modes of network activity: The initial association of pairs of LTM memory activations in STM (left column), WM Maintenance through spontaneous STM-paced activations of bound LTM memory pairs (middle column), and cue-driven associative recall of previously paired stimuli (right column). **Bottom-Half**: Multi-trial peri-stimulus activity traces from the three cortical patches across 100 trials (495 traces, as each trial features 5 activated and maintained LTM memory pairs and very few failures of paired activation). Shaded areas indicate a standard deviation from the underlying traces. Vertical dashed lines denote mean onset of each network's activity, as determined by activation half-width (*Online Methods*), also denoted by a box underneath the traces. Error bars indicate a standard deviation from activation onset and offset. Mean peak activation is denoted by a triangle on the box, and shaded arrows to the left of the box denote targeted pattern stimulation of a network at time 0. As there are no external cues during WM maintenance (aka delay period), we use detected STM activation onset to align firing rate traces of 5168 STM-paced LTM-reactivations across trials and reactivation events for averaging. White arrows annotate feedforward (FF) and feedback (FB) delay, as defined by respective network onsets.

Next, we explore the ability of the closed STM-LTM loop system to flexibly bind co-active pairs of long-term memories from different modalities (LTMa and LTMb respectively). As both LTM activations trigger cells in STM via feedforward projections, a unique joint STM cell assembly with shared pattern-selectivity is created. Forward-activations include excitation and inhibition and combine non-linearly with each other (*Online Methods*) and with prior STM content. **Figure 4** illustrates how this new index then supports WM operations, including delay maintenance through STM-paced co-activation events and stimulus-driven associative memory pair completion. The three columns of **Figure 4** illustrate three fundamental modes of the closed STM-LTM loop: stimulus-driven encoding, WM maintenance, and associative recall. The top three rows show sampled activity of a single trial (see also **Supplementary Figures 6,7**), whereas the bottom row shows multi-trial averages.

During stimulus-driven association, we co-activate memories from both LTM's by brief 50 ms cues that trigger activation of the corresponding memory patterns. The average of peri-stimulus activations reveals a 45 ± 7.3 ms LTM attractor activation delay, followed by a 43 ± 7.8 ms feedforward delay (about half of which is explained by axonal conduction time due to the spatial

distance between LTM and STM) from the onset of the LTM activations to the onset of the input-specific STM response (**Figure 5 top-left** and **bottom-left**).

During WM maintenance, a 10 s delay period, paired LTM memories reactivate together. Onset of these paired activations is a lot more variable than during cued activation with a feedback delay mean of 41.5 ± 15.3 ms, mostly because the driving STM-activations are of variable size and strength.

Following the maintenance period, we test the memory system's ability for associative recall. To this end, we cue LTMa, again using a targeted 50 ms cue for each memory, and track the systems response across the STM-LTM loop. We compute multi-trial averages of peri-stimulus activations during recall testing (**Figure 4 bottom-right).** Following cued activation of LTMa, STM responds with the related joint cell assembly activation as the input is strongly correlated to the learned inputs as a result of the simultaneous activation with LTMb earlier on. Similar to the mnemonic function of an index, the completed STM pattern then triggers the associated memory in LTMb through its backprojections. STM activation now extends far beyond the transient activity of LTMa because STM recurrent connectivity and the STM-LTMb backprojection re-excite it. Temporal overlap between associated LTMa and LTMb memory activations peaks around 125 ms after the initial stimulus to LTMa.

We collect distributions of feedforward and feedback delays during associative recall (**Figure 5)**. To facilitate a more immediate comparison with biological data we also compute the Bottom-Up and Top-Down response latency of the model in analogy to Tomita et al.[23]. Their study explicitly tested widely held beliefs about the executive control of PFC over ITC in memory retrieval. To this end, they identified and recorded neurons in ITC of monkeys trained to memorize several visual stimulus-stimulus associations. They employed a posterior-split brain paradigm to cleanly disassociate the timing of the bottom-up (contralateral stimuli) and top-down response (ipsilateral stimuli) in 43 neurons significantly stimulus-selective in both conditions. They observed that the latency of the top-down response (178 ms) was longer than that of the bottom-up response (73 ms).

Our simulation is analogous to this experimental setup with respect to some key features, such as the spatial extent of memory areas (STM/dlPFC about 289 mm²) and inter-area distances (40 mm cortical distance between PFC and ITC). These measures heavily influence the resulting connection delays and time needed for information integration. In analogy to the posterior-split brain experiment our model's LTMa and LTMb are unconnected. However, we now have to consider them as ipsi- and contralateral visual areas in ITC. The display of a cue in one hemi-field in the experiment then corresponds to the LTMa-sided stimulation of an associated memory pair in the model. This arrangement forces any LTM interaction through STM (representing PFC), and allows us to treat the cued LTMa memory activation as a Bottom-up response, whereas the much later activation of the associated LTMb representation is related to the Top-down response in the experimental study. **Figure 5** shows the distribution of these latencies in our simulations, where we also marked the mean latencies measured by Tomita et al. The mean of our bottom-up delay (72.9 ms) matches the experimental data (73 ms), whereas the mean of the broader top-down latency distribution (155.2 ms) is a bit lower than in the monkey study (178 ms). Of these 155.2 ms, only 48 ms are explained by the spatial distance between networks, as verified by a fully functional alternative model with 0 mm distance between networks.

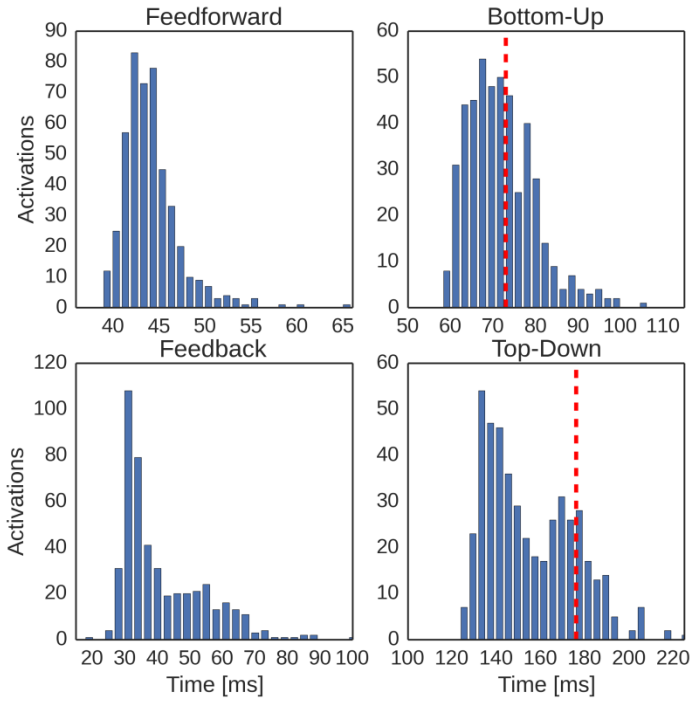


Figure 5. Comparison of key activation delays during associative recall in model and experiment following a cue to LTMa. Top-Left: Feedforward delay distribution in the model, as defined by the temporal delay between LTMa onset and STM onset (as shown in Figure 4, Bottom-right). **Top-Right**: Bottom-up delay distribution in the model, as defined by the temporal delay between stimulation onset and LTMa peak activation. The red line denotes the mean bottom-up delay, as measured by Tomita et al.[23]. **Bottom-Left**: Feedback delay distribution in the model, as defined by the temporal delay between STM onset and LTMb onset (measured by half-width, as shown in Figure 4, Bottom-right). **Bottom-Right**: Top-Down delay distribution in the model, as defined by the temporal delay between stimulation onset and LTMb peak activation. The red line denotes the mean bottom-up delay, as measured by Tomita et al.[23]. Model delays were averaged over 100 trials with 5 paired stimuli each.

Discussion

We have in this work presented and tested a novel theory for WM. It hypothesizes that activity in parieto-temporal LTM stores targeting PFC via fixed or slowly plastic and patchy synaptic connections trigger an activity pattern in PFC that gets rapidly encoded by means of fast Hebbian plasticity to form a cell assembly displaying attractor dynamics. Equally plastic backprojections from PFC to the LTM stores are enhanced as well and connects the formed “index” specifically with the active cell assemblies there. This rapidly but temporarily enhanced connectivity produces a functional WM capable of encoding and maintaining individual LTM items, i.e. to bring these LTM representations “on-line”, and to form novel associations within and between several connected LTM areas and modalities. The PFC cell assemblies themselves do not encode much information but act as indices into LTM stores containing information that is more permanent. The underlying highly plastic connectivity and thereby the WM itself is flexibly remodeled and updated as new incoming activity gradually over-writes previous WM content.

We further successfully demonstrated the functional implications of this theory by implementing and evaluating a special case of a biologically plausible large-scale spiking neural network model representing PFC and two reciprocally connected LTM stores in parieto-temporal cortex. We showed how a small number of single LTM items could be encoded and maintained “on-line” and how pairs of simultaneously activated items could become jointly indexed and associated. Activating one pair member now also activates the other one indirectly via PFC with a short latency. We further demonstrated that this kind of WM could readily be updated such that as new items are encoded, old ones are fading away whereby the active WM content is replaced.

Recall dynamics in the presented model are in most respects identical to our previous cortical associative memory models[24]. Any activated memory item, whether randomly or specifically triggered, is subject to known and previously well characterized associative memory dynamics, such as pattern completion, rivalry, bursty reactivation dynamics, oscillations in different frequency bands,

etc.[19,20,25]. Moreover, sequential learning and recall could readily be incorporated[26]. This could for example support encoding of sequences of items in WM rather than unrelated single ones, resulting in reactivation dynamics reminiscent of e.g. the "phonological loop"[27].

