A Hot-Coal theory of Working Memory

Mikael Lundqvist[1, 2,*], Jonas Rose[2, 3], Melissa R. Warden[2, 4], Tim Buschman[2, 5], Earl K. Miller[2], Pawel Herman[6,*]

1 Department of Psychology, Biological psychology, Stockholm University, SE-10691, Stockholm, Sweden

2 The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA, 02139, USA

3 Faculty of Psychology, Neural Basis of Learning, Ruhr University Bochum, 44801, Bochum, Germany

4 Department of Neurobiology and Behavior, Cornell University, Ithaca, NY, 14853, USA

5 Princeton Neuroscience Institute, Princeton University, Washington Rd., Princeton, NJ 08540, USA

6 Computational Brain Science Lab, Department of Computational Science and Technology, KTH Royal Institute of Technology, Stockholm, 100 44, Sweden

*corresponding authors

Abstract

Working memory allows us to selectively remember and flexibly manipulate a limited amount of information. Importantly, once we learn a certain operation, it generalizes to any memory object, not just the objects it has been trained on. Here we propose a conceptual model for how this might be achieved on the neural network level. It relies on spatial computing, in which sensory information flows spatially within the network over time. As a result, information about, for instance, object order can be retrieved agnostically to the detailed synaptic connectivity responsible for encoding specific memory items. This spatial flow is reflected in low-dimensional brain activity complementing high-dimensional activity that accounts for storing the sensory information itself. By comparing the dimensionality of local field potentials and spiking activity from prefrontal cortex of rhesus macaques performing multi-item working memory tasks we verify predictions from this model. We discuss how spatial computing may be a principle to aid generalization and zero-shot learning by utilizing spatial dimensions as an additional information encoding dimension. The new model also helps explain several aspects of neurophysiological activity related to working memory control, including dimensionality, context-dependent selectivity as well as persistent and non-persistent delay activity.

Introduction

Working memory (WM) is a key aspect of higher-order cognition (Baddeley, 1992; Miller & Cohen, 2001). It is a mental sketchpad for the short-term storage and top-down control of information. An essential feature of WM is flexibility. A wide variety of information, sensory inputs, decisions, recalled memories

etc can be selected, maintained, manipulated, and read out as needed (Warden and Miller, 2010; Chatham et al., 2014; Wolff et al., 2017; Yu et al., 2020; Lewis-Peacock et al., 2015; Lundqvist et al., 2018a; van Ede et al., 2017). How are neural networks wired to achieve such flexibility? In many models of WM, information about both the items held in WM and the control operations performed on them are determined by network synaptic connectivity (Masse et al., 2019). This, however, leads to an issue to resolve: if anything changes about the task, e.g., the rules, the timing of events, or the set of items used as memoranda, training is required to modify the network weights. This perennial need for re-training stands in contrast to the general and flexible use of WM in humans and other primates.

Here we propose a new conceptual "Hot-Coal" theory of WM. It targets fundamental challenges for explaining how WM can be general purpose and flexibly controlled. It builds on our previous model of working memory (Lundqvist et al., 2011; Lundqvist et al., 2016; Miller et al., 2018), in which bursts of beta band oscillations act as a control signal that inhibits (and thus gates and controls) bursts of gamma-band oscillations and associated spiking that encodes and maintains WM content. The key new feature added here is *Spatial Computing*, which uses the spatial components of network activity as an additional dimension of encoding task-related information. Consequently, moving information about the stimulus from one part of the network to another is considered as memory computation by itself. In the model this is achieved by a spatio-temporal pattern of excitation shaping a particular expression of sensory information encoded redundantly across the cortical sheet, like the flow of oxygen shaping the glow on a bed of hot coal. This way the imposed pattern controls where in the network a specific memorandum can be accessed at any given point in time without knowing the precise full synaptic connectivity underlying the storage.

Learning a new task corresponds to learning a new spatio-temporal pattern of excitation that controls the flow of WM items across the network and is reflected in a low-dimensional component of neural activity. Because learning does not hinge on changing synaptic weights related to memory storage, the model is flexible and general purpose in the sense that it can use new items, novel to the task, without any re-training.

Here, we find supporting evidence for the proposed conceptual Hot-Coal model of WM in neurophysiological recordings from monkey's prefrontal cortex (PFC), which is implicated in both storage and flexible control aspects of WM (D'Esposito et al., 1995; Miller & Cohen, 2001; Badre and Wagner, 2004). By comparing beta, gamma and spiking activity, we show that task-related information and WM items are represented on different spatial scales in line with the spatial computing hypothesis. We also find evidence that low-dimensional spatial component of beta bursting directs information about the same item in different parts of the network when it is used in different contexts (e.g., as a remembered item vs an item used for comparison). Finally, we discuss how the new theory can provide an account for generalizable control and zero-shot learning. It can simultaneously tie together a large number of experimental findings about the intrinsic dimensionality of neural activity, task-dependent selectivity as well as persistent and non-persistent WM activity.

Results

The conceptual Hot-Coal Model

The proposed Hot-Coal model is compatible with our recent WM model with respect to the reported spiking and oscillatory burst manifestations (Lundqvist et al., 2011; 2016; Miller et al., 2018). In particular, according to the model sensory information held in WM is carried by spiking associated with brief gamma bursts, primarily in superficial-layer cortex. In between spiking/gamma bursts, WM contents are maintained by temporary increases in synaptic weights induced by the spiking (Wang et al., 2006). Gamma/spiking bursting is interleaved with beta bursting (and lower levels of spiking), primarily originating in deep cortical layers. The beta is inhibitory to the gamma bursting and thus regulates its expression.

