Fast Hebbian Plasticity explains Working Memory and Neural Binding

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

We have extended a previous spiking neural network model of prefrontal cortex with fast Hebbian plasticity to also include interactions between short-term and long-term memory stores. We investigated how prefrontal cortex could bind and maintain multi-modal long-term memory representations by simulating three cortical patches in macaque brain, i.e. in prefrontal cortex together with visual and auditory temporal cortex.

Our simulation results demonstrate how simultaneous brief multi-modal activation of items in long-term memory could build a temporary joint cell assembly linked via prefrontal cortex. 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 short-term memory network flexibly updates as new stimuli arrive thereby gradually over-writing older representation. In a wider context, this working memory model suggests a novel solution to the "binding problem", a long-standing and fundamental issue in cognitive neuroscience.

Introduction

By working memory (WM), we typically understand a flexible but volatile kind of memory capable of holding a small number of items over shorter time span. This allows acting beyond the immediate here and now and bringing on-line task relevant information in long-term memory (LTM). WM is a key component in cognition and often affected early on in neurological and psychiatric conditions, like e.g. Alzheimers disease and schizophrenia[1]. Prefrontal cortex (PFC) has been repeatedly implicated as a key structure for WM in humans and non-human primates[2,3].

The prevalent theory of WM for thirty years has been that of persistent activity in PFC recurrent networks, proposed by Goldman-Rakic and colleagues[4,5] and supported by computational models[6,7]. Although there is solid support for delay activity in PFC during WM, recent experimental observations of e.g. the variability of this activity[8,9] and the observation of "activity silent working memory"[10] have cast doubt on persistent activity as the main neural mechanism underlying WM.

Proposals for and simulations of synaptic plasticity as a main neural mechanism behind WM were made earlier using abstract attractor neural networks with non-spiking units[11–15]. In 2008 Mongillo et

al.[16] suggested a "synaptic working memory theory" and demonstrated a spiking model of a piece of cortex which displayed WM in the form of self-sustained bursty delay activity, thus also allowing single WM memory items to remain silent for short periods of time. However, this model was based on non-Hebbian forms of synaptic plasticity where synaptic weight changes have no dependence on postsynaptic activity. Such processes can only facilitate memory items already encoded in a pre-structured synaptic matrix of sufficient stimulus-specificity, which precludes encoding of genuinely novel memories.

Building on an earlier abstract model[18], we recently modeled a human word-list learning task using a spiking short-term memory (STM) model of PFC based on fast Hebbian synaptic plasticity[19]. We demonstrated our model's capability to encode novel items, to quantitatively reproduce key characteristics of human cued and non-cued recall, and to reproduce several important network dynamic features of memory replay and recall. To function in this context, synaptic plasticity needs to operate on a time-scale of a few hundred milliseconds. Although processes like LTP has typically been seen as a slowly inducing and expressing, early[20] as well as more recent research has shown the existence also of faster forms, e.g. Short-Term Potentiation (STP)[21,22], which opens up the possibility for this type of WM model.

In the following we present an extended WM model which integrates STM in PFC[19] with network components representing multi-modal LTM areas. Thus, this new model factors in that STM does not operate in isolation to support WM but interacts closely with processes like attention and LTM. For instance, unattended stimuli rarely end up in WM or get consolidated in LTM, so there is an "attentional gate" that controls that only relevant items get encoded[23]. Furthermore, WM is often described as activating or bringing LTM representations on-line[24]. Our extended model allows for a quantitative analysis of such key interactions between STM and LTM.

Results

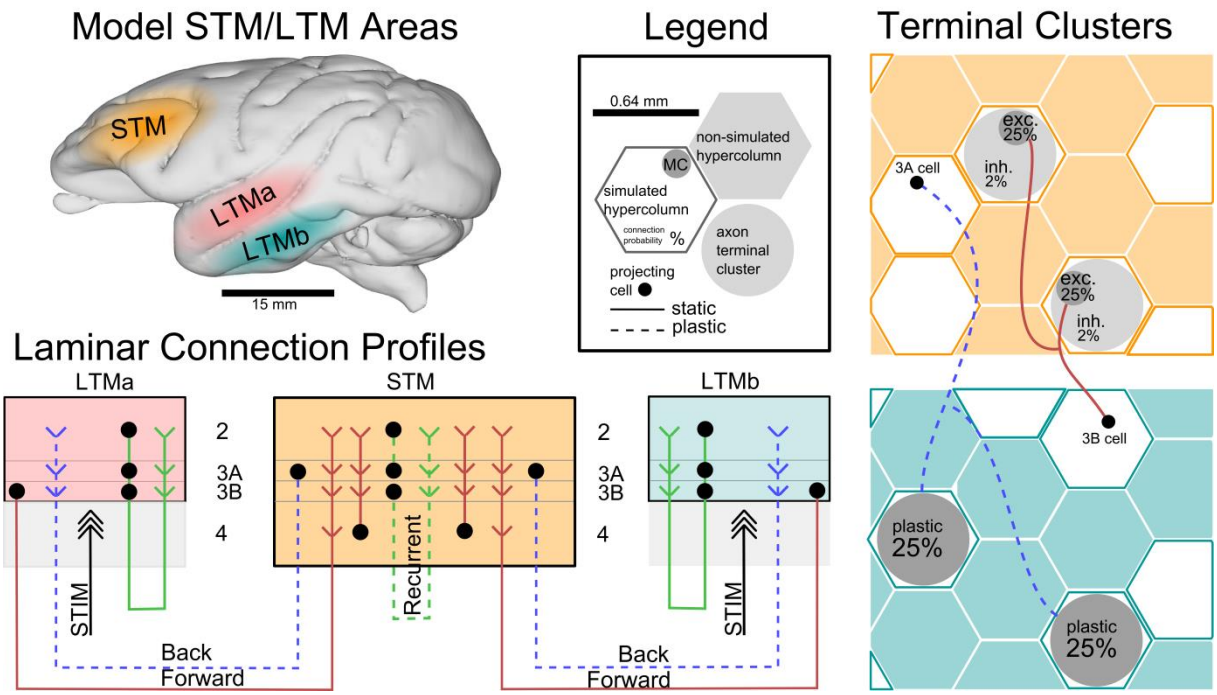


Figure 1. Schematic of modeled connectivity within and across representative STM and LTM areas in macaque. The model organizes cells into grids of nested hypercolumns (HC) and minicolumns (MC), sometimes referred to as macro columns, and "functional columns" respectively. Each network spans several hundred mm² and the simulated columns constitute a spatially distributed subsample, which is important due to the implied conduction delays. Pyramidal cells in the simulated supragranular layers form connections both within and across columns of their respective networks. Unlike STM, LTM areas do not feature an input layer 4, but are instead stimulated directly to cue the activation of previously learned long-term memories. Additional corticocortical connections are sparse (<1% connection probability) and implemented with terminal clusters and specific laminar connection profiles. The connection schematic illustrates laminar connections realizing a direct supragranular forward-projection, as well as a common supragranular back-projection. Layer 2/3 recurrent connections in STM and corticocortical backprojections feature fast Hebbian plasticity. For an in-depth model description, including the columnar microcircuits, please refer to *Online Methods* and **Supplementary Figure 1**.

We simulate three cortical patches, i.e. an STM and two slightly smaller LTM networks (LTMa, LTMb) representing PFC and temporal visual and auditory cortical areas respectively (**Figure 1**). The computational network model used here represents a modular cortical microcircuit architecture in line with previous models[25,26]. In particular, the current model is an extension of a recent STM model[19]. The abstract, sub-sampled 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 to allow for the study of memory binding and interaction. This composite model synthesizes many different kinds of cortical data and produces complex output dynamics. It is beyond the scope of this work to explore and analyze all its properties in detail. We provide references to particularly interesting aspects that have been intensively studied in preceding related models. Additional model details are found in the *Online Methods*.