The Case for Hebbian Plasticity

The underlying mechanism of our model is fast Hebbian plasticity, not only in the intrinsic PFC connectivity, but also in the projections from PFC to LTM stores. The former has some experimental support[12,13,28,29] whereas the latter remains a prediction of the model. Dopamine D1 receptor (D1R) activation by dopamine (DA) is strongly implicated in reward learning and synaptic plasticity regulation in the basal ganglia[30]. In analogy we propose that D1R activation is critically involved in the synaptic plasticity intrinsic to PFC and in projections to LTM stores, which would also explain the comparatively dense DA innervation of PFC and the prominent WM effects of PFC DA level manipulation[31,32]. In our model, the parameter κ represents the level of DA-D1R activation, which in turn regulates its synaptic plasticity. We typically increase kappa 4-8 fold temporarily in conjunction with stimulation of LTM and WM encoding, in a form of attentional gating. Larger modulation limits WM capacity to 1-2 items, while less modulation diminishes the strength of cell assemblies beyond what is necessary for reactivation and LTM maintenance.

When the synaptic plasticity WM hypothesis was first presented and evaluated, it was based on synaptic facilitation[8,20]. However, such non-Hebbian plasticity is only capable of less specific forms of memory. Activating a cell assembly, comprising a subset of neurons in an untrained STM network featuring such plasticity, would merely facilitate all outgoing synapses from active neurons. Likewise, an enhanced elevated resting potential resulting from intrinsic plasticity would make the targeted neurons more excitable. In either case, there would be no coordination of activity specifically within the stimulated cell assembly. Thus, if superimposed on an existing LTM, such forms of plasticity may well contribute to WM, but they are by themselves not capable of supporting encoding of novel memory items or the multi-modal association of already existing ones. In contrast, in our previous work[9] we showed that fast Hebbian plasticity similar to STP[12] allows effective one-shot encoding of novel STM items. In the current extended model, by also assuming the same kind of plasticity in backprojections from PFC to parieto-temporal LTM stores, PFC can also bind and bring on-line existing but previously unassociated LTM items across multiple modalities.

Our implementation of a fast Hebbian plasticity reproduces a remarkable aspect of STP: it decays in an activity-dependent manner[28,29]. The decay is not noticeably time-dependent, and silence preserves synaptic information. The typically detrimental effects of distractors on performance in virtually all kinds of WM tasks suggest an activity-dependent update, as does the duration of "activity-silent WM" in recent experiments[33]. Although we used the BCPNN learning rule to reproduce these effects, we expect that other Hebbian learning rules allowing for neuromodulated fast synaptic plasticity could give comparable results.

Experimental support and Testable predictions

Our model was built from available relevant microscopic data on neural and synaptic components as well as modular structure and connectivity of selected cortical areas in macaque monkey. When challenged with specific stimulus items, the network so designed generates a well-organized macroscopic dynamic working memory function, which can be interpreted in terms of manifest behavior and validated against cognitive experiments and data.

Unfortunately, the detailed neural processes and dynamics of our new model are not easily accessible experimentally and it is therefore quite hard to find direct and quantitative results to validate it. Yet, in analyzing our resulting bottom-up and top-down delays, we drew an analogy to a split-brain experiment[23] because of its clean experimental design (even controlling for subcortical pathways) and found similar temporal dynamics in our highly subsampled cortical model. The timing of inter-area signals also constitutes a testable prediction for multi-modal memory experiments. Furthermore, reviews of intracranial recordings conclude that theta band oscillations play an important role in long-range communication during successful retrieval[34]. With respect to theta band oscillations in our model, STM leads the rest of cortex during maintenance, engages bi-directionally during recall (due to the STM-LTM loop), and lags during stimulus-driven encoding and LTM activation, reflecting experimental observations[35]. These effects are explained by our model architecture, which imposes delays due to the spatial extent of networks and their distances from each other. Fast oscillations in the gamma band, while often theta-nested, are strongly linked to local processing and activated memory items in our model, also matching experimental findings[34]. Local frequency coupling is abundant with significant phase-amplitude coupling (e.g. **Figure 2B**), and was well characterized in related models[21].

The most critical requirement and thus prediction of our theory and model is the presence of fast Hebbian plasticity in the PFC backprojections to parieto-temporal memory areas. Without such plasticity, our model cannot explain the necessary STM-LTM binding. This plasticity is likely to be subject to neuromodulatory control, presumably with DA and D1R activation involvement. Since STP decays with activity, a high noise level could be an issue since it could shorten WM duration (see *The Case for Hebbian Plasticity*). The evaluation of this requirement is hampered by little experimental evidence and a general lack of experimental characterization of the synaptic plasticity in long-range corticocortical projections.

Our model also makes specific predictions about the density of corticocortical long-range connectivity. For example, as few as six active synapses (*Online Methods*) onto each coding pyramidal neuron is sufficient to transfer specific memory identities across the cortical hierarchy and to support maintenance and recall.

Finally, our model suggests the occurrence of a double peak of frontal network activation in executive control of multi-modal LTM association (see STM population activity during WM Maintenance in **Figure 4**). The first one originates from the top-down control signal itself, and the second one is a result of corticocortical reentry and a successful activation of one or more associated items in LTM. As such, the second peak should also be correlated with successful memory maintenance or associative recall.

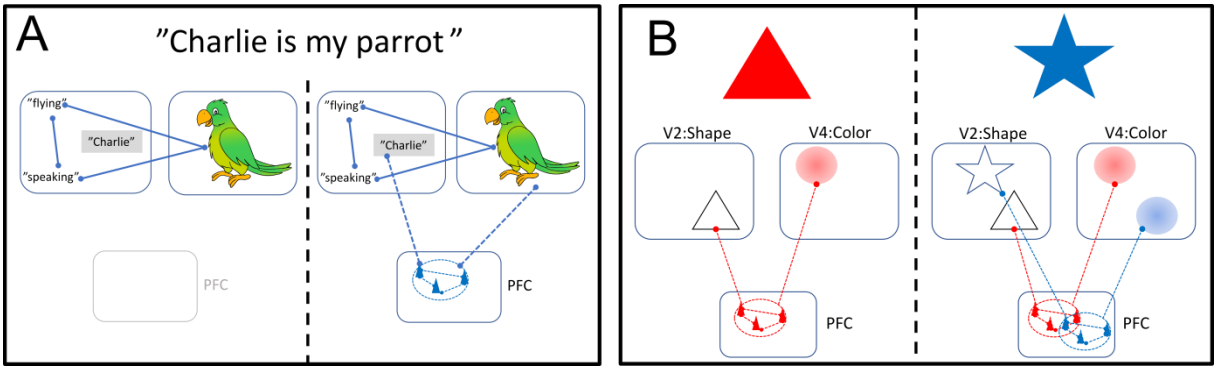


Figure 6. Name – Object binding and memorized feature binding via PFC. A: Name-Object binding: Initially the representation of "parrot" exists in LTM comprising symbolic and sub-symbolic components. When it is for the first time stated that "Charlie is my parrot", the name "Charlie" is bound reciprocally by fast Hebbian plasticity via PFC to the parrot representation, thus temporarily extending the composite "parrot" cell assembly. Pattern completion now allows "Charlie" to trigger the entire assembly and "flying" or the sight of Charlie to trigger "Charlie". If important enough or repeated a couple of times this association could consolidate in LTM. **B: Memorized Feature Binding:** When a red triangle followed by a blue star is shown and attended, these shape-color bindings are encoded by fast Hebbian plasticity via PFC to create a composite cell assembly. It supports pattern completion meaning that stimulation with shape will trigger the color representation and vice versa.

Solving the Binding Problem

The "binding problem" is a classical and extensively studied problem in perceptual and cognitive neuroscience (see e.g. Zimmer et al.[36]). Binding occurs in different forms and at different levels, from lower perceptual to higher cognitive processes[37,38]. At least in the latter case, WM and PFC feature quite prominently[14] and this is where our WM model may provide further insight.

Variable binding is a special case and a cognitive kind of neural binding in the form of a variable – value pair association of items previously not connected by earlier experience and learning[14]. A simple special case is the association of a mathematical variable and its value "The value of x is 2", i.e. x = 2. More generally, an object and a name property are bound like in "Charlie is my parrot" such that <name> = "Charlie" (**Figure 6-A**). This and other more advanced forms of neural binding are assumed to underlie complex functions in human cognition including logical reasoning and planning[39], but has been a challenge to explain by neural network models of the brain[15,40].

Based on our WM model, we propose that fast Hebbian plasticity provides a neural mechanism that solves this variable binding problem. The joint index to LTM areas formed in PFC/STM during presentation of a name – object stimulus pair, serves to bind the corresponding LTM stored variable and value representations in a specific manner that avoids mixing them up. Turning to **Figure 4** above, imagine that one of the LTMa patterns represent the image of my parrot and one pattern in LTMb, now a cortical language area, represents his name "Charlie". When this and two other image – name pairs are presented they are each associated via specific joint PFC indices. Thereafter "Charlie" will trigger the visual object representation of a parrot, and showing a picture of Charlie will trigger the name "Charlie" with a dynamics as shown in the right-most panels of **Figure 4**. Here as well, flexible updating of the PFC index will avoid confusion even if in the next moment my neighbor shouts "Charlie" to call his dog, also named Charlie. Work is in progress to uncover how such variable binding mechanisms can be used in neuro-inspired models of more advanced human logical reasoning[16].