The key novel concept of the Hot-coal model (Figure 1) is spatial computing. It makes the control of sensory items independent on the detailed synaptic connectivity that maintains them in order to facilitate generalization. Spatial computation rests on two main principles: *i)* Items held in WM are supported by redundant connectivity such that they can be accessed independently in different parts of a network; *ii)* access to WM items is mediated by spatio-temporal activation maps. A spatio-temporal pattern of beta bursts, inhibitory to gamma bursting, serves as a dynamic template shaping the activation map. The gamma and spiking that carry WM content are thus expressed wherever and whenever there are "pockets" of low beta bursting. This way gating, manipulation, and reading out WM content is controlled by the pattern of beta bursting.

Metaphorically, WM storage can be seen as a bed of hot coal glowing with memories of recent inputs, which constitute the high-dimensional component of the activity. Task-related control, the rules that act on the items (such as selective prioritization and/or access to an items held in WM) come from spatial-temporal patterns of excitatory bursts (like puffs of oxygen) that re-heat and shape the item information across the network. This controls where in the network a specific memorandum can be accessed at any given point in time without requiring knowledge of the precise synaptic connectivity underlying the storage. To achieve this there are two principal drivers of neural activity underlying WM control.

The first driver comes from local recurrent excitatory connectivity shaped by temporary changes in synaptic weights that reflect the WM traces. These circuits mediate the gamma bursting and associated spiking that carries the WM contents. This is preserved from our previous model (Lundqvist et al., 2011). In the Hot-Coal model, however the recurrent connectivity is not strong enough to support WM retention with attractor-like activity once the sensory input is removed. Its main role in the model is instead to supply a fading "glow" of information (carried by the synaptic weights changes) that can be reactivated.

The second driver of activity is hypothesized to originate from sub-cortical inputs reflecting task demands. It manifests itself as a spatio-temporal pattern of beta bursting that determines the content-related bursts of gamma/spiking of various parts of the cortical sheet (Figure 1A). When gamma bursts, controlled by the beta pattern, ignite they turn the local 'glow' of information into information expressed in spiking by activating the local winner-take-all circuitry of the superficial layers (Figure 1B; Lundqvist et al., 2010). Importantly, this control spatio-temporal beta pattern, which we propose to emerge in the loop between deep cortical layers, thalamus and the basal ganglia and reflect subcortical excitation (O'Reilly and Frank, 2006; Chatham et al., 2014; Chatham and Badre, 2015; Ketz et al., 2015; Schmitt et al., 2017; Lusk et al.,

2020). In cortex, it is reflected in a task-related low-dimensional component in the neural activity, shared by many neurons based on their spatial location.

The subcortical control of the task-dependent flow of activity in the network mediates "*spatial computing*" in the WM network. This allows control of *where* WM content (spiking) is located in the network without knowing *what* the content is per se or its precise format (i.e., the specific set of synaptic weights that underlies representation of that content). Thus, for example, in a WM sequence the first or second item can be selected (activated) or discarded (suppressed) without having to know and thus target the exact pattern of connectivity between individual neurons that represent that item. One key manifestation of this model is a low dimensional signal shared by spiking of many neurons, even those that represent different WM contents (Figure 1B, C). Each unique set of task demands corresponds to a unique pattern of low-dimensional spatiotemporal activity. A key feature that enables spatial computing is redundancy in cortical connectivity that stores the WM content (Figure 1A; Goldman-Rakic, 1996). This means that several parts of the cortical sheet are capable of holding the same information to support the spatial transfer of information about WM items. Most models do not capitalize on representations that are reported to be highly redundant across cortical sites (Goldman-Rakic, 1996; Dotson et al., 2018).

In motivation of the model, the principle of spatial computing can help explain observations that the main effect of training WM tasks is a more pronounced low-dimensional activity rather than, as many existing models predict, improved retention of sensory information (Tang et al., 2019). Further, the beta bursting that gives rise to the observed low-dimensional spatio-temporal profiles (e.g. ramping activity) leads at the same time to highly variable activity on single trials (Figure 1B). These two effects are otherwise hard to reconcile. The model can also explain observations of context dependent selectivity of individual neurons as an interplay between the spatio-temporal changes in excitability interacts with the stimulus preference of individual neurons (Figure 1D). In this view, a neuron that responds to a certain stimulus may, for example, be prevented from doing so at certain times by inhibition of that neuron's local network.

Below, we present the results of testing the model predictions in neural activity recorded from the lateral prefrontal cortex (PFC) of rhesus monkeys while they perform working memory tasks.

Prediction: Distinct neural sources of low and high-dimensional activity

One key prediction of the model is that (low-dimensional) task information and (high-dimensional) information about WM content have different sources: the PFC-thalamus-basal ganglia loop vs recurrent connectivity within PFC itself, respectively. To test this, we used PFC activity recorded during performance of a serial two-item working memory task. Monkeys had to remember the identity and order of two visual objects presented in sequence (Task 1, Figure 2A). After a brief memory delay, there was a two-object test sequence that could either be identical to the encoded sequence or a mismatch (either the temporal order or the identity of the objects was changed).