We introduce the performance of the corticocortical model in several steps. First, we take a brief look at ground and attractor states in the isolated networks. Second, we describe the effect of active LTM memories on STM with- and without plasticity. Third, we add the plastic backprojections from STM to LTM and monitor the encoding and propagation of several memories in the resulting closed cortical loop. We track the evolution of acquired memory indices in STM and their driving role in WM maintenance (aka. delay activity). We then demonstrate that the emerging WM network system stays

plastic and capable of updating its maintained set of memories. Finally, we stimulate several unconnected (i.e. multi-modal) pairs of memories in parallel and study how they flexibly bind in STM. We explore temporal characteristics of network activations during WM encoding, maintenance, and cue-driven associative recall of memory pairs associated in this manner. Our simulation output then allows us to analyze cross-cortical activation delay distributions and compare these with biological data on bottom-up and top-down response delays in a related study by Tomita et al.[27].

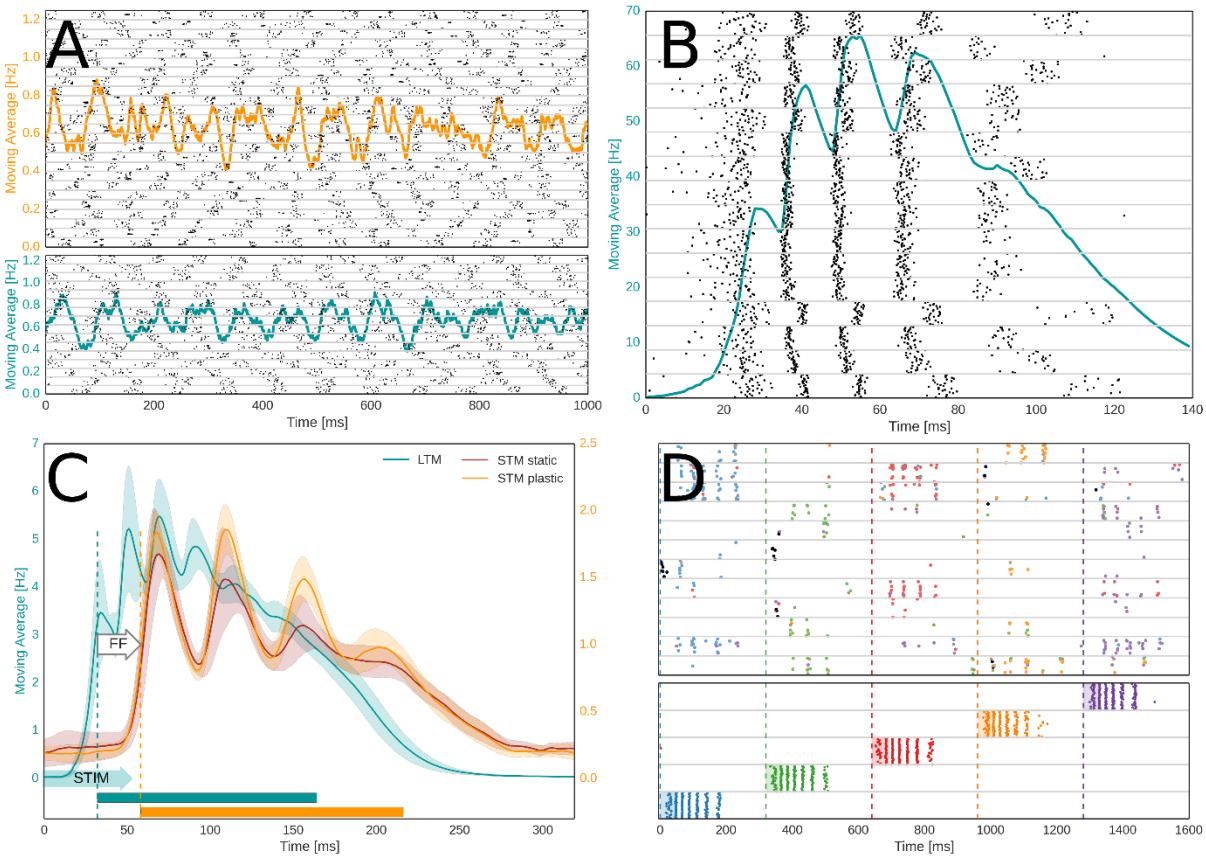


Figure 2. Basic Network behavior in spike rasters and exponential moving averages. A: Alpha/beta oscillations characterize the ground state of both STM (top) and LTM (bottom) in the absence of attractor activity. The underlying spike raster shows layer 2/3 activity of one MC from each HC (separated by grey horizontal lines) in the simulated network, revealing weak transient spatial synchronization of activity. **B:** LTM attractor activations express as theta-nested gamma bursts. The underlying spike raster shows layer 2/3 activity of the activated MC in each HC, revealing spatially-dependent, transient synchronization. **C:** LTM-to-STM forward dynamics as shown in exponential moving averages of STM and LTM activity following LTM-activation through 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 plasticity (dark orange). Horizontal bars at the bottom indicate activation half-width. Onset is denoted by vertical dashed lines. A shaded arrow denotes the stimulus **D:** Subsampled spike raster of STM (top) and LTM (bottom) during forward activation of the naïve STM by five different LTM-attractors, triggered via specific 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 simple spike counts in each patterns initial activation time window).

Ground state and Attractor activity

In the untrained network, fluctuations in membrane voltages evoke a ground state with low-rate, irregular, asynchronous spiking that transitions into an alpha/beta oscillation at medium input rates (**Figure 2-A**). 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.[28] If the network has embedded attractors, a slightly stronger background excitation (+39%, r_{bg}^{L23}, **Supplementary Table 2**) can occasionally trigger memory item reactivations accompanied by oscillations in the form of theta-nested gamma bursts[29] (**Figure 2-B**). Because attractor activity is self-

limiting through neural adaptation and synaptic depression, this can result in a random walk across stored memories when background excitation is strong. Any attractors, whether randomly or specifically triggered, are subject to known and previously well characterized associative memory dynamics, such as pattern completion, rivalry, perceptual blink, bursty reactivation dynamics, oscillations in different frequency bands, etc.[28–30].

Open loop STM Dynamics

We now consider cued activation of several memories embedded in LTM. Each HC in LTM features a selectively coding MC for any given memory pattern. These distributed memory representations activate as a whole (**Figure 2-B**), so we only need to observe the pattern-specific component MCs in a single HC to get a sense of LTM activations. Five different LTM attractors are triggered by brief cues (*Online Methods*), 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**).

As seen in **Figure 2**, STM activations are sparse (ca 5%) and random. The spatial and temporal structure of STM activity is caused by the tendency of cells in the same MC to fire together (e.g. see Ground state). The distributed, but patchy character of the STM response is further a result of branching forward-projections from LTM layer 3B cells, which tend to activate close-by cells. 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 also features strong recurrent weights between cells in embedded attractors. STM neurons in **Figure 2-D** are colored according to their dominant memory pattern selectivity, which reveals that STM activations are mostly non-overlapping as well. Although MC activity is not exclusive to any given input pattern (unlike the strictly orthogonal LTM patterns), they are clearly clustered. 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 were recently activated by a strong, busting 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 activations. **Figure 2-C** shows a peri-stimulus exponential moving average (EMA) of both networks (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 (which we will scrutinize in more detail later on). Initial oscillatory activity in STM has a lower frequency than LTM because the naïve STM is less active than LTM (compare left and right y-axis) in our model and subsequently does not trigger its basket cells as quickly (the main driver of alpha/beta and gamma). STM oscillations speed up later as new cell assemblies become stronger (e.g. **Figure 3-A** and **Supplementary Figure 1**). As seen in **Figure 2-B**, co-activated MCs in LTM are not always in phase with respect to gamma and large, dispersed forward axonal conduction delays further decorrelate gamma input to STM. Activating strong plasticity in STM (κ_p, *Online Methods* and **Supplementary Table 1**) has a noticeable effect on the amplitude of stimulus-locked oscillatory STM activity after as little as 100ms.