With regard to perceptual feature binding, e.g. of object color and shape, memory is required as soon as the task demands retention of the result of the feature binding for selection of a response after the stimulus itself is gone, as illustrated in **Figure 6-B**. Recent experiments have provided support for the involvement of PFC in such memory related forms of feature binding[41]. Gamma band oscillations, frequently implicated when binding is observed, are also a prominent output of our model[42].

Conclusions

We have formulated a novel indexing theory for working memory and tested it by means of computer simulations, which demonstrated the versatile WM properties of a large-scale spiking neural network model implementing key aspects of the theory. Our model provides a novel mechanistic understanding of the targeted WM and variable binding phenomena, which connects microscopic neural processes with macroscopic observations and functions in a way that only computational models can do. While we designed and validated this model based on macaque data, the theory itself is quite general and we expect our findings to apply also to mammals including humans, commensurate with changes in key model parameters (cortical distances, axonal conductance speeds, etc.).

WM dysfunction has an outsized impact on mental health, intelligence, and quality of life. Progress in mechanistic understanding of function and dysfunction is therefore very important for society. We hope that our theoretical and computational work provides inspiration for experimentalists to scrutinize the theory and model, especially with respect to neuromodulated fast Hebbian synaptic plasticity and large-scale network architecture and dynamics. Only in this way can we get closer to a more solid understanding and theory of working memory and position future computational research and development appropriately even in the clinical and pharmaceutical realm.

Acknowledgements

This work was supported by the EuroSPIN Erasmus Mundus doctoral program, SeRC (Swedish e-science Research Center), and StratNeuro (Strategic Area Neuroscience at Karolinska Institutet, Umeå University and KTH Royal Institute of Technology). The simulations were performed using computing resources provided by the Swedish National Infrastructure for Computing (SNIC) at PDC Centre for High Performance Computing. We are grateful for helpful comments and suggestions from Drs Jeanette Hellgren Kotaleski, and Arvind Kumar.

Conflict of Interest

Nothing to declare

Online Methods

Neuron Model

We use an integrate-and-fire point neuron model with spike-frequency adaptation[43] which was modified[44] for compatibility with a custom-made BCPNN synapse model in NEST (see *Simulation Environment*) through the addition of the intrinsic excitability current I_{β_j}. The model was simplified by excluding the subthreshold adaptation dynamics. Membrane potential V_m and adaptation current are described by the following equations:

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

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

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 by b, which decays with time constant τ_{I_w}. Simulated basket cells feature neither the intrinsic excitability current I_{β_j} nor this spike-triggered adaptation.

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

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

Synapse Model

Excitatory AMPA and NMDA synapses have a reversal potential $E^{AMPA} = E^{NMDA}$, while inhibitory synapses drive the membrane potential toward E^{GABA}. In addition to BCPNN learning (next Section), plastic synapses are also subject to synaptic depression (vesicle depletion) according to the Tsodyks-Markram formalism[45]:

$$\frac{dx_{ij}^{dep}}{dt} = \frac{1 - x_{ij}^{dep}}{\tau_{rec}} - U x_{ij}^{dep} \sum_{sp} \delta(t - t_{sp}^i - t_{ij}) \qquad (4)$$

The fraction of synaptic resources available at each synapse x_{ij}^{dep} is depleted by a synaptic utilization factor U with each spike transmission and recovers with τ_{rec} back towards its maximum value of 1. Every presynaptic input spike (at t_{sp}^i with transmission delay t_{ij}) thus 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 ($\tau^{AMPA} \ll \tau^{NMDA}$).

$$g_{ij}^{syn}(t) = x_{ij}^{dep}(t) w_{ij}^{syn} e^{-\frac{t - t^i - t_{ij}}{\tau^{syn}}} H(t - t_{sp}^i - t_{ij}) \qquad (5)$$

$H(\cdot)$ denotes the Heaviside step function, and w_{ij}^{syn} is the peak amplitude of the conductance transient, learned by the following *Spike-based BCPNN Learning Rule*.

Spike-based BCPNN Learning Rule

Plastic AMPA and NMDA synapses are modeled to mimic short-term potentiation (STP)[12] with a spike-based version of the Bayesian Confidence Propagation Neural Network (BCPNN) learning rule[44,46]. For a full derivation from Bayes rule, deeper biological motivation, and proof of concept, see Tully et al. (2014) and the earlier STM model implementation[9].

Briefly, the BCPNN learning rule makes use of biophysically plausible local traces to estimate normalized pre- and post-synaptic firing rates, as well as co-activation, which can be combined to implement Bayesian inference because connection strengths and MC activations have a statistical interpretation[44,47,48]. Crucial parameters include the synaptic activation trace Z, which is computed from spike trains via pre- and post-synaptic time constants $\tau_{z_i}^{syn}, \tau_{z_j}^{syn}$, which are the same here but differ between AMPA and NMDA synapses:

$$\tau_{z_i}^{AMPA} = \tau_{z_j}^{AMPA} = 5ms, \;\; \tau_{z_i}^{NMDA} = \tau_{z_j}^{NMDA} = 100ms \qquad (6)$$

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 coactivation:

$$\tau_p \frac{dP_i}{dt} = \kappa(Z_i - P_i), \qquad \tau_p \frac{dP_j}{dt} = \kappa\left(Z_j - P_j\right), \qquad \tau_p \frac{dP_{ij}}{dt} = \kappa\left(Z_i Z_j - P_{ij}\right) \quad (7)$$

These traces constitute memory itself and decay in a palimpsest fashion. STP decay is known to take place on timescales that are highly variable and activity dependent[29]; see Discussion – The case for Hebbian plasticity.

We make use of the learning rule parameter κ (**Equation 7**), which may reflect the action of endogenous neuromodulators, e.g. dopamine acting on D1 receptors, that signal relevance and thus modulate learning efficacy. It can be dynamically modulated to switch off learning to fixate the network, or temporarily increase plasticity (κ_p, κ_{normal}, **Supplementary Table 1**). In particular, we trigger a transient increase of plasticity concurrent with external stimulation.

Tully et al.[44] show 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 is derived from the post-synaptic firing rate estimate p_j and implemented in the NEST neural simulator[49] as an individual neural current I_{β_j} with scaling constant β_{gain}

$$I_{\beta_j} = \beta_{gain} \log(P_j) \quad (8)$$

I_{β_j} is thus an activity-dependent intrinsic membrane current to the neurons, similar to the A-type K+ channel [50] or TRP channel [51]. Synaptic weights are modeled as peak amplitudes of the conductance transient (**Equation 5**) and determined from the logarithmic BCPNN weight, as derived from the P-traces with a synaptic scaling constant w_{gain}^{syn}.

$$w_{ij}^{syn} = w_{gain}^{syn} \log \frac{P_{ij}}{P_i P_j} \qquad (9)$$

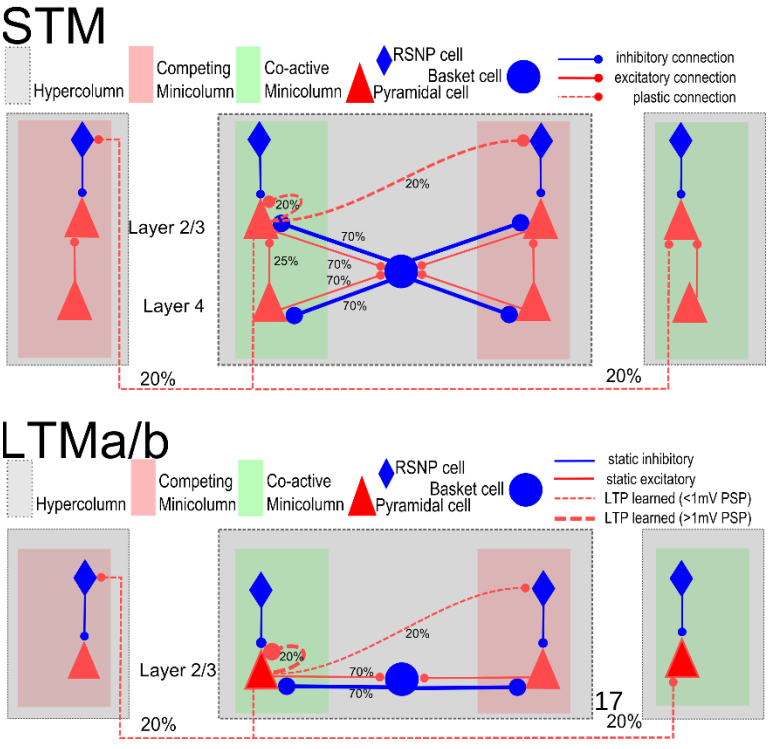
In our model, AMPA and NMDA synapses make use of w_{gain}^{AMPA} and w_{gain}^{NMDA} respectively. The logarithm in **Equations 8,9** is motivated by the Bayesian underpinnings of the learning rule, and means that synaptic weights w_{ij}^{syn} multiplex both the learning of excitatory and di-synaptic inhibitory interaction. The positive weight component is here interpreted as the conductance of a monosynaptic excitatory pyramidal to pyramidal synapse (**Supplementary Figure 1**, plastic connection to the co-activated MC), while the negative component (**Supplementary Figure 1**, 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[52–55]. Accordingly, BCPNN connections with a negative weight use a GABAergic reversal potential instead, as in previously published models[9,17,44]. Model networks with negative synaptic weights have been shown to be functionally equivalent to ones with both excitatory and inhibitory neurons with only positive weights[56].

Code for the NEST implementation of the BCPNN synapse is openly available (see **Simulation Environment**).