In previous analysis of the same data, the spiking of neurons reflecting a given object ramped up prior to its expected usage but not before tests where the particular item was not behaviorally relevant (Warden and Miller, 2010; Lundqvist et al., 2018a). On a single neuron level, we observe however that this flexible ramping behavior is limited. A maximally flexible ramping neuron would only ramp up activity for its

preferred stimulus, and not for others (Figure 2B). This would be easily achievable in a trained RNN. Instead, single neurons with ramping behavior before a particular test tend to ramp up to varying degree following all stimuli, not just the preferred one (illustrated in Figure 2B, single neuron examples in Figure S1). This could be due to generally broad tuning, but the model suggests that this is due to ramping of sub-cortical item-independent excitatory inputs in the spatial vicinity of this neuron before a particular test cue in a trial. Thus neurons in a given network location have the flexibility to ramp up spiking activity prior to specific tests, but at the cost of selectivity as they respond to varying degree to all stimuli. This setup therefore sacrifices stimulus decodability but instead facilitates generalization of the task relevant activity to memory items that have not been previously encountered in the task. Despite a mix of neurons with distinct stimulus preferences in a given location, all neurons ramp up. In our model, the ramp up comes from stimulus independent excitation, elevated in a particular cortical location. This gives a different prediction for how beta, gamma and spiking should relate to one another compared to the "broad tuning model".

In the task, there is a sequence of two images to be held in WM and then compared to a sequence of two test stimuli. Thus, there two aspects of WM information, the images themselves as WM items and their task context, i.e., whether an image is first or second in the sequence and thus whether it is expected as the first or second test stimulus (Figure 2). Our model predicts that these two different types of information should be reflected differently in neural activity (whereas "broad tuning model" does not predict this distinction). Spiking should primarily carry the stimulus item information. It reflects the connectivity of single neurons within each "pocket" of gamma excitation and thus is the high-dimensional component that carries image information (Figure 2C "Identity"). The gamma is associated (correlated) with this spiking because it reflects the shared local excitation that controls the spiking. At the same time, gamma bursting in a given network location is part of a larger pattern of anti-correlated gamma/beta pattern that encodes the task operations. Thus gamma (and beta) power should reflect the lower dimensional task component, i.e., the task context (Figure 2C "Order"), which then gets inherited by the spiking such that it expresses a mix of high and low-dimensional activity.

To test these predictions, we first examined spiking vs gamma power in the delay (1000 ms) leading up to the first and second test stimuli. The data was labelled by the task context (Test 1 or Test 2) and by stimulus identity (which stimulus was expected based on the sequence held in WM, i.e., the first stimulus in the sequence as Test 1 and the second as Test 2). We then calculated the percentage of variance (PEV) explained by these labels. As predicted by the Hot-Coal model, the ramping up of gamma bursting carried information mainly about the task component, i.e., whether it was the lead up to the first or second test stimulus, and not the identity of the stimuli. The spiking instead carried a mixture of the two factors. This suggests that the task information is expressed in spatial patterns of excitation on a larger spatial scale (at the level of a few hundred micrometers contributing to the gamma signal) compared to the stimulus information (individual neurons), which is consistent with the model (Figure 1). We also found a similar difference between spiking and beta activity in line with the prediction about the beta as a correlate of task information (Figure S2).

We further tested this prediction by making use of the fact that monkeys performed also another WM task (Figure 2A, D). In Task 2, two cues were again serially presented, just as in Task 1, but then tested in parallel in a single test. Even though both tasks are identical up until the testing period, earlier analysis of this data has shown that neurons have different spiking ramp up leading up to the first test cue (Test 1) depending on what is required at the test itself (Warden and Miller, 2010). The model predicts further

that this difference between the tasks should also be observed in gamma bursting, whereas the pre-test ramp up in spiking should not only reflect the task difference but also carry information about stimulus identity. This is indeed what we found when analyzing PEV (Figure 2E).

We further elaborated on these results by using demixed principal component analysis (dPCA) of the two tasks (Figure 3), which decomposes population activity into principal components dependent on task parameters (Tang et al., 2016). We identified low-dimensional task and condition-independent components as well as high-dimensional stimulus dependent components (measures differences across stimuli) for both gamma and spiking. The task components captured differences between the two tasks, and the condition-independent components accounted for shared patterns of activity over time. As predicted, these low-dimensional components were more dominant in gamma bursting (Figure 3 left), whereas spiking had additionally strong stimulus-related components (Figure 3 right).

To control that the observed difference between gamma and spiking did not simply reflect quality in the measures making stimulus information difficult to read out from gamma, we performed the same analysis on a third task (Figure 4A). In this task, the items were presented in distinct locations of the visual field and not foveally, as in the previous two tasks. Due to this, also item information should have a spatial organization and the model predicts that there will not be a distinction between gamma and spike information. This is indeed what we found (Figure S3), This suggests that the observed differences between the gamma and spiking above reflect distinct spatial organization of task and item information, beyond a common notion that gamma is merely a low-pass filtered spiking activity. Consistent with the model, we also found dominance of the low-dimensional task as well as condition-independent activity in beta bursting (Figure S4).

Low-dimensional spatial excitation gives rise to task-dependent selectivity

Neurons often respond non-linearly to a combination of experimental factors, with increased activation only to a specific combination of task context and sensory inputs (Goldman-Rakic, 1996; Rigotti et al., 2013; Stringer et al., 2019). As a result of such rich representations, it is possible to find a simple linear read out from population activity to extract task and stimulus information in order to solve various tasks. This type of selectivity could in principle be achieved in a RNN model by adding random connections that "scramble" inputs such that each neuron receive a unique combination of contextual and sensory information. However, according to the Hot-Coal model the scrambling instead happens due to low-dimensional external excitation interfering with each neuron's selectivity (Figure 1D). We argue that this separation has functional advantages as 1) learning a low-dimensional pattern for one task does not change representations for another task thereby facilitating task-switching, and 2) training does not become object dependent but generalizes to novel objects.