Closed-loop STM-LTM Dynamics

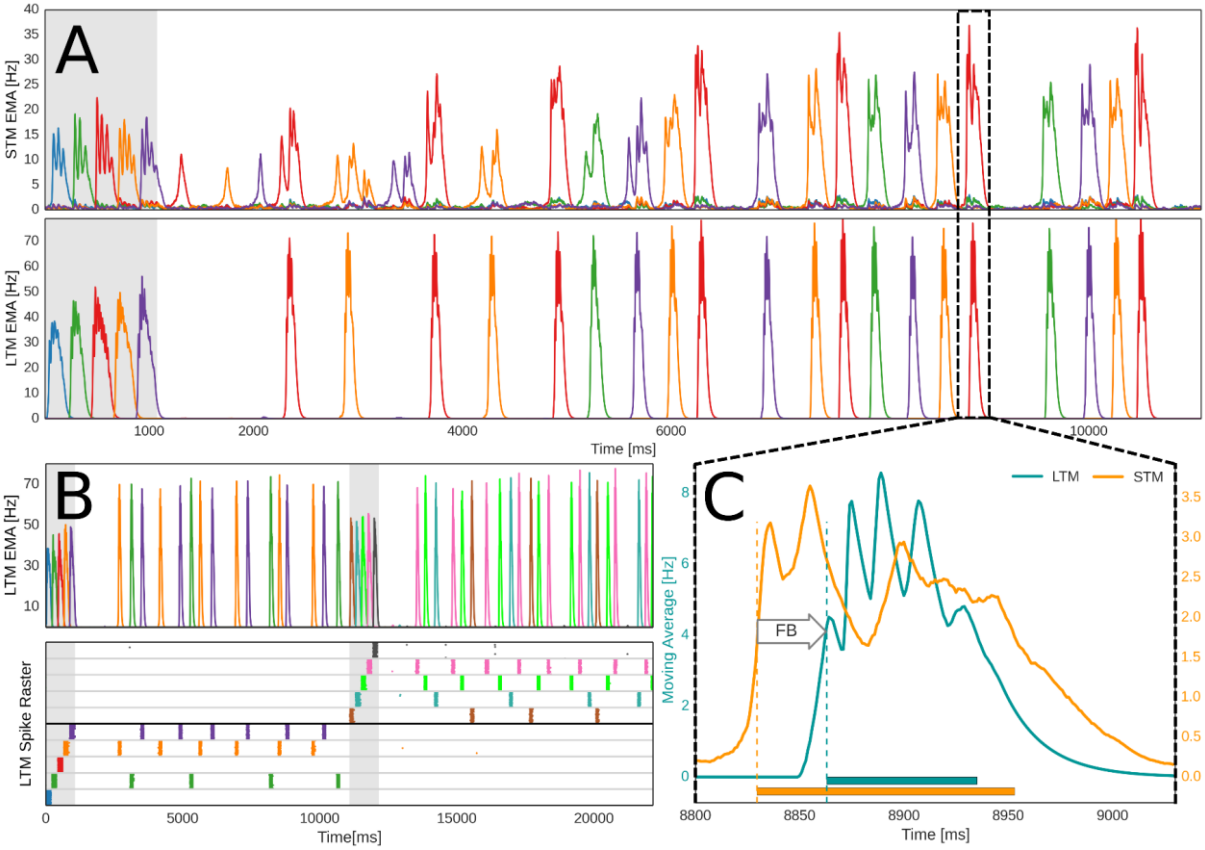


Figure 3. Indexing and reactivation of LTM. A: Exponential moving averages of pattern-specific subpopulations in STM (top) and LTM (bottom) during indexing and subsequent maintenance of five memories (the first pattern is often not maintained). From 0 to 1000 ms (shaded grey), five LTM memories are cued via targeted 50 ms stimuli. Plasticity is temporarily boosted during this time and activated attractors are thus indexed in STM. Thereafter, strong noise drive to STM causes random activations and consolidation of pattern-specific subpopulations in STM. Backprojections reactivate associated LTM memories. **B:** Updating of WM. Rapid encoding and subsequent maintenance of a second group of memories following an earlier set. Times of external input and plasticity modulation shaded grey. Bottom: Spike raster shows layer 2/3 activity of one LTM HC (MCs separated by grey horizontal lines). **C:** STM-to-LTM feedback dynamics during a reactivation event. STM-driven activations of LTM memories are characterized by a feedback delay. Horizontal bars at the bottom indicate activation half-width. Onset is denoted by vertical dashed lines.

In **Figure 2-D** we already demonstrated a relatively simple way to extract pattern-specific subpopulations in STM from forward input spiking activity. By enabling backprojections we can now track activated memories in the closed cortical loop**. In Figure 3-A**, we show network activity following targeted activation of five LTM memories (Spike raster in **Supplementary Figure 1**). Under an increased unspecific noise-drive (r_{WMbg}^{L23}, **Supplementary Table 2**), STM cell assemblies, formed during the brief input-period, may activate randomly. These brief bursts of activity are initially weak and different from full-fledged theta-nested gamma-bursts seen in LTM attractor activity.

The closed cortical loop allows for additional amplification of these brief STM activations if the indexed LTM memory becomes active. Whilst memory pattern specificity in STM is random (but emergent from circuit architecture, short-term dynamics and plasticity), backprojections are plastic and thus acquire specificity from STM-LTM co-activations during forward stimulation. Given a high enough STM firing rate, the sparse but potentiated backprojections can trigger indexed LTMs. In turn, LTM activation will

again feed into STM, typically causing a second peak of activation in STM following the first by about 40 ms (**Figure 3-C**). A small set of memories are now maintained and intermittently active, so the closed cortical loop now constitutes a fully functional WM. Furthermore, there is a noticeable ramp-up in the strength of STM pattern-specific activity over the course of the delay period. Since STM remains plastic throughout our simulation, reactivation events further strengthen pattern-specific cell assemblies in STM and their link to associated LTM attractors. Eventually, STM assemblies become strong attractors themselves, capable of reliably activating their LTM associate. This effect may be called auto-consolidation and it is an emergent feature of the plastic cortical 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.

A crucial feature of any WM system is its flexibility, and **Figure 3-B** highlights an example of rapid updating., i.e. the maintained set of activated memories can be quickly overwritten by yet another set of input memories. As LTM attractors are non-overlapping, two sets of five memories imply that we used a network with at least ten simulated MC per HC here (The original model features only nine but was slightly scaled up). Generally speaking, earlier items will be reliably displaced from active maintenance in our model if activation of those new items is accompanied by the same transient elevation of plasticity (κ_p/κ_{normal}, **Supplementary Table 1**) used during the original indexing of the first five memories. For STM rates and subsampled spike raster see **Supplementary Figures 2**, and **3**.

In analogy with earlier results[19], cued activation can often still recall previously maintained items. The rate of memory decay depends critically on the amount of noise in the system, which erodes learned correlations 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.[31].

Analogous to forward delays in the LTM-driven scenario, we observe a feedback delay in the STM-driven closed-loop mode of operation (**Figure 3-C**). The forward delay is still evident here in a delayed secondary increase of the STM activation following LTM onset, which extends the STM activation (we will scrutinize these delays in more detail later on).