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 15% of the mean in order to account for individual arborization differences. 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:

$$\overline{t_{ij}} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{V} + t_{min}^{syn}\ ms \qquad t_{ij} \sim N(\overline{t_{ij}},\ .15\overline{t_{ij}}) \qquad (10)$$



Supplementary Figure 1. Local columnar connectivity within STM and LTM. Connection probabilities are given by the percentages, further details in **Supplementary Tables 1-3**. The strength of plastic connections develops according to the synaptic learning rule described in *Spike-based Bayesian Learning*. Initial weights are low and distributed by a noise-based initialization procedure (*Stimulation protocol*). LTM however, dashed connections are not plastic in LTM (besides the STD of **Equation 4**), but already encode memory patterns previously learned through an LTP protocol, and loaded before the simulation using receptor-specific weights found in **Supplementary Table 2**.

STM Network Architecture

We simulate $n_{HC}^{STM} = 25$ HCs on a grid with spatial extent of 17x17 mm. This spatially distributed network of columns has sizable conduction delays due to the distance between columns and can be interpreted as a spatially distributed subsampling of columns from the extent of dorsolateral PFC (such as BA 46 and 9/46, which also have a combined spatial extent of about 289 mm² in macaque).

Each of the non-overlapping HCs has a diameter of about 640 µm, comparable to estimates of cortical column size[57], contains 24 basket cells, and its pyramidal cell population has been divided into twelve functional columns (MC). This constitutes another sub-sampling from the roughly 100 MC per HC when mapping the model to biological cortex. We simulate 20 pyramidal neurons per MC to represent roughly the layer 2 population of an MC, 5 cells for layer 3A, 5 cells for layer 3B, and another 30 pyramidal cells for layer 4, as macaque BA 46 and 9/46 have a well-developed granular layer[58]. The STM model thus contains about 18.000 simulated pyramidal cells in four layers (although layers 2, 3A, and 3B are often treated as one layer 2/3).

STM Network Connectivity

The most relevant connectivity parameters are found in **Supplementary Tables 1-3**. 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 [59]. 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.

Pyramidal cells in layer 2/3 form connections both within and across HCs at connection probability $p_{L23e-L23e}$. These projections are implemented with plastic synapses and contain parallel AMPA and NMDA components, as explained in subsection *Spike-based BCPNN Learning Rule.* Connections across columns and areas may feature sizable conduction delays due to the implied spatial distance between them (**Supplementary Table 1**)

Pyramidal cells in layer 4 project to pyramidal cells of layer 2/3, targeting 25% of cells within their respective MC only. Experimental characterization of excitatory connections from layer 4 to layer 2/3 pyramidal cells have confirmed similarly high fine-scale specificity in rodent cortex[60] and in-turn, full-scale cortical simulation models without functional columns have found it necessary to specifically strengthen these connections to achieve defensible firing rates[61].

In summary, the STM model thus features a total of 16.2 million plastic AMPA- and NMDA-mediated connections between 18.000 simulated pyramidal cells in STM, as well as 67.500 static connections

from 9.000 layer 4 pyramidals to layer 2/3 targets within their respective MC, and 604.800 static connections to and from 600 simulated basket cells.

LTM network

We simulate two structurally identical LTM networks, referred to as LTMa, and LTMb. LTM networks may be interpreted as a spatially distributed subsampling of columns from areas of the parieto-temporal cortex commonly associated with modal LTM stores. For example Inferior Temporal Cortex (ITC) is often referred to as the storehouse of visual LTM[62]. Two such LTM areas are indicated in **Figure 1.**

We simulate $n_{HC}^{LTM} = 16$ HCs in each area and nine MC per HC (further details in **Supplementary Tables 1-3**). Both LTM networks are structurally very similar to the previously described STM, yet they do not feature plasticity beyond short-term dynamics in the form of synaptic depression. Unlike STM, LTM areas also do not feature an input layer 4, but are instead stimulated directly to cue the activation of previously learned long-term memories (*Stimulation Protocol*). Various previous models with identical architecture have demonstrated how attractors can be learned via plastic BCPNN synapses[9,17,44,63]. We load each LTM network with nine orthogonal attractors (ten in the example of **Figure 3-B**, which features two sets of five memories each). Each memory pattern consists of 16 active MCs, distributed across the 16 HCs of the network. We load-in BCPNN weights from a previously trained network (**Supplementary Table 2**), but thereafter set $\kappa = 0$ to deactivate plasticity of recurrent connections in LTM stores.

In summary, the two LTM models thus feature a total of 7.46 million connections between 8.640 pyramidal cells, as well as 13.608 static connections to and from 576 basket cells.

Corticocortical Connectivity

Our model implements supragranular feedforward and feedback pathways, as inspired by recent characterizations of such pathways by Markov et al.[64] between cortical areas that are at a medium distance in the cortical hierarchy. The approximate cortical distance between Inferior Temporal Cortex (ITC) and dlPFC in macaque is about 40 mm and with an axonal conductance speed of 2 m/s, distributed conduction delays in our model (**Equation 9**) average just above 20 ms between these areas[65–67].

In the forward path, layer 3B cells in LTM project towards STM (**Figure 1**). We do not draw these connections one-by-one, but as branching axons targeting 25% of the pyramidal cells in a randomly chosen MC (the chance of any layer 3B cell to target any MC in STM is only 0.15%). The resulting split between targets in layer 2/3 and 4 is typical for feedforward connections at medium distances in the cortical hierarchy[64] and has important functional implications for the model (*LTM-to-STM Forward Dynamics*). To increase the information contrast in the forward response and balance the total current delivered to STM we also branch off some inhibitory corticocortical connections as follows: for every excitatory connection within the selected targeted MC, an inhibitory connection is created from the same pyramidal layer 3B source cell onto a randomly selected cell outside the targeted MC, but inside the local HC. This is best understood as di-synaptic inhibition via a vertically projecting inhibitory interneuron like a double bouquet and/or bipolar cell[52–55]. Although we do not explicitly simulate such cells, such an interneuron would be local to an MC and targeted by incoming excitatory connections (same arrangement as Tully et al.[17,44]). Simultaneous inputs add in non-trivial ways, as

excitation and inhibition from several inputs can interfere with each other. This way of drawing random forward-projections retains a degree of functional specificity due to its spatial clustering and yields patchy sparse forward-projections with a resulting inter-area connection probability of only 0.0125% (648 axonal projections from L3B cells to STM layers 2/3 and 4 results in ~20k total connections after branching as described above.

Long feedback pathways across the cortical hierarchy are dominated by infra-granular projections (projections from large cells in layer 5 and 6), yet especially between association cortices and at short and medium range there are reciprocal projections from layer 3A cells to the cortical areas below[64]. In our model we draw sparse plastic connections from layer 3A cells in STM to layer 2/3 cells in LTM: branching axons target 25% of the pyramidal cells in a randomly chosen HC in LTM, again simulating a degree of axonal branching found in the literature[68]. Using this method, we obtain biologically plausible sparse and structured feedback projections with an inter-area connection probability of 0.66%, which – unlike the forward pathway – do not have any built-in MC-specificity but may develop such through activity-dependent plasticity. More parameters on corticocortical projections can be found in **Supplementary Table 3.** On average, each LTM pyramidal cell receives about 120 corticocortical connections from STM. Because about 5% of STM cells fire together during memory reactivation, this means that a mere 6 active synapses per target cell are sufficient for driving (and thus maintaining) LTM activity from STM (there are 96 active synapses from coactive pyramidal cells in LTM).

Notably LTMa and LTMb have no direct pathways connecting them in our model since plasticity of biological connections are likely too slow (LTP timescale) to make a difference in WM dynamics. This arrangement also guarantees that any binding of long-term memories across LTM areas must be the result of interaction via STM instead. Corticocortical connectivity is very sparse, below 1% total network connectivity.

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, two independent Poisson sources generate spikes at rate r_{bg}^{layer} , and connect onto all pyramidal neurons in that layer, via non-depressing conductances $\pm g_{bg}^{layer}$ (**Supplementary Table 2**). Before each simulation, we distribute the initial values of all plastic weights in the process of learning induced by 1.5 s low background activity (**Supplementary Table 2**, $\boldsymbol{r_{bg-low}^{L23}}$). To cue the activation of a specific memory pattern (i.e. attractors), we excite LTM pyramidal cells belonging to a memory patterns component MC with an additional excitatory Poisson spike train (rate r_{cue}, length t_{cue}, conductance g_{cue}). As LTM patterns are strongly encoded in each LTM, a brief 50 ms stimulus is usually sufficient to activate any given memory.

Spike Train Analysis and Memory Activity Tracking

We track memory activity in time by analyzing the population firing rate of pattern-specific and network-wide spiking activity usually using an exponential moving average filter time-constant of 20 ms. We do not use an otherwise common low-pass filter with symmetrical window, because we are particularly interested in characterizing activation onsets and onset delays. As activations are

characterized by sizable gamma-like bursts, a simple threshold detector can extract candidate activation events and decode the activated memory. This is trivial in LTM due to the orthogonal and known nature of its patterns. In STM we decode the stimulus-specificity of each cell individually by finding the maximum correlation between input pattern and the untrained STM spiking response in the 320 ms (which is the stimulation interval during plasticity-modulated stimulation period, shown in e.g. **Figure 2D**) following the pattern cue to LTM. Thereafter we can filter the population response of in STM with the same selectivity on that basis to obtain a more robust readout. We validate the specificity by means of cross-correlations, which reveal that the pattern specific populations are rather orthogonal according to the covariance matrix (off-diagonal magnitude < 0.1). In all three networks, we measure onset and offset of pattern activity by thresholding each individual activation at half of its population peak firing rate. In LTM, we further check pattern completion by analyzing component MC activation. Whenever targeted stimuli are used, we analyze peri-stimulus activation traces. When activation onsets are less predictable, such as during free STM-paced maintenance, we extract activation candidates via a threshold detector trained at the 50th percentile of the cumulative distribution of the population firing rate signal.