To test the idea that low-dimensional excitation explains mixed selectivity in single neurons we analyzed spiking and LFPs from the third WM task performed by a different pair of monkeys (Task 3, Figure 4A). The model predicts that changes in excitation levels resulting from sub-cortical inputs should alter selectivity of single neurons between task epochs. We have previously shown that beta oscillations are anti-correlated with gamma and spiking (Lundqvist et al., 2016; 2018a), with beta bursting being suppressed during cues (Figure 4B). We analyzed beta and gamma bursting as proxies for local excitation.

Particularly beta, since it expresses less cross-talk between spiking and field activity than gamma, should signal excitation levels distinguishable from spiking itself.

Interestingly, the levels of beta activity had a distinct shift downwards before and during test periods compared to encoding cue periods (Figure 4B), suggesting a general shift in excitability. It was not due to saccades, as it was seen equally during test epochs in which animals responded or withheld the response (compare Test 1 and Test 2 in Figure 4B). This implies that excitation levels indexed by LFPs were generally elevated during test epochs compared to encoding. Sites with neurons selective to encoding cues had lower beta during encoding than sites with non-selective neurons (Figure 4B). However, during the test epochs beta for non-selective sites was consistently lower compared to the lowest beta observed at selective sites during encoding. Indeed, in trials where the same items were presented during encoding and test epochs, gamma power (Figure S5) and firing rates were increased during the test period. Importantly, information about the objects was confined to a smaller part of the network during cue encoding compared to the test epochs (Figure 4B, right). A significantly larger portion of the neurons responded in test epochs even in trials with identical cue and test stimuli (62/496 had significant variance explained by cue identity compared to 152/496 by test identity, at p=0.05, corrected for multiple comparisons). This suggests that controlling the spatio-temporal patterns of excitation levels could be an easy way to distinguish sample and test information in the selective neural population, in line with the proposed model.

Discussion

We have presented a conceptual model of flexible working memory, metaphorically referred to as a Hot-Coal model, and tested predictions in neurophysiological data from rhesus macaques. The model relies on spatial computing that utilize spatial dimensions to perform control/related computations. It suggests independent sources for task-related and WM item-related activity in PFC. We propose that this scheme allows for separate learning of task rules and memory representations themselves, and therefore facilitates transfer learning to new stimuli in the spirit of zero-shot learning, i.e. not used during task training. Importantly, we have found evidence for the independent sources by comparing gamma and spiking activity. It suggested that task related activity is shared among neurons within a few hundred micrometers, whereas stimulus selectivity has more of a distributed organization. This is consistent with the suggested model, where stimulus selectivity arises from the neuron-to-neuron recurrent connectivity, and task activity from spatially structured subcortical inputs.

How can this lead to generalized working memory? In the Hot-Coal model, sustained memory traces in PFC rely not only on recurrent excitation and synaptic traces, but also on the increased excitation from sub-cortical structures including thalamus and basal ganglia (O'Reilly and Frank, 2006; Schmitt et al, 2017). While specific content of the delay activity relies on precise PFC connectivity, it is not self-sustained in the absence of subcortical inputs. In support of this view, there is growing evidence suggesting that excitation from mediodorsal thalamus is needed for sustained working memory and attention activity in PFC (Bolkan et al., 2017; Schmitt et al., 2017). Importantly, this does not rule out attractor-like activity in PFC, which has so far been very successful in describing several aspects of delay activity (Wimmer et al., 2014, Panichello et al., 2019; Lundqvist et al., 2016). However, such attractor activity may require content

independent excitation from thalamus and basal ganglia that provides stable dynamics in part the PFC network. The recurrent excitation within PFC on its own is not strong enough to sustain attractors and thus memory traces quickly decay. In this way, our model suggests that subcortical excitation can help control when and where on the cortical sheet information is expressed with the attractor-like dynamics.

As an example, let's consider Tasks 1 and 2 in our data where two items were presented and tested in sequence. If the pattern of excitation onto PFC repeats from trial to trial, the physical locations of sites carrying information about the item presented as first and second cues are predictable and therefore the items can be selectively accessed. This is independent of the identity of these items and details in how they are maintained by the PFC connectivity. We refer to this as spatial computing. Apart from spatio-temporal patterns of excitation spatial computing also requires redundancy in cortex, such that the same information about a stimulus can be maintained in multiple parts of PFC. Such redundancy seems to be a hallmark of cortical activity (Siegel et al., 2015; Stringer et al., 2019), and could be supported by horizontal connections between neurons with shared stimulus preference (Goldman-Rakic, 1996; Kisvarday et al., 1997; Ko et al., 2013).

Our Hot-coal model tackles a challenging problem of WM generalization, which is evidenced by human and animal behavior. Importantly, besides the functional aspects, it also ties together many aspects of known cortical dynamics, including redundancy of coding, the dominance of low-dimensional activity, mixed selectivity, the role of bursty gamma and beta oscillatory activity, as well as transitions between near silent and sustained delay activity. Below we expand on this.

The dimensionality of cortical representations has recently become a topic of increased interest (Rigotti et al., 2013; Stringer et al., 2019; Badre et al., 2020; Cueva et al., 2019; Wolff et al., 2020; Kobak et al., 2016; Tang et al., 2019; MacDowell et al., 2020). Low-dimensional activity, i.e. activity shared across many neurons and conditions, has been implicated in generalizability (Badre et al., 2020) while high-dimensional activity is needed for selective encoding of events. Low-dimensional activity often reflects task-related activity such as modulation between task epochs (Tang et al., 2019). It has recently been shown to correlate with task learning and has been considered to provide important temporal information to working memory (Tang et al., 2019; Cueva et al., 2019; Wolff et al., 2020). The key novel aspect of the Hot-Coal model and the supporting experimental findings is the proposed mechanisms for how high- and low-dimensional activities arise and interact. Importantly, we introduce a spatial aspect of the low-dimensional signal and suggest that it arises from sub-cortical excitation. We argue that these original components of our theory explain how desirable generalization can be obtained in WM function.