Multi-modal, Multi-item Working Memory

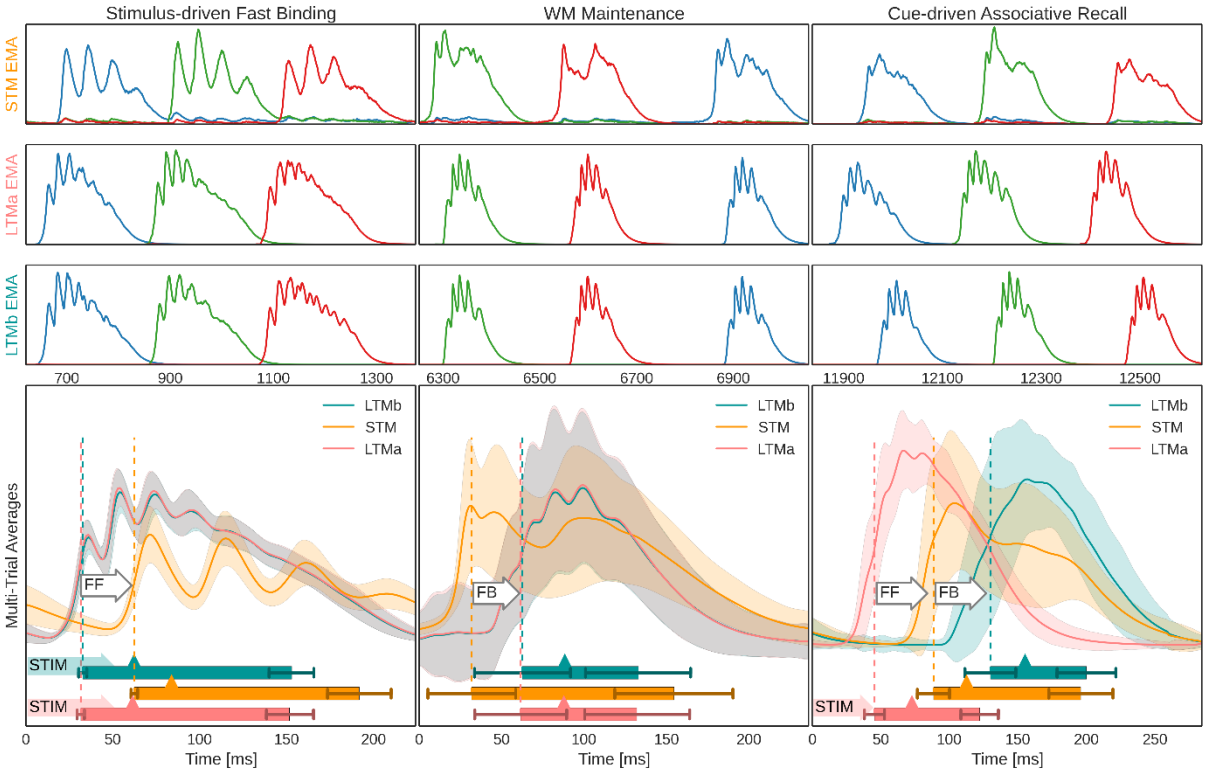


Figure 4. Exponential Moving Averages (EMA) of memory- and network-activations during three different modes of network activity : **Top-Half**: Activation of three memories (blue,green,red respectivly) in STM (1st row), LTMa (2nd row), and LTMb (3rd row) during three different modes of network activity: The initial binding of pairs of LTM memory activations in STM (left column), WM Maintenance through STM-driven 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 attractor half-width, also denoted by a rectangle underneath the traces. Error bars indicate a standard deviation from activation onset and offset. Mean peak activation is denoted by a triangle on the rectangle, and shaded arrows denote targeted pattern stimulation of a network. As there are no external cues during WM maintenance (aka delay period), we use detected STM activation onset to average EMA network traces of 5168 STM-driven LTM-reactivations across trials and reactivation events. 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 propagate into PFC a unique joint STM index of both memories is created when forward-activations combine non-linearly with each other (see **Online Methods)** as well as with prior STM content. **Figure 4** illustrates how this index then supports WM operations, including cue-driven associative pair completion and delay maintenance through STM-driven discrete co-activation events. The three columns of **Figure 4** illustrate three fundamental modes of the closed loop system: Stimulus driven binding, WM maintenance, and associative recall. The top three rows show sampled activity of a single trial (see also **Supplementary Figures 4,5**), whereas the bottom shows multi-trial averages.

During stimulus-driven binding, we co-activate pairs of LTM memories. As before, when we used only one LTM network, this involves a brief 50 ms cue to each patterns specific sub-population, which then triggers activation of the corresponding attractor. As all long-term memories are encoded equally well in LTM, activations are almost perfectly synchronized. The multi-trial (and multi-pattern) average of peri-stimulus activations reveals a 45 ± 7.3 ms attractor activation delay, followed by a 43 ± 7.8 ms feedforward delay (about half of which is explained by the spatial distance between LTM and STM)

from the practically simultaneous onset of the two LTM activations to the onset of the input-specific STM response (also **Figure 5 top-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, mostly because the driving STM-indices are of variable size and strength, which has an impact on the feedback delay distribution, with mean 41.5 ± 15.3 ms.

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 cortical 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 index as the input is strongly correlated to learned inputs from the simultaneous activation with LTMb earlier on, and input-specific STM activation patterns have become temporarily stable attractors in their own right. The completed STM index then triggers the associated memory in LTMb through its backprojections. STM activation extends beyond the transient activity of LTMa because the cortical loop keeps it active (now driven from LTMb-inputs). 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 computed the Bottom-Up and Top-Down response latency of the model in analogy to Tomita et al.[27]. Their study explicitly tested widely held believes 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 and 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 (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. The display of a cue in one hemi-field in the experiment 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.

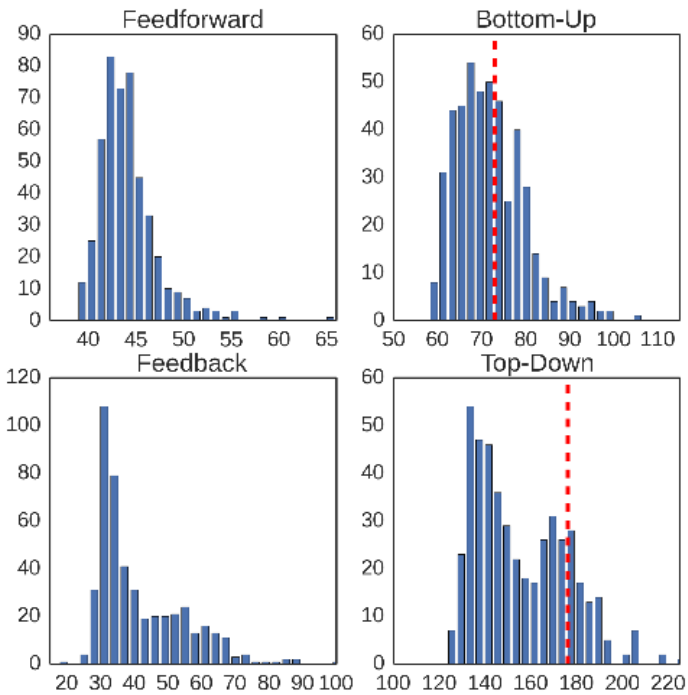


Figure 5. Distributions of key delays in the trained system during associative recall, averaged over 100 trials with 5 paired stimuli each.: Top-Left: Feedforward delays distribution, as defined by the temporal delay between LTMa onset and STM onset (as shown in Figure 4, Bottom-right). Top-Right: Bottom-up delay distribution, as defined by the temporal delay between stimulation onset and LTMa peak activation. The red lines denotes the mean bottom-up delay, as measured by Tomita et al.[27]. Bottom-Left: Feedback delays distribution, 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, as defined by the temporal delay between stimulation onset and LTMb peak activation. The red lines denotes the mean bottom-up delay, as measured by Tomita et al.[27].

Discussion

We have presented the design and evaluation of a spiking neural network WM model, featuring STM (PFC) and LTM components (temporal cortex) as well as fast Hebbian synaptic plasticity. We demonstrated how single memory items could be encoded by one-shot learning and how PFC rapidly enhances synaptic connections intrinsically as well as those targeting LTM representations, which are thereby "brought on-line". The PFC cell assembly serves as a temporary index into LTM in a similar manner as suggested by the "Hippocampal Indexing Theory" for the relation between hippocampus and neocortex in episodic memory and memory consolidation[17,32]. When a pair of LTM items are bound together by being simultaneously indexed in this manner, activating one will also activate the other indirectly via PFC with a short latency. We further demonstrated that this kind of WM is readily updated such that as new items get encoded, old ones are gradually over-written.

Recall dynamics in the model presented is in most respects identical to our previous cortical associative memory models[33]. From that we can deduce e.g. that sequential learning and recall could readily be incorporated in the model[34]. This could support encoding of e.g. sequences of items in WM rather than unrelated ones, resulting in a reactivation dynamic reminiscent of the "phonological loop"[35].

The Case for Hebbian Plasticity

A requirement 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[21,22,36,37] 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[38]. In analogy we propose that D1R activation is also critically involved in the synaptic plasticity intrinsic to PFC and in projections to LTM stores which would also explain prominent WM effects of PFC DA level manipulation[39,40]. In our model, kappa 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 LTM stimulation and WM encoding, i.e. a form of as 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[16,29]. However, such non-Hebbian plasticity is only capable of less specific forms of memory. Activating a cell assembly comprising a subset of neurons in a naïve STM network 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 alone not capable of supporting encoding of novel memory items or the multi-modal binding of already existing ones. In our previous paper[19] we showed that fast Hebbian plasticity in the form of STP[21] allows such one-shot encoding of new 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 will also be capable to bind and bring on-line previously not associated LTM items.