Simulation Environment

We use the NEST simulator[49] 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 available on github: https://github.com/Florian-Fiebig/BCPNN-for-NEST222-MPI

Supplementary Tables

Adaptation current	b	86 pA	Depression time constant	τ_{rec}	500 ms	BCPNN AMPA gain	w_{gain}^{AMPA}	3.93 nS
Adaptation time constant	τ_{I_w}	500 ms	AMPA synaptic time constant	τ^{AMPA}	5 ms	BCPNN NMDA gain	w_{gain}^{NMDA}	0.21 nS
Membrane Capacity	C_m	280 pF	NMDA synaptic time constant	τ^{NMDA}	100 ms	BCPNN bias current gain	β_{gain}	90 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	14 pS	AMPA Reversal Potential	E^{AMPA}	0 mV	BCPNN highest rate	f_{max}	20 Hz
Upstroke slope factor	Δ_T	3 mV	NMDA Reversal Potential	E^{NMDA}	0 mV	BCPNN lowest probability	ε	0.01
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	-80 mV	Dopaminergic Modulation	κ_p	6.0	P-Trace time constant	τ_p	5 s
Utilization factor	U	.33	Regular Plasticity	κ_{normal}	1.0			

Supplementary Table 1. Neurons, synapses, and plasticity.

STM patch size	17 x 17 mm		Initialization Input rate layer 2/3	r_{bg-low}^{L23}	550 Hz
Simulated HCs	n_{HC}^{STM}	25	Background activity rate layer 2/3	r_{bg}^{L23}	625 Hz
Simulated MC per HC	n_{MC}^{STM}	12	Background activity rate layer 4	r_{bg}^{L4}	300 Hz
LTM patch size	25 x 25 mm		High Background activity rate layer 2/3 (e.g. STM Maintainance)	$r_{bg-high}^{L23}$	950 Hz
Simulated HCs	n_{HC}^{LTM}	16			
Simulated MC per HC	n_{MC}^{LTM}	9	Background conductance	g_{bg}	±1.5 nS
Axonal Conduction Speed	V	$2\frac{m}{s}$			
Minimal conduction delay	t_{min}^{syn}	1.5 ms	Cue stimulus duration	t_{cue}	50 ms
STM – LTM distance	$d_{STM-LTM}$	40 mm	Stimulation rate	r_{cue}	650 Hz
Hypercolumn diameter	d_{HC}	0.64 mm	Cue stimulus conductance	g_{cue}	+1.5 nS
Layer 2 pyramidal per MC	n_{MC}^{PYR-L2}	20	LTM Intra HC – Intra MC weight	$w_{IntraMC}^{IntraHC}$	3.36 w_{gain}^{syn}
Layer 3A pyramidal per MC	$n_{MC}^{PYR-L3A}$	5	LTM Intra HC – Inter MC weight	$w_{InterMC}^{IntraHC}$	-4.82 w_{gain}^{syn}
Layer 3B pyramidal per MC	$n_{MC}^{PYR-L3B}$	5			
Layer 4 pyramidal per MC	n_{MC}^{PYR-L4}	30	LTM Inter HC –Coactive MC weight	$w_{CoactiveMC}^{InterHC}$	3.08 w_{gain}^{syn}
Basket cells per MC	n_{MC}^{basket}	2	LTM Inter HC – Competing MC weight	$w_{CompetingMC}^{InterHC}$	-4.28 w_{gain}^{syn}

Supplementary Table 2. Network size, Conduction delay, Stimulation, LTM Preload BCPNN weights. Layer 4 not simulated in LTM.

Scope	Source	Target	Type	Symbol	Value
Cortical Area	Pyramidal	Basket	probability	p_{P-B}	0.7
	Pyramidal	Basket	condnductance (static)	g_{P-B}	+3.5 nS
	Basket	Pyramidal	probability	p_{B-P}	0.7
	Basket	Pyramidal	conductance (static)	g_{B-P}	-40 nS
	L23e	L23e	probability	$p_{L23e-L23e}$	0.2
	L23e	L23e	AMPA gain (BCPNN)	w_{gain}^{AMPA}	3.93nS
	L23e	L23e	NMDA gain (BCPNN)	w_{gain}^{NMDA}	0.21nS
	L4e	L23e	probability	$p_{L4e-L23e}$	0.25
	L4e	L23e	conductance (static)	$g_{L4e-L23e}$	25 nS
Feed forward	LTM L3Ae	STM MC	probability	$p_{L3Ae-MC}^{FF}$	0.0015
	LTM L3Ae	STM MC	branching factor	$b_{L3Ae-MC}^{FF}$	0.25
	LTM L3Ae	STM L23e	conductance (static)	$g_{L3Ae-L23e}^{FF}$	±7.2 nS
	LTM L3Ae	STM L4e	conductance (static)	$g_{L3Ae-L4e}^{FF}$	±7.2 nS
Feedback	STM PYR	LTM PYR	probability	p_{P-P}^{FB}	0.0066
	STM L3Be	LTM HC	branching factor	$b_{L3Be-HC}^{FB}$	0.25
	STM L3Be	LTM L23e	AMPA gain (BCPNN)	w_{FB}^{AMPA}	7.07 nS
	STM L3Be	LTM L23e	NMDA gain (BCPNN)	w_{FB}^{NMDA}	0.4 nS

Supplementary Table 3. Projections

Model Robustness

Our model incorporates a plethora of biological constraints, such as estimates on the extent and distance of areas (e.g. STM patch size approximates macaque dlPFC, and is 40mm from ITC), laminar cell distributions (n_{MC}^{PYR-L2}, $n_{MC}^{PYR-L3b}$,…), hypercolumnar size, etc. The model also abides by various electrophysiological constraints, such as plausible EPSP, IPSP sizes, estimates on laminar connection densities, characterization of cortical FF/FB pathways, estimates on axonal conductance speeds, dendritic arbor sizes (branching factors), commonly accepted synaptic time-constants for various receptor types, depression, adaptation, and builds on top of established models we adapted, such as

the neuron model or the synaptic resource model. References to many of these constraints can be found throughout the Method Section.

Because our model is quite complex and synthesizes many different components and processes it is beyond the scope of this work to perform a detailed parameter sensitivity analysis. However, from our extensive simulations we conclude that it is robust and degrades gracefully. Almost all uncertain parameters can be varied ±30% without breaking WM function. The model is dramatically subsampled and scaling up would be possible. This could be expected to further improve overall robustness. Highly related modular cortical network models have been studied extensively elsewhere[9,20,26,44,69], so here we prioritize novel aspects, namely the parameterization of corticocortical connectivity and spatial scale.

In the feedback pathway, a mere 0.6% connectivity is sufficient to support LTM activation in maintenance and recall. As rigorous testing (not shown here) revealed, lower connectivity degrades WM capacity, unless we increase the total number of co-active STM cells by other means. Forward connectivity can be even lower (0.015% in this model), because terminal clusters in STM are smaller and provide more information contrast (*Corticocortical Connectivity*). In both cases, our model uses these low density values, but they could be increased or decreased if single synaptic currents are reduced/increased respectively. Somewhat peculiarly, we also found that we needed to increase the corticocortical conductance of the backprojections (w_{FB}^{syn}) by the same factor 1.8 (over the local conductance gain w_{gain}^{syn}) as another detailed model account of macaque visual cortex[70] to achieve functional WM at the stated long-distance connection probabilities.

There is an upper, but no lower limit on corticocortical distances in our model. When conduction delays exceed 65 ms (130 mm), STM feedback can no longer activate the LTM network, because bursts desynchronize before they arrive. On the other hand, STM and LTM could even be adjacent as we briefly mentioned at the end of the result section. Additionally, there is a minimum spatial scale to each component network. If we reduce the spatial extent (and thus the connection delays between HCs) by 45%, theta-like oscillations degrade and break at 20%, when the largest inter-HC delays fall below 5 ms. Spiking activity of activated memories collapses into a single brief burst (**Supplementary Figure 8,** cf. **Figure 2D, Supplementary Figure 7**), which degrades learning and effective information transmission both within and across networks. Networks may be much smaller however, if this is compensated by slower axonal conductance velocities (<2 mm/ms).

References

1. Slifstein, M. *et al.* Deficits in prefrontal cortical and extrastriatal dopamine release in schizophrenia a positron emission tomographic functional magnetic resonance imaging study. *JAMA Psychiatry* **72,** 316–324 (2015).

2. Fuster, J. M. Cortex and Memory: Emergence of a New Paradigm. *J. Cogn. Neurosci.* **21,** 2047–2072 (2009).

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

4. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61,** 331–349 (1989).

5. Goldman-Rakic. Cellular Basis of Working Memory Review. *Neuron* **14,** 477–485 (1995).

6. Camperi, M. & Wang, X. J. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *J. Comput. Neurosci.* **5,** 383–405 (1998).

7. Compte, A. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cereb. Cortex* **10,** 910–923 (2000).

8. Mongillo, G., Barak, O. & Tsodyks, M. SynaptiC Theory of Working Memory. *Science (80-.).* **319,** 1543–1546 (2008).

9. Fiebig, F. & Lansner, A. A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *J. Neurosci.* **37,** 83–96 (2017).

10. Teyler, T. J. & DiScenna, P. The hippocampal memory index theory. *Behav. Neurosci.* **100,** 147–154 ST–The hippocampal memory index theory (1986).

11. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus* **17,** 1158–1169 (2007).

12. 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).

13. Park, P. *et al.* NMDA receptor-dependent long-term potentiation comprises a family of temporally overlapping forms of synaptic plasticity that are induced by different patterns of stimulation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369,** 20130131 (2014).

14. Cer, D. M. & O'Reily, R. C. in *Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience* (eds. Zimmer, H. D., Mecklinger, A. & Lindenberger, U.) 193–220 (Oxford University Press, USA, 2012). doi:10.1093/acprof:oso/9780198529675.003.0008

15. van der Velde, F. & de Kamps, M. The necessity of connection structures in neural models of variable binding. *Cognitive Neurodynamics* **9,** 359–370 (2015).

16. Pinkas, G., Lima, P. & Cohen, S. Representing, binding, retrieving and unifying relational knowledge using pools of neural binders. in *Biologically Inspired Cognitive Architectures* **6,** 87–95 (Elsevier B.V., 2013).

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

18. Lundqvist, M., Rehn, M., Djurfeldt, M. & Lansner, A. Attractor dynamics in a modular network model of

neocortex. *Netw. Comput. Neural Syst.* **17,** 253–276 (2006).

19. Lundqvist, M., Compte, A. & Lansner, A. Bistable, irregular firing and population oscillations in a modular attractor memory network. *PLoS Comput Biol.* **6,** (2010).

20. 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).

21. 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 *Brain Research* **1536,** 68–87 (2013).

22. Mi, Y., Katkov, M. & Tsodyks, M. Synaptic Correlates of Working Memory Capacity. *Neuron* **93,** 323–330 (2017).

23. Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I. & Miyashita, Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401,** 699–703 (1999).

24. Lansner, A. Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. *Trends in Neurosciences* **32,** 178–186 (2009).

25. Silverstein, D. N. & Lansner, A. Is Attentional Blink a Byproduct of Neocortical Attractors? *Front. Comput. Neurosci.* **5,** (2011).

26. Tully, Lindén, H., Hennig, M. H. & Lansner, A. Probabilistic computation underlying sequence learning in a spiking attractor memory network. *BMC Neurosci.* **14,** P236 (2013).

27. Baddeley, A., Gathercole, S. & Papagno, C. The Phonological Loop as a Language Learning Device. *Psychol. Rev.* **105,** 158–173 (1998).

28. Volianskis, A. & Jensen, M. S. Transient and sustained types of long-term potentiation in the CA1 area of the rat hippocampus. *J. Physiol.* **550,** 459–92 (2003).

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

30. Wickens, J. R. Synaptic plasticity in the basal ganglia. *Behavioural Brain Research* **199,** 119–128 (2009).

31. Arnsten, A. F. T. & Jin, L. E. Molecular influences on working memory circuits in dorsolateral prefrontal cortex. *Prog. Mol. Biol. Transl. Sci.* **122,** 211–231 (2014).

32. Goto, Y., Yang, C. R. & Otani, S. Functional and Dysfunctional Synaptic Plasticity in Prefrontal Cortex: Roles in Psychiatric Disorders. *Biological Psychiatry* **67,** 199–207 (2010).

33. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci.* **19,** 394–405 (2015).

34. Johnson, E. L. & Knight, R. T. Intracranial recordings and human memory. *Current Opinion in Neurobiology* **31,** 18–25 (2015).

35. Anderson, K. L., Rajagovindan, R., Ghacibeh, G. A., Meador, K. J. & Ding, M. Theta oscillations mediate interaction between prefrontal cortex and medial temporal lobe in human memory. *Cereb. Cortex* **20,** 1604–1612 (2010).

36. Zimmer, H. D., Mecklinger, A. & Lindenberger, U. *Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience*. *Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience* (Oxford University Press, 2012). doi:10.1093/acprof:oso/9780198529675.001.0001

37. Reynolds, J. H. & Desimone, R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24,** 19–29 (1999).

38. Zimmer, H. D., Mecklinger, A. & Lindenberger, U. in *Handbook of binding and memory: Perspectives from cognitive neuroscience* 3–22 (Oxford University Press, 2006).

39. Pinkas, G., Lima, P. & Cohen, S. A dynamic binding mechanism for retrieving and unifying complex predicate-logic knowledge. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7552 LNCS,** 482–490 (2012).

40. Legenstein, R., Papadimitriou, C. H., Vempala, S. & Maass, W. Variable binding through assemblies in spiking neural networks. in *CEUR Workshop Proceedings* **1773,** (2016).

41. Zmigrod, S., Colzato, L. S. & Hommel, B. Evidence for a role of the right dorsolateral prefrontal cortex in controlling stimulus-response integration: a transcranial direct current stimulation (tDCS) study. *Brain Stimul* **7,** 516–520 (2014).

42. Tallon-Baudry, C. & Bertrand, O. Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences* **3,** 151–162 (1999).

43. Brette, R. & Gerstner, W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. Neurophysiol.* **94,** 3637–42 (2005).

44. Tully, P. J., Hennig, M. H. & Lansner, A. Synaptic and nonsynaptic plasticity approximating probabilistic inference. *Front. Synaptic Neurosci.* **6,** (2014).

45. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl. Acad. Sci.* **94,** 719–723 (1997).

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

47. Sandberg, A., Lansner, A., Petersson, K. M. & Ekeberg, O. A Bayesian attractor network with incremental learning. *Network* **13,** 179–194 (2002).

48. Fiebig, F. & Lansner, A. Memory consolidation from seconds to weeks: a three-stage neural network model with autonomous reinstatement dynamics. *Front. Comput. Neurosci.* **8,** 1–17 (2014).

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

50. 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).

51. Petersson, M. E., Yoshida, M. & Fransén, E. A. Low-frequency summation of synaptically activated transient receptor potential channel-mediated depolarizations. *Eur. J. Neurosci.* **34,** 578–93 (2011).

52. Tucker, T. R. Recruitment of Local Inhibitory Networks by Horizontal Connections in Layer 2/3 of Ferret Visual Cortex. *J. Neurophysiol.* **89,** 501–512 (2002).

53. Ren, M., Yoshimura, Y., Takada, N., Horibe, S. & Komatsu, Y. Specialized inhibitory synaptic actions between nearby neocortical pyramidal neurons. *Science* **316,** 758–61 (2007).

54. Silberberg, G. & Markram, H. Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells. *Neuron* **53,** 735–46 (2007).

55. 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,** 743–53 (2007).

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

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

58. Petrides, M. & Pandya, D. N. Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in

the human and the macaque brain and corticocortical connection patterns. *Eur. J. Neurosci.* **11,** 1011–1036 (1999).

59. Douglas, R. J. & Martin, K. A. C. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27,** 419–51 (2004).

60. Yoshimura, Y. & Callaway, E. M. Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nat. Neurosci.* **8,** 1552–1559 (2005).

61. Potjans, T. C. & Diesmann, M. The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model. *Cereb. Cortex* **24,** 785–806 (2014).

62. Miyashita, Y. Inferior Temporal Cortex: Where Visual Perception Meets Memory. *Annu. Rev. Neurosci.* **16,** 245–263 (1993).

63. Lansner, A., Marklund, P., Sikström, S. & Nilsson, L. Reactivation in Working Memory: An Attractor Network Model of Free Recall. *PLoS One* **8,** e73776 (2013).

64. Markov, N. T. *et al.* Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* **522,** 225–259 (2014).

65. Girard, P., Hupe, J. M. & Bullier, J. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J Neurophysiol* **85,** 1328–31. (2001).

66. Thorpe, S. J. & Fabre-Thorpe, M. Seeking categories in the brain. *Science* **291,** 260–263 (2001).

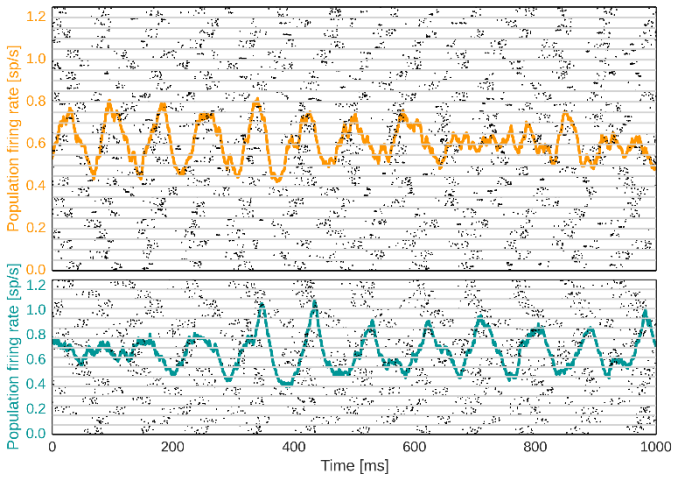
67. Caminiti, R. *et al.* Diameter, Length, Speed, and Conduction Delay of Callosal Axons in Macaque Monkeys and Humans: Comparing Data from Histology and Magnetic Resonance Imaging Diffusion Tractography. *J. Neurosci.* **33,** 14501–14511 (2013).

68. Zufferey, P. D., Jin, F., Nakamura, H., Tettoni, L. & Innocenti, G. M. The role of pattern vision in the development of cortico-cortical connections. *Eur. J. Neurosci.* **11,** 2669–2688 (1999).