It was noted already by Goldman-Rakic et al., 1996 that single neurons display task-dependent activity, where some for instance only respond only to sample cues, others only to the test cues. Furthermore, it has recently been shown that recurrent neural networks trained on more complex tasks such as delayed-match-to-sample tasks form multiple functional neuronal sub-groups (Dubreuil et al., 2020). In particular, one group was proposed to encode the sample cue and another – group test cue to facilitate the comparison. We observed a similar functional organization, where excitation controlled which sub-population of neurons encoded sample and test objects even if they had the same identity. As predicted by our model, these sub-groups of neural ensembles were spatially organized. Theoretically, this solution is functionally appealing since the recurrently connected sub-groups do not have to be expanded to account for new stimuli, unseen in the task training process, but all neurons, regardless of stimulus preference, can become part of a sub-group based on their spatial location.

The concept of spatial computing extends and complements our prior work (Lundqvist et al., 2016; Miller et al., 2018). There we reported that interspersed lower-frequency alpha/beta bursts in LFPs are negatively correlated with the bursts of spiking (and gamma) that are informative about WMs (Lundqvist et al., 2016). At times when information had to be encoded or accessed, beta bursting was suppressed and gamma bursting elevated. Here the beta and gamma bursting has a unique spatio-temporal pattern for each task, giving rise to the associated low-dimensional activity but at the same time highly dynamic activity on single trials. Indeed, the low-dimensional components were clearly visible in the beta and gamma bursting, more pronounced than in spiking itself.

There is an ongoing debate whether WM storage relies solely on spiking or a combination of spiking and synaptic mechanisms (Sreenivisan and D'Esposito, 2019; Constantinidis et al., 2018, Lundqvist et al., 2011; 2018b, Mongillo et al., 2008; Wolff et al., 2017; Sandberg et al., 2003). Our model suggests the latter. However, depending on the current needs spiking activity may become more or less persistent depending on the sub-cortical excitation. This could be used, for instance, to prioritize one of the items held in working memory, which is currently relevant (Warden and Miller, 2010, Lundqvist et al., 2018a, Wolff et al., 2017; Yu et al., 2020). In this scheme, a subgroup of neurons in part of the network encoding the relevant item receive stronger excitation resulting in the elevated spiking, which expresses the WM information. At the same time, information about non-prioritized items should to a larger degree be maintained by synaptic mechanisms.

In sum, the Hot-Coal theory postulates a novel mechanism for how networks may achieve flexible working memory with powerful generalization capabilities. At the same time the proposed mechanism of spatial computing reconciles multiple seemingly contradictory neurophysiological findings.