Our implementation of a fast Hebbian plasticity reproduces a remarkable aspect of STP: It decays in an activity-dependent manner[36,37]. Its decay is not noticeably time-dependent, and silence preserves synaptic information. The detrimental effects of distractors on performance in virtually all kinds of WM tasks suggests an activity-dependent update, as does the duration of "activity-silent WM" in recent experiments[10]. 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.

Model Robustness

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 generally robust and degrades gracefully. Almost all uncertain parameters can be varied ±30% without breaking WM function. The model is dramatically sub-sampled 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[19,29,34,41,42], so we here prioritize novel aspects, namely the parameterization of corticocortical connectivity and spatial scale.

In the feedback pathway, a mere 0.66% connectivity is sufficient to support STM-driven LTM activation in maintenance and recall. Even 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%), because terminal clusters in STM are smaller and provide more information contrast (see Online Methods). In both cases, our model uses these low density values, but they could be increased/decreased if single synaptic currents are reduced/increased proportionally. In fact, we found that we needed to increase the corticocortical conductance of the back-projections (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[43] 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 end, 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. Theta oscillations degrade at 45% and completely break at 20% spatial scale, when the largest inter-HC delays fall below 5 ms. Attractor activity then collapses into a single gamma cycle (rather than longer bursts), which degrades learning and effective information transmission both within and across networks. Networks may be much smaller however, but this needs to be compensated by slower axonal conductance velocities (<2 mm/ms).

Experimental support and testable predictions

Unfortunately, the detailed neural processes represented in 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[27] because of its clean experimental design (even controlling for subcortical pathways). The timing of inter-area signals also constitute a testable prediction for multi-modal memory experiments. Furthermore, reviews of intracranial recording conclude that theta band oscillations plays an important role in long-range communication during successful retrieval[44]. With respect to theta in our model, STM leads the rest of cortex during maintenance, engages bi-directionally during recall (due to the closed cortical loop), and lags with respect to the theta phase during stimulus-driven encoding and LTM activation, reflecting experimental observations[45]. These effects can be explained by our model architecture, which imposes delays due to the spatial extent of networks and their distances from each other. Gamma power, while often theta-nested, is more strongly linked to local processing and individual foreground items in our model, also matching experimental findings[44]. Cross frequency coupling is abundant with significant phase-amplitude coupling (e.g. **Figure 2B**), and was well characterized in related models[46].

The most critical requirement and thus prediction of our model is the presence of fast Hebbian plasticity in the PFC back-projections to temporal memory areas. Without such plasticity, our model cannot explain STM-LTM binding. This is likely to co-exist with neuromodulatory control over that plasticity, likely with DA and D1R activation involvement. Since the STP decays with activity a high noise level could be an issue since it could shorten WM duration.

Our model also makes specific predictions about corticocortical long-range connectivity. For example, as few as six active synapses (see **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 activation in executive control of LTM memories (e.g. see STM EMA 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 of a successfully activated LTM. As such, the second peak should also be correlated to successful memory maintenance or associative recall.

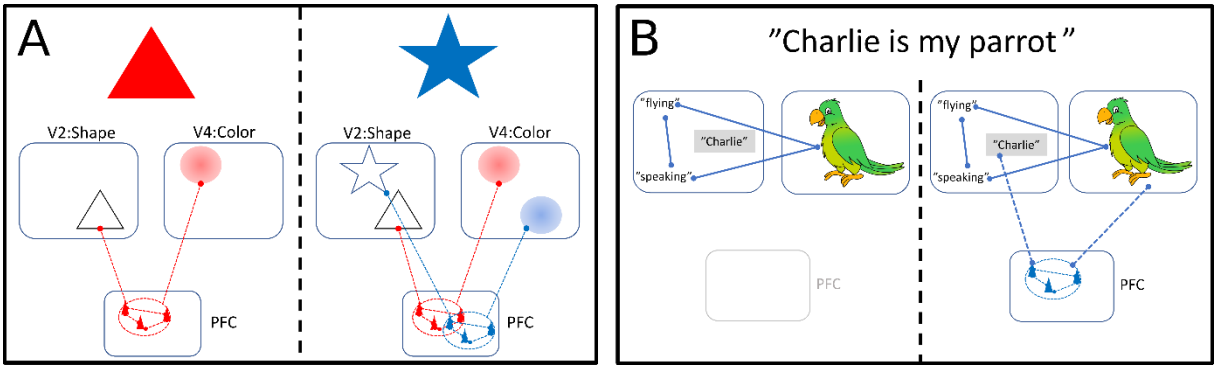


Figure 6. Feature binding and variable binding via PFC. A: Feature Binding: When a red triangle and followed by a blue star is shown and attended these shape – color binding is formed by fast Hebbian plasticity via PFC to form a composite cell assembly. It supports pattern completion meaning that stimulation with shape will trigger the color representation and vice versa. **B: 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.

The Neural Binding Problem

The so called "binding problem" refers to the brain's ability to temporarily connect items previously not connected by earlier experience and learning[47]. The fast Hebbian plasticity proposed here as a mechanism for WM offers without any modification an additional possible neural mechanism for binding that solves the binding problem. For instance, visual feature binding refers to the instant association of e.g. shape and color of multiple objects without mixing them up during recall (**Figure 6-A**). Suggested neural mechanisms include temporal binding by synchronization[48] or feature integration[49]. Filling roles with instances, like e.g. in "Charlie is my parrot", is a second form of variable-value binding (**Figure 6-B**). This and other forms of neural binding underlie even more complex functions in human cognition including logic reasoning[50]. That PFC may be involved is supported by the fact that its temporary inactivation in humans affect feature binding[51], and that successful feature binding requires attention[23]. Gamma band oscillations, frequently implicated when binding is observed, are also a prominent output of our model's WM networks[52].

Conclusions

Our simulations have demonstrated the versatile WM properties of a spiking neural network composed of STM and LTM components and we have partly connected them to existing experimental data. Although our model is still quite abstract, and theory and simulation models can never prove anything about biology, it nevertheless offers a biologically plausible mechanistic understanding of WM which connects microscopic processes with macroscopic observations and function in a way that only computational models can do. While we applied this model to macaque, it is quite generic and we expect model findings to apply also to mammals including human, commensurate with changes in key parameters (cortical distances, axonal conductance speeds, etc.).

WM dysfunction has an outsized impact on mental health, intelligence, and quality of life, progress in this respect is very important for society. Our hope is that our new computational model, which reproduces several important aspects of human WM, will trigger many more targeted experiments

aimed to evaluate and improve it. More adequate models and understanding can help position future 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). The simulations were performed using computing resources provided by the Swedish National Infrastructure for Computing 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[53] which was modified[42] for compatibility with a custom-made BCPNN synapse model in NEST 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[54]:

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

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 decays 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)[21] with a spike-based version of the Bayesian Confidence Propagation Neural Network (BCPNN) learning rule[42,55]. 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[19].

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[13,17,42]. 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:

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

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[37]; 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, like 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. [42] 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[56] 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 [57] or TRP channel [58]. 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}.