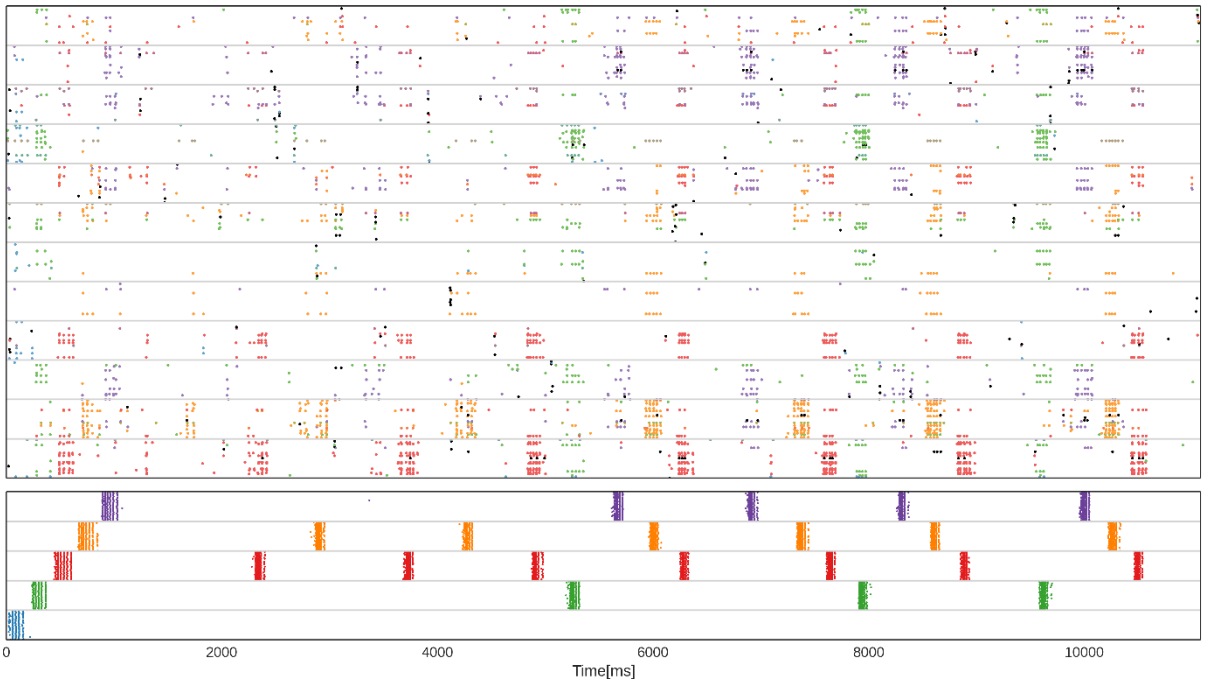
69. Lundqvist, M., Compte, A. & Lansner, A. Bistable, irregular firing and population oscillations in a modular attractor memory network. *PLoS Comput. Biol.* **6,** 1–12 (2010).

70. Schmidt, M. *et al.* Full-density multi-scale account of structure and dynamics of macaque visual cortex. (2015).

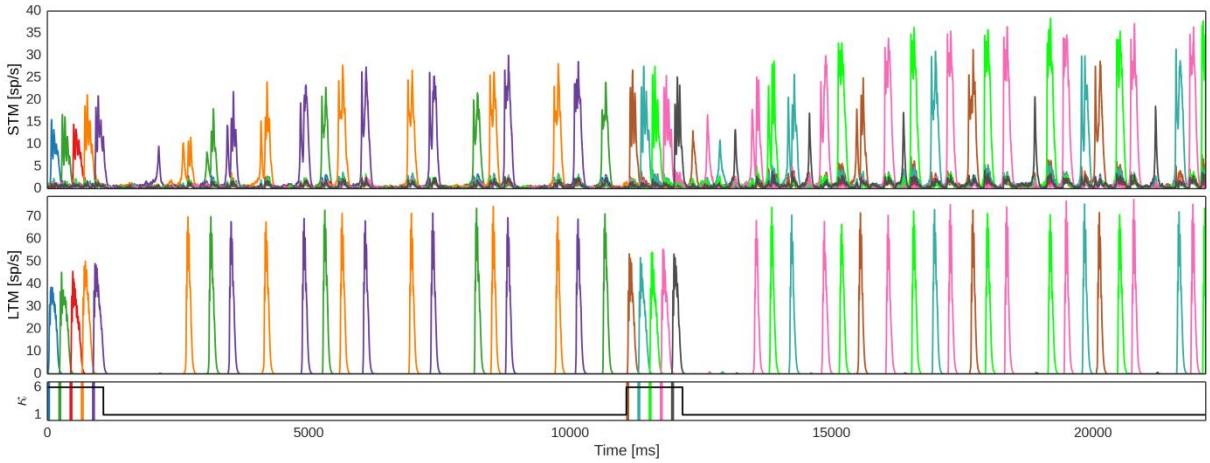
Supplementary Figures



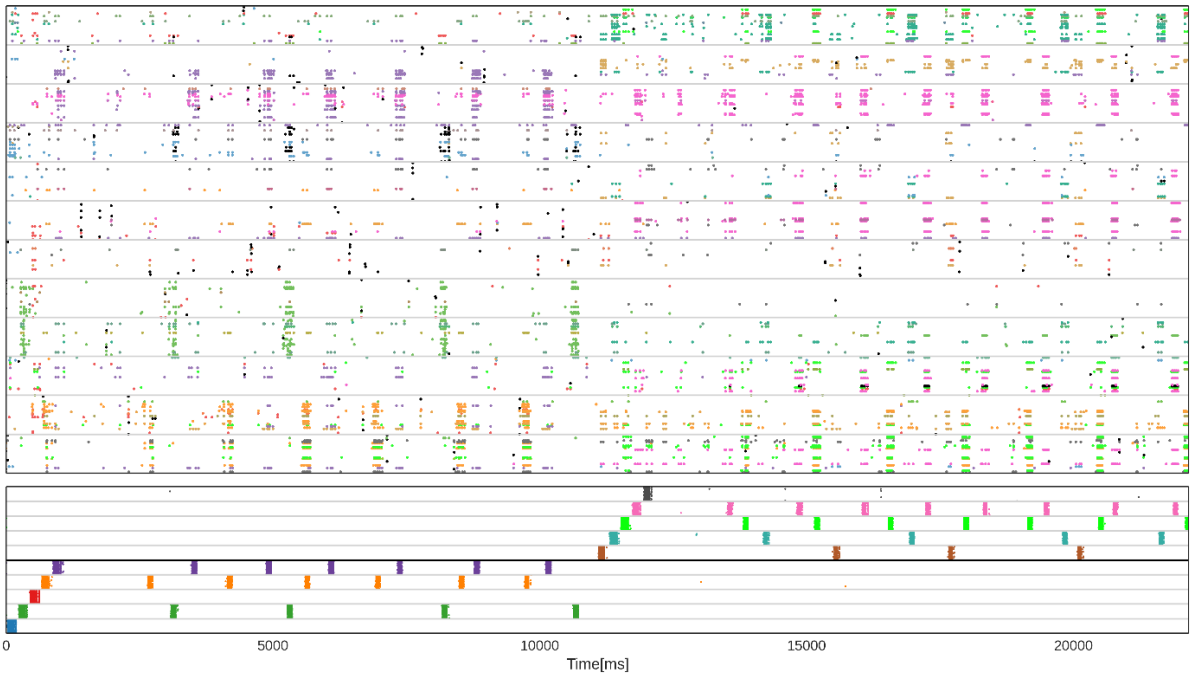
Supplementary Figure 2. Activity in the untrained network under strong background input. Subsampled spike raster of STM (top) and LTM (bottom) layer 2/3 activity. HCs are separated by grey horizontal lines. Global oscillations in the alpha range (10-13 Hz) characterize this activity state in both STM (top) and LTM (bottom) in the absence of attractors.