Figures

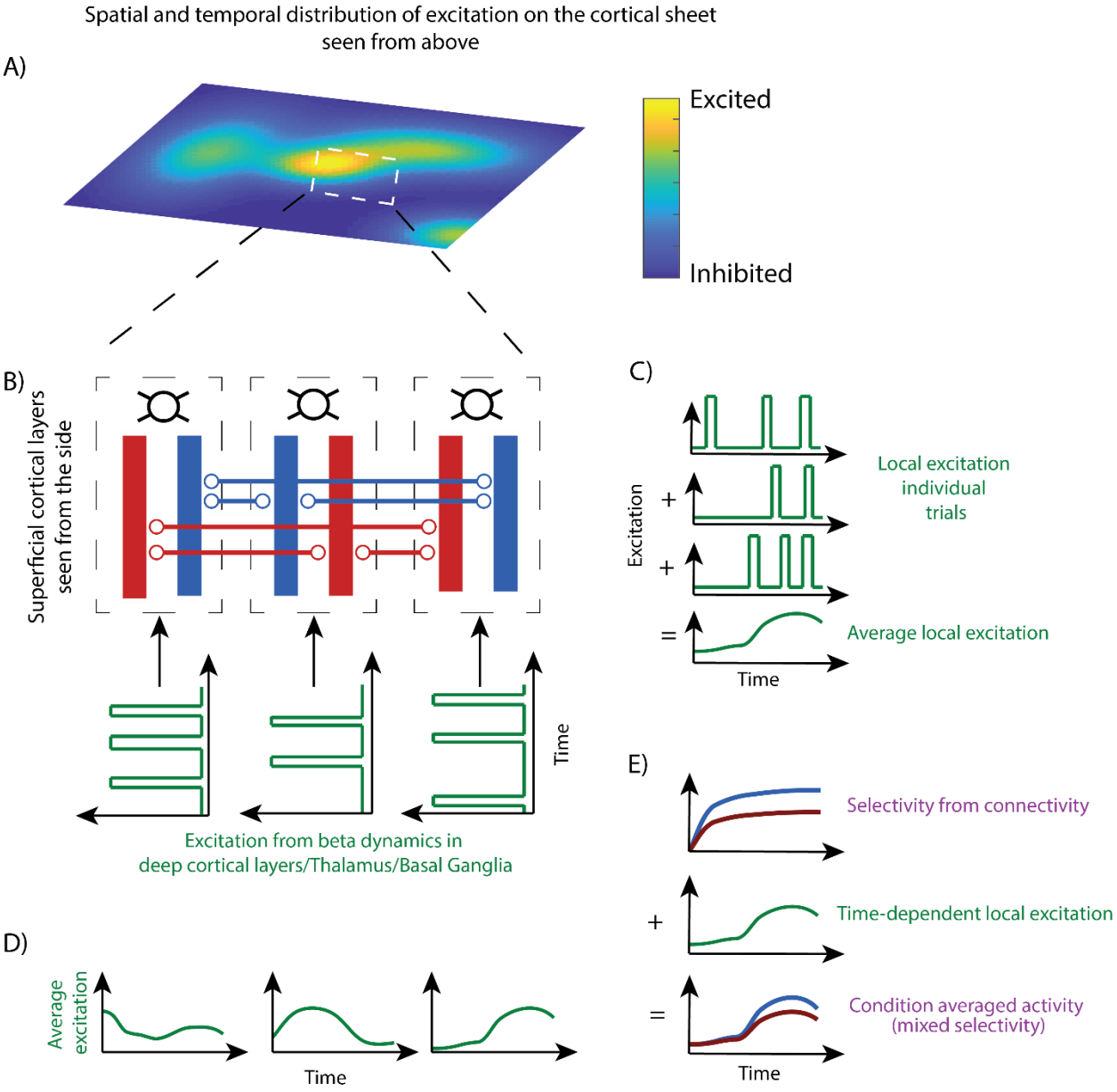
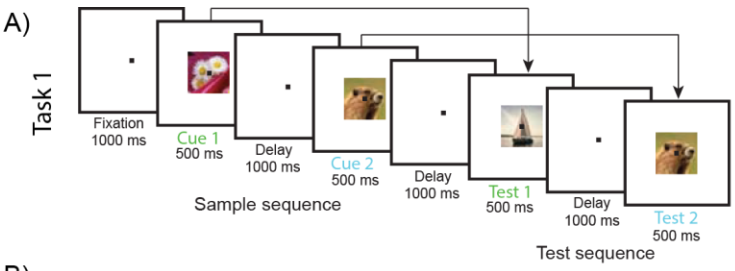
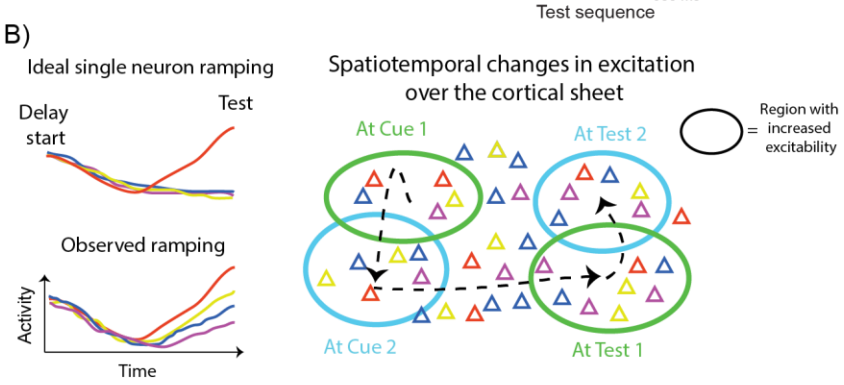
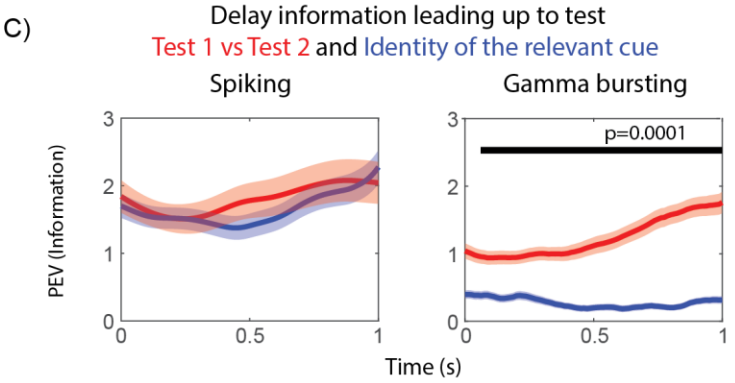


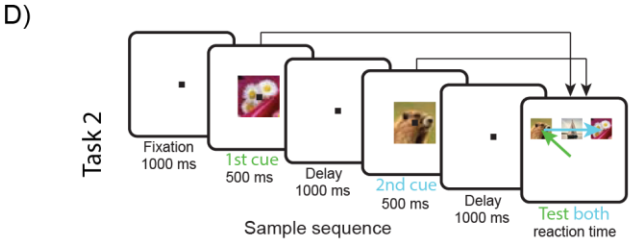
Figure 1. *Conceptual figure of the Hot-Coal model. A) The excitation levels in different parts of the cortical sheet have time-varying, task-dependent patterns of activity. B) The model consists of a cortical sheet of 'modules' (dotted rectangles) containing multiple local populations (red and blue rectangles) coding for distinct stimuli. There are lateral connections between neurons with similar object coding preferences (Goldman-Rakic, 1996). This creates a redundancy where the same information (red vs blue) can be represented in multiple modules. Locally, within each 'module' red and blue populations compete through shared inhibitory populations (black neurons). This connectivity scheme implements a local winner-take-all (WTA) dynamics and generates gamma oscillations (Lundqvist et al., 2010; 2006). Each module also receives trains of excitatory bursts (bottom), like oxygen puffs. These puffs contain no information about the stimulus held in WM, but 'ping' the local network, which responds by 'ringing' back with the most excitable local population through the WTA mechanism. C) The excitatory 'oxygen puffs' occur at different*

times, even for identical trials (top rows). But for a given module within the cortical sheet there is a tendency for the puffs to occur at specific parts of the trial, giving rise to a shared (low-dimensional) time-dependent average activity pattern within each module (middle row). D) There is a unique stream of puffs to each module of the network leading to a unique local trial-averaged activity pattern over time. This corresponds to learned task-related activity. It directs the flow of information spatially within the redundant network such that objects can be held spatially segregated in an object-independent and predictable way. E) Each neuron has a stimulus preference given by the network connectivity (top row). The model predicts that complex behavior (bottom row) can arise from combining preference with the low-dimensional excitation (middle row).