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

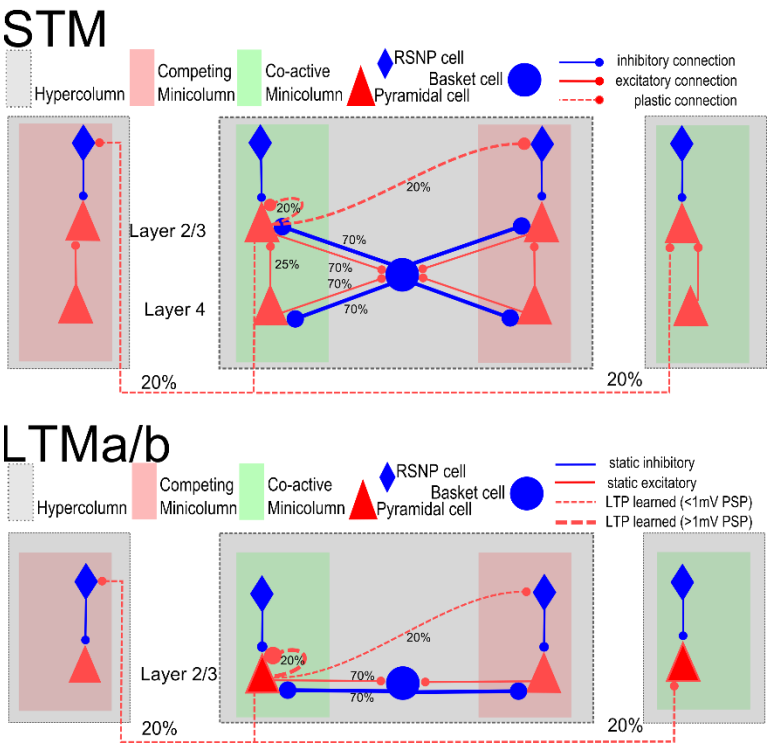
In our model, AMPA and NMDA synapses make use of w_{gain}^{AMPA} and w_{gain}^{NMDA} respectively. The logarithm in **Equations 8,9** is motivated by the Bayesian underpinnings of the learning rule, and means that synaptic weights w_{ij}^{syn} multiplex both the learning of excitatory and di-synaptic 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[59–62]. Accordingly, BCPNN connections with a negative weight use a GABAergic reversal potential instead, as in previously published models[19,25,42]. 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[63].

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:

$$\overline{t_{ij}} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{V} + t_{min}^{syn}\, ms \qquad t_{ij} \sim N(\overline{t_{ij}}\,,\ .15\overline{t_{ij}}) \qquad (7)$$



Supplementary Figure 1. Local columnar connectivity within STM and LTM. Connection probabilities are given by the percentages, further details in **Supplementary Tables 1 and 3**. The strength of plastic connections develops according to the synaptic learning rule described in *Spike-based Bayesian Learning*. Dashed connections are not plastic in LTM (besides the STD of **Equation 4**), but already encode attractors previously learned through an LTP protocol, and loaded at simulation using effective weights found in **Supplementary Table 2**.

STM Network Architecture

We simulate $n_{HC}^{STM} = 25$ HC 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[64], 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 a 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[65]. 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 [66]. Local basket cells fire in rapid bursts, and induces 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 an 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 (*Online Methods*, and **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[67] 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[68].

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 temporal cortex commonly associated with auditory and visual LTM. For example Inferior Temporal Cortex (ITC) is often referred to as the storehouse of visual LTM[69]. Two such LTM areas are indicated in **Figure 1.**

We simulate $\mathrm{n}_{\mathrm{HC}}^{\mathrm{LTM}} = 16$ HC 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 (*Online Methods*). Various previous models with identical architecture have demonstrated how attractors can be learned via plastic BCPNN synapses[18,19,25,42]. We load each LTM network with nine orthogonal attractors. 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 in all LTM stores.

In summary, the two LTM models thus features 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.[70] 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 macque 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[71–73].

In the forward path, layer 3B cells in LTM project towards STM (**Figure 1**). We do not draw these connection 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 0.0015). The resulting split

between targets in layer 2/3 and 4 is typical for feedforward connections at medium distances in the cortical hierarchy[70] and has important functional implications for the model (*Cortical 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[59–62]. Although we do not explicitly simulate such cells, such an interneuron would be local to a MC and targeted by incoming excitatory connections (same arrangement as Tully et al. 2014, 2016). 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 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 [70]. 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[74]. 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**).

To cue the activation of a specific attractor, we drive 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 attractors are already strongly encoded in each LTM, a single, targeted 50 ms cue is usually sufficient to activate any given memory.

Spike Train Analysis and Memory Activity Tracking

We track memory activity in time by analyzing the exponential moving average (EMA) of pattern-specific and network-wide spiking activity usually using a filter time-constant of 20 ms. As activations are characterized by sizable gamma-bursts, a simple threshold detector can extract candidate activation events and decode the activated pattern (which is trivial due to the orthogonal and known nature of LTM activation patterns). We measure onset and offset by thresholding each individual activation by half of its peak value. In LTM, we further assess pattern completion by checking for component MC activation. Whenever targeted stimuli are used, we analyze peri-stimulus activation traces. When activations onsets are less predictable, such as during WM maintenance, we extract activations candidates via a threshold detector trained at the 50th percentile of the cumulative distribution of the EMA signal.

Simulation Environment

We use the NEST simulator[56] 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

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		Axonal Conduction Speed	V	$2\,\frac{m}{s}$
Simulated HC	n_{HC}^{STM}	25	Minimal conduction delay	t_{min}^{syn}	1.5 ms
Simulated MC per HC	n_{MC}^{STM}	12	STM – LTM distance	$d_{STM-LTM}$	40 mm
			Background activity rate layer 4	r_{bg}^{L4}	300 Hz
			Background activity rate layer 2/3	r_{bg}^{L23}	625 Hz
LTM patch size	25 x 25 mm		Attractor activity threshold	r_{bg-low}^{L23}	450 Hz
Simulated HCs	n_{HC}^{LTM}	16	Maintenance STM background activity rate layer 2/3	r_{WMbg}^{L23}	950 Hz
Simulated MC per HC	n_{MC}^{LTM}	9	Background conductance	g_{bg}	±1.5 nS
			Cue stimulus duration	t_{cue}	50 ms
			Stimulation rate	r_{cue}	850 Hz
Layer 2 pyramidal per MC	n_{MC}^{PYR-L2}	20	Cue stimulus conductance	g_{cue}	+1.5 nS
Layer 3A pyramidal per MC	$n_{MC}^{PYR-L3A}$	5	LTM Intra HC – Intra MC weight	$w_{IntraMC}^{IntraHC}$	3.36 w_{gain}^{syn}
Layer 3B pyramidal per MC	$n_{MC}^{PYR-L3B}$	5	LTM Intra HC – Inter MC weight	$w_{InterMC}^{IntraHC}$	-4.82 w_{gain}^{syn}
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 for not simulated in LTM.

Scope	Source	Target	Type	Symbol	Value
Corticocortical	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

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, P. S. Cellular basis of working memory. *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. Shafi, M. *et al.* Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* **146,** 1082–1108 (2007).

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

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

11. O'Reilly, R. C., Braver, T. S. & Cohen, J. D. A Biologically-Based Computational Model of Working Memory. *Model. Work. Mem. Mech. Act. Maint. Exec. Control* 375–411 (1997).

12. Lansner, Sandberg, Petersson & Ingvar. On forgetful attractor network memories. in *Artificial neural networks in medicine and biology: Proceedings of the ANNIMAB-1 Conference* 54–62 (Springer, London, 2000). doi:10.1007/978-1-4471-0513-8_7

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

14. Sandberg, A., Tegnér, J. & Lansner, A. A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems* **14,** 789–802 (2003).

15. Amit, D. J. & Mongillo, G. Spike-Driven Synaptic Dynamics Generating Working Memory States. *Neural Comput.* **15,** 565–596 (2003).

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

17. 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).

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

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

20. Hanse, E. & Gustafsson, B. Onset and stabilization of NMDA receptor-dependent hippocampal long-term potentiation. *Neurosci. Res.* **20,** 15–25 (1994).

21. 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).

22. 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).

23. Adam, G. The relationship between attention and working memory. *Front. Hum. Neurosci.* **5,** 1–45 (2011).

24. Norris, D. Short-term memory and long-term memory are still different. *Psychol. Bull.* **143,** 992–1009 (2017).

25. 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).

26. 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).

27. 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).

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

29. 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).

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

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

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

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

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

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

36. 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).

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

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

39. 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).

40. 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).

41. 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).

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

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

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

45. 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).

46. 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).

47. 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

48. Gray, C. M. & Singer, W. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci.* **86,** 1698–1702 (1989).

49. Treisman, A. The binding problem Abbreviations FIT feature integration theory IC illusory conjunction IT inferior temporal cortex PET positron emission tomography. *Curr. Opin. Neurobiol.* **6,** 171–78 (1996).

50. 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).