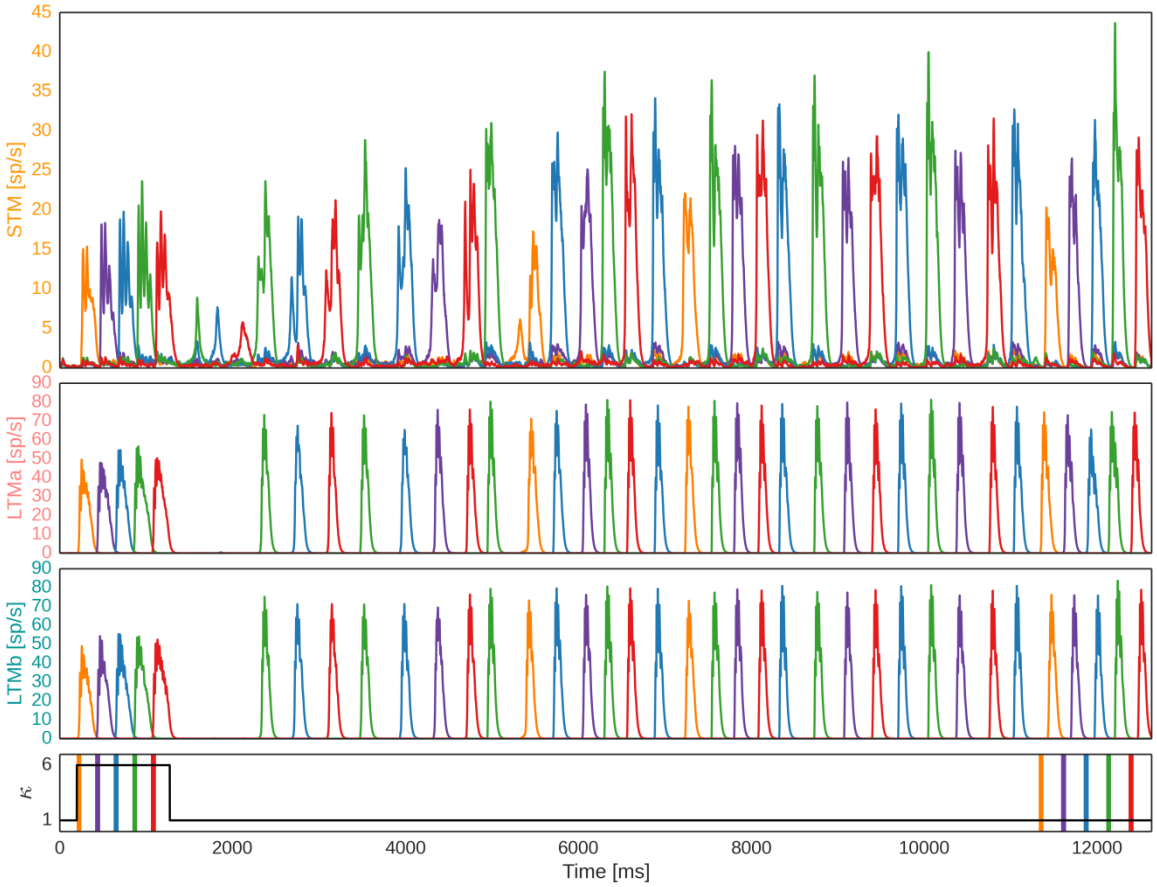
Supplementary Figure 3. Encoding and feedback-driven reactivation of long-term memories. Subsampled spike raster of STM (top) and LTM (bottom) during encoding and subsequent maintenance of five memories (the first pattern is not maintained in this simulation). During the initial plasticity-modulated stimulation phase, five LTM memories are cued via targeted 50 ms stimuli (shown underneath). Plasticity of STM and its backprojections is modulated during this initial memory activation (cf. **Figure2D**). Thereafter, a strong noise drive to STM causes spontaneous activations and plasticity-induced consolidation of pattern-specific subpopulations in STM. Backprojections reactivate associated LTM memories. **Top:** STM spike raster shows layer 2/3 activity in a single HC. MCs are separated by grey horizontal lines. STM spikes are colored according to each cell's dominant LTM pattern-correlation, similar to Figure 2D. **Bottom:** LTM spike raster only shows the activity of five coding MC in a single LTM HC, but indicates the activation of distributed LTM memory patterns. LTM spikes are colored according to the pattern-specificity of each cell.



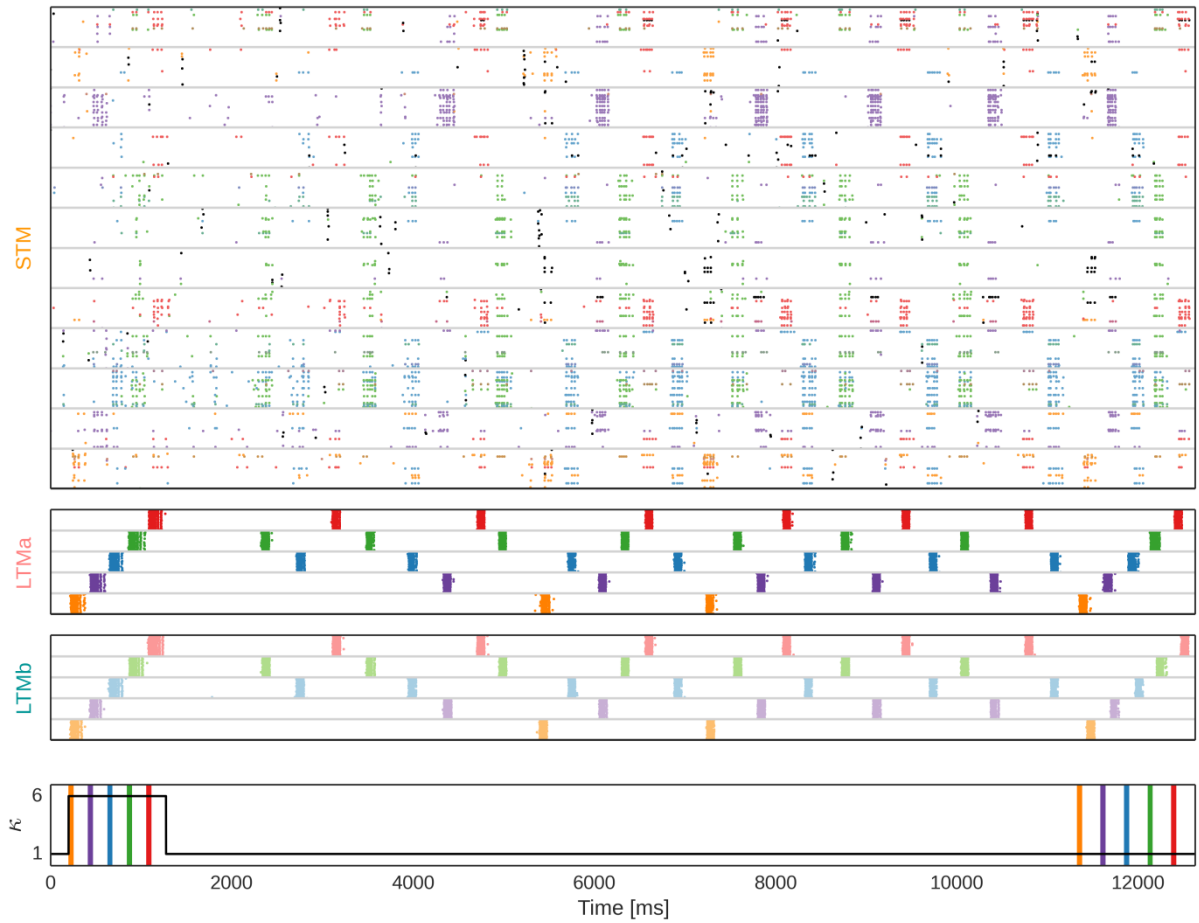
Supplementary Figure 4. WM updating. Population firing rates of pattern-specific subpopulations in STM and LTM during encoding and subsequent maintenance of two sets of five LTM memories. After encoding and 10 s maintenance of the first set, WM contents are overwritten with the second set of memories, maintained thereafter in spontaneous reactivation events. Bottom: Stimuli to LTM and modulation of plasticity.



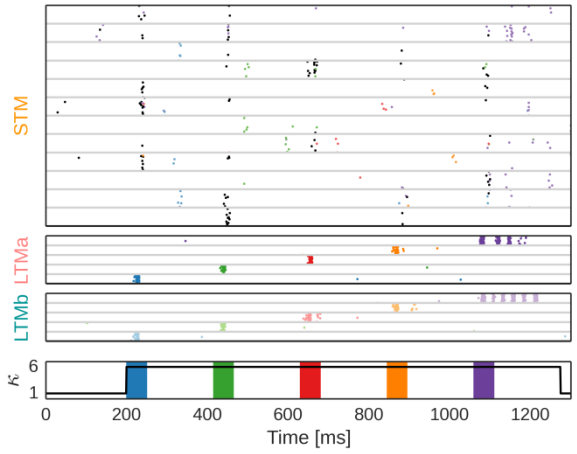
Supplementary Figure 5. WM updating. Subsampled spike raster of the layer 2/3 population in a Hypercolumn of STM (top) and LTM (bottom) respectively during encoding and subsequent maintenance of two sets of five LTM memories. STM spikes are colored according to each cells dominant pattern-selectivity. LTM Spikes are colored according to the pattern-specificity of each cell. After encoding and 10 s maintenance of the first set, WM contents are overwritten with the second set of memories, maintained thereafter. Plasticity is temporarily boosted during the initial activation of LTM attractors (see preceding figure). Strong noise drive to STM causes spontaneous reactivations and consolidation of pattern-specific subpopulations in STM following each stimulation period.



Supplementary Figure 6. Population firing rates of patter-specific subpopulations in the three networks, during the Feature binding task. Three memory pairs (blue, green, red respectivly) in STM (1st row), LTMa (2nd row), and LTMb (3rd row) during three different modes of network activity: The initial binding of pairwise LTM memory activations in STM (0 – 1 s), WM Maintenance (1 s – 10 s), and cue-driven associative recall of previously paired stimuli (11 s – 12 s). Bottom: Stimuli to LTM and modulation of plasticity. Note the cued recall of all five memories after 10sec of maintenance.



Supplementary Figure 7. Spiking activity in the three networks, during the multi-modal LTM binding task. Subsampled spike raster of the layer 2/3 population in a Hypercolumn of STM (top), and five coding minicolumns in LTMa (2nd row) and LTMb (3rd row) respectively during plasticity-modulated stimulation (i.e. encoding), subsequent maintenance, and associative cued recall of five paired LTM patterns (orange,purple,blue,green,red). Minicolumns are separated by grey horizontal lines. STM spikes are colored according to each cells dominant memory pair-selectivity. LTM Spikes are colored according to the memory pair-specificity of each cell in slightly shifted hues to illustrate that LTMa and LTMb code for different, but associated memories. Bottom: Stimuli to LTM and modulation of plasticity. Note the cued recall of all five memories at the end.



Supplementary Figure 8. Network activity during plasticity-modulated stimulation with 20% spatial extent. Subsampled spike raster of the layer 2/3 population in a Hypercolumn of STM (top), and five coding minicolumns in LTMa (2nd row) and LTMb (3rd row) respectively during plasticity-modulated stimulation (i.e. encoding) of five paired LTM patterns. Without sufficient conduction delays, memory activations collapse into very brief bursts (with the exception of the last pattern here) and STM cannot effectively activate from or subsequently encode such brief activations (cf. **Figure 2D**, and **Supplementary Figure 7**).