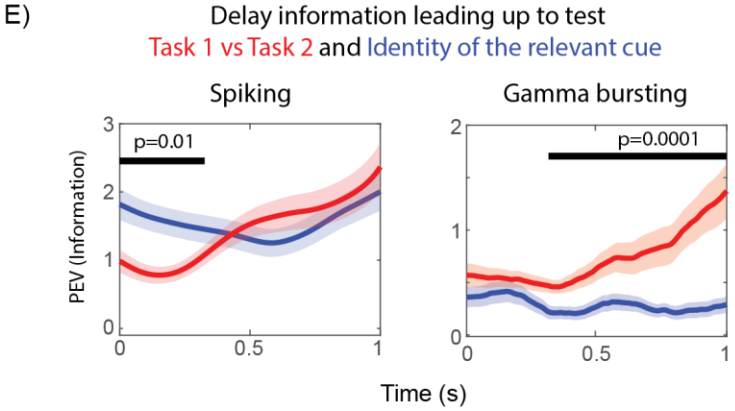


Figure 2. Disassociation between gamma and spiking. *A) In Task 1, two cue objects are presented and then tested sequentially. The order as well as identity of the objects have to be preserved between cue and test in order for the trial to be a match trial. B) Left: Activity ramps up before the information about cue identity is used in the task. In the idealized case, single neuron activity only ramps up for one specific cue. However, neurons with ramping activity respond to all cues to varying levels. Right: the model explains the observed phenomenon of post stimulus ramping after all stimuli to varying degrees by external time-dependent excitation that travels over the cortical sheet (see also Figure 1A). Neurons preferring the red or blue cues are scattered throughout the network. Some sites exhibit ramping excitation, reflected in increased activity of all neurons at that site, though to the lesser degree when the preferred cue was not provided. C) Red plots show PEV accounting for test order effects estimated over two groups of epochs, preceding either Test 1 or Test 2. Blue curves reflect PEV wrt. cue identity (information about identity about Cue 1 prior to Test 1 and Cue 2 leading up to Test 2). Black bars demonstrate when blue and red plots differ, using cluster-based statistics ($p<0.05$). D) Task 2 structure, which is identical to Task 1 until Test 1. Unlike in Task 1 however, at Test 1 both Cue 1 and Cue 2 are tested in parallel. E) PEV information pooled over Task 1 and 2 in the period leading up to Test 1. Red plots show PEV estimated over epochs grouped based on task, whereas blue curves reflect PEV information about the cue identity (average of information about Cue 1 and Cue 2).*