51. 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).

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

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

54. 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).

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

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

57. 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).

58. 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).

59. 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).

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

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

62. 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).

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

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

65. 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).

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

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

68. 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).

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

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

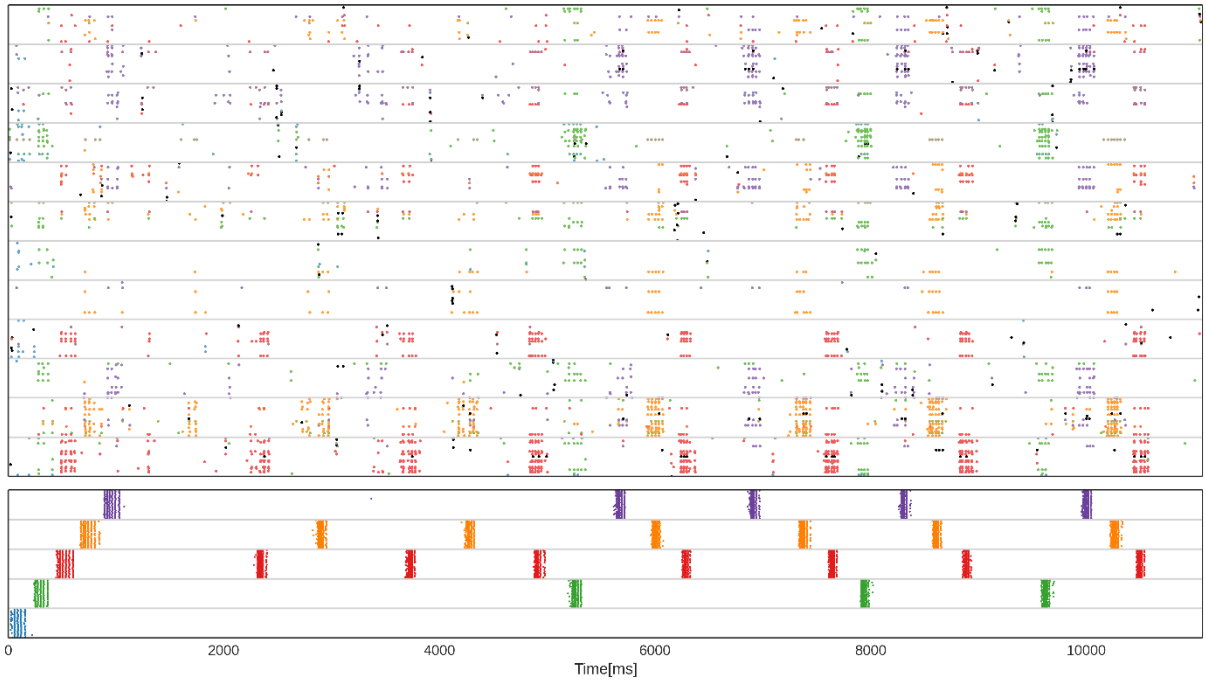
71. Girard, P., Hupe, J. M. & Bullier, J. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J Neurophysiol* **85,** 1328–31. (2001).

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

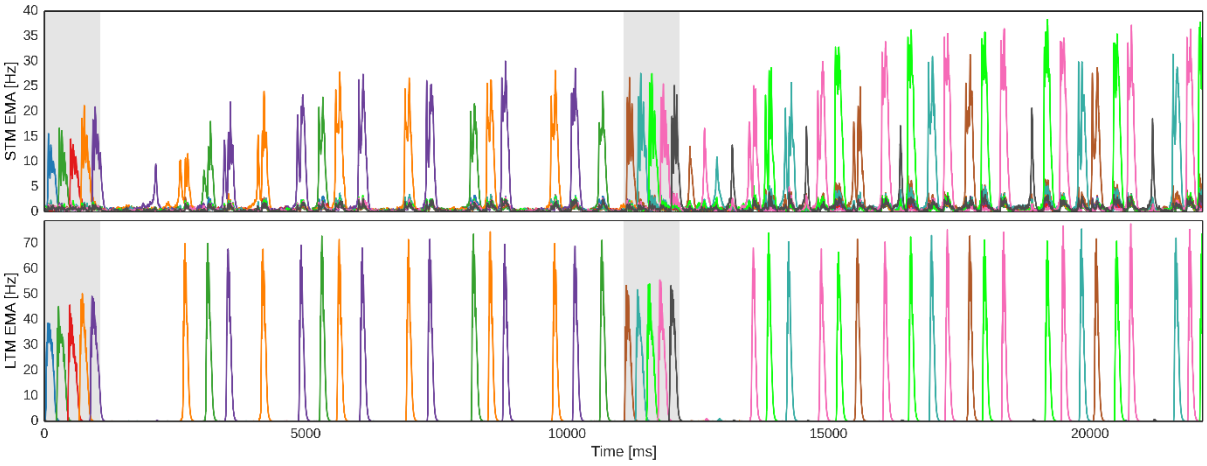
73. Caminiti, R. *et al.* Diameter, Length, Speed, and Conduction Delay of Callosal Axons in Macaque Monkeys and Humans: Comparing Data from Histology and Magnetic Resonance Imaging Diffusion Tractography. *J. Neurosci.* **33,** 14501–14511 (2013).

74. 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).

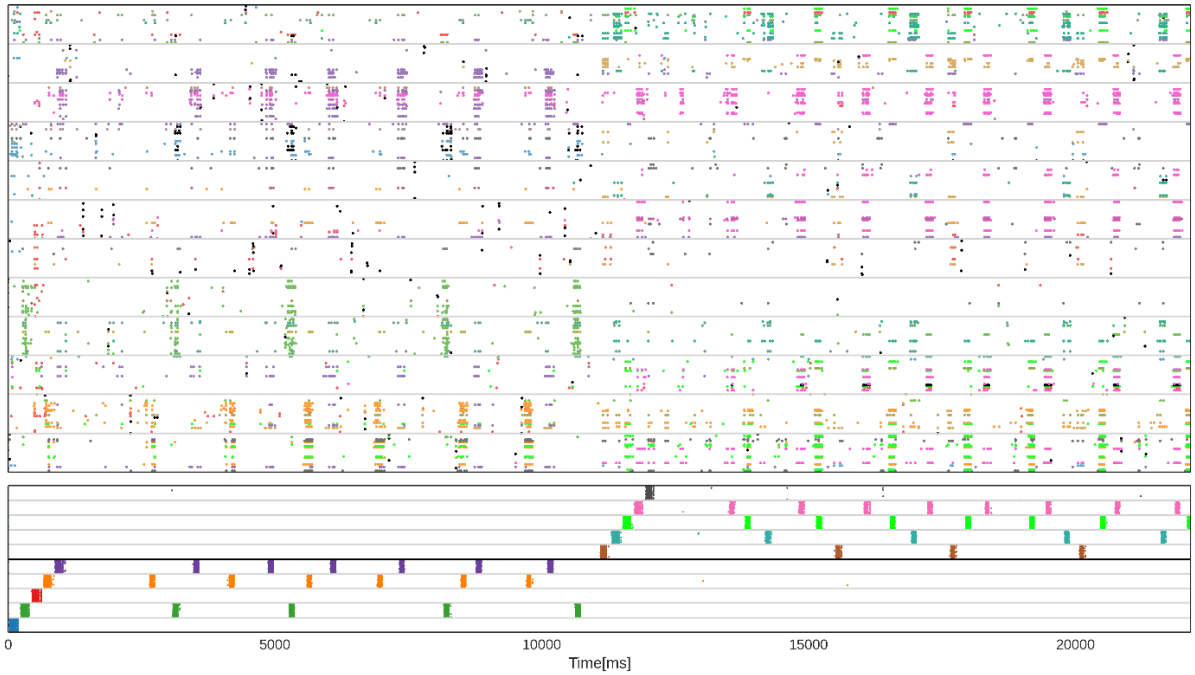
Supplementary Material



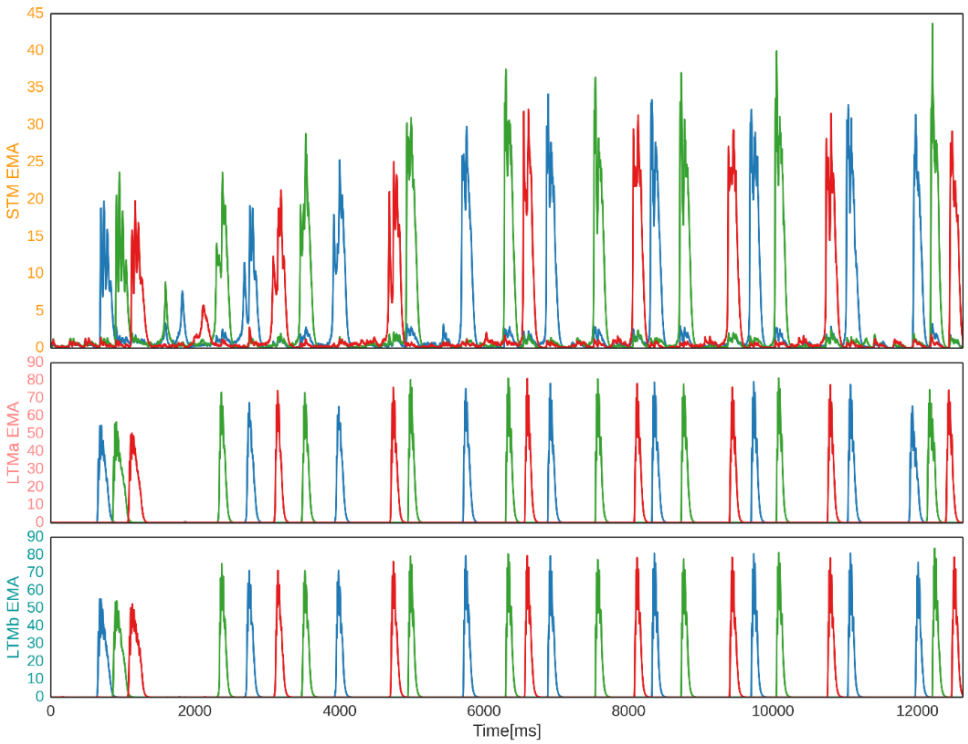
Supplementary Figure 1. Indexing and feedback reactivation of long-term memories. A: Subsampled spike raster of STM (top) and LTM (bottom). From 0 to 1000 ms (shaded grey), five LTM memories are cued via targeted 50 ms stimuli. Plasticity is temporarily boosted during this time and activated attractors are thus indexed in STM. Thereafter, strong noise drive to STM causes random activations and consolidation of pattern-specific subpopulations in STM. Backprojections reactivate associated LTM memories. Top: STM spike raster shows layer 2/3 activity in a single hypercolumn (HC). Minicolumns (MC) are separated by grey horizontal lines. STM spikes are colored according to each cells dominant pattern-selectivity (based on simple spike counts in each patterns initial activation time window), similar to **Figure 2**. Bottom: LTM spike raster only shows the activity of five coding MC in a single LTM HC, but indicates the activation of distributed LTM attractors. LTM Spikes are colored according to the pattern-specificity of each cell.



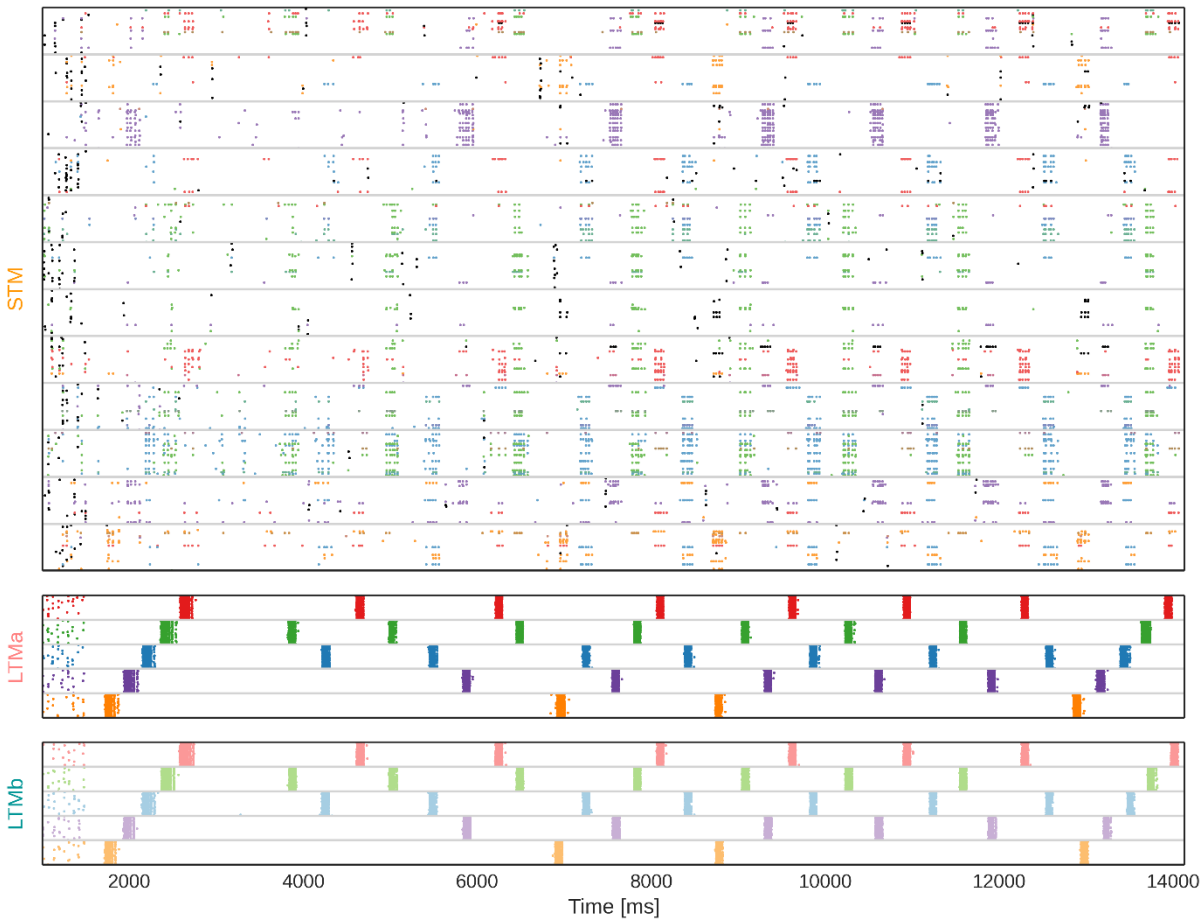
Supplementary Figure 2. WM updating. Exponential moving averages of pattern-specific subpopulations in STM (top) and LTM (bottom) during indexing and subsequent maintenance of two sets of five LTM memories. After encoding and 10 s maintenance of the first set, WM contents are overwritten with the second set of memories, maintained thereafter. The grey shading indicates input-driven updates to the WM system. Plasticity is temporarily boosted during these times and activated attractors are thus indexed in STM. Subsequent strong noise drive to STM causes random activations and consolidation of pattern-specific subpopulations in STM following each stimulation period. Backprojections reactivate associated LTM memories.



Supplementary Figure 3. WM updating. Subsampled spike raster of the layer 2/3 population in a Hypercolumn of STM (top) and LTM (bottom) respectively during indexing and subsequent maintenance of two sets of five LTM memories. STM spikes are colored according to each cells dominant pattern-selectivity (based on simple spike counts in each patterns activation time window). 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. Strong noise drive to STM causes random activations and consolidation of pattern-specific subpopulations in STM following each stimulation period. Backprojections reactivate associated LTM memories.



Supplementary Figure 4. Exponential Moving Averages 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 pairs 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).



Supplementary Figure 5. 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 (middle) and LTMb (bottom) respectively during indexing, 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 (based on simple spike counts in each patterns activation time window). 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. The spike rasters shows a brief period of noisy activity in all three networks, before the targeted sequential activation of five LTM pairs. From 3000 ms to 13000 ms the rasters show STM-driven WM maintainance of the encoded pairs. Finally, at 13.000 ms LTMa alone is cued and retrieves the corresponding LTMb activation with some delay through STM activation.