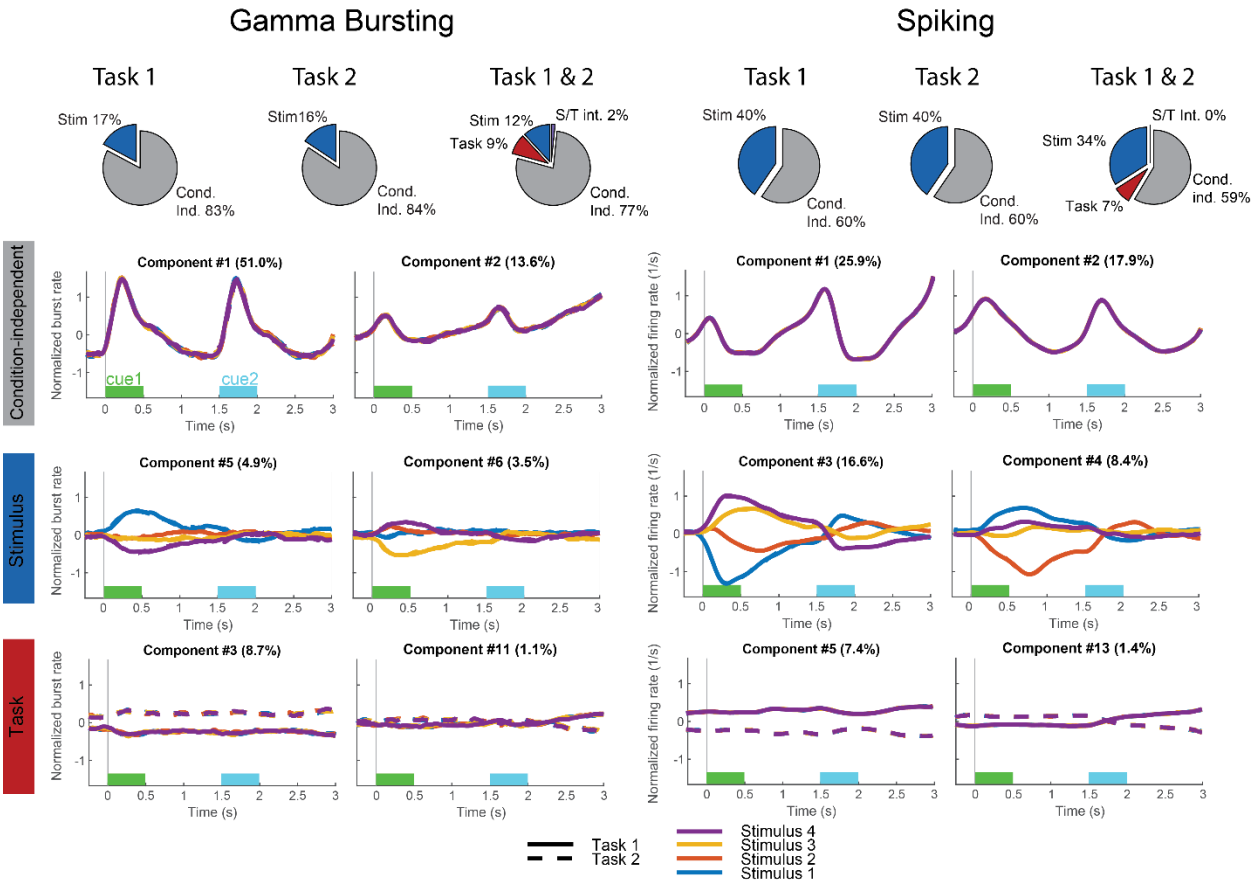


Figure 3. Demixed PCA of gamma and spiking. *The model predict that task information (task and condition independent components) should be more prevalent in gamma bursting as compared to spiking (see Figure 2). Here we used dPCA (Kobak et al., 2016) to extract the principal components and attribute them to task and cue specific activity. "Condition-independent" components correspond to low-dimensional time-dependent task activity, "Task" components explain the variance that originates from the differences between Tasks 1 and 2. "Stimulus" components account for the variance between four different stimuli.* The bottom half of the figure show example components for Task 1 and Task 2 combined.

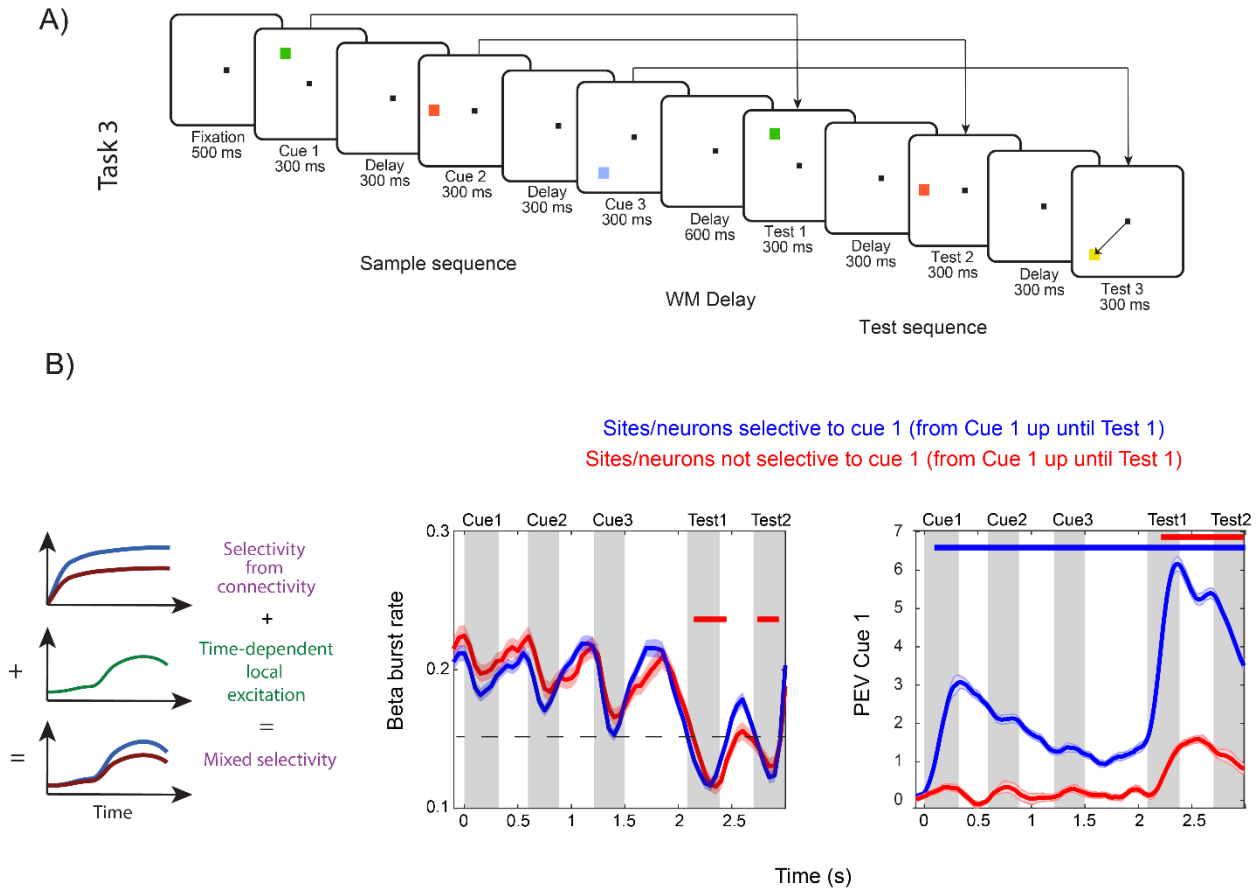


Figure 4. Mixed selectivity arising from time-dependent excitation. *A) In Task 3, two or three colored squares are presented in a sequence. Their location and color are to be remembered. Following the 0.6 s delay, the sequence of squares is repeated. Monkeys should saccade to the square in the test sequence with the color changed relative the sample sequence. B) Left: the model predicts that mixed selectivity arises from a combination of stimulus selectivity (given by connectivity) and the unique pattern of local excitation. Middle: beta bursting tracks (is anti-correlated with) the excitation in the network. Blue curves correspond to sites in which neurons are selective to Cue1-Cue3, red curves describe sites with no such selective neurons. Here trials with three sequential cues and two test stimuli (monkeys correctly respond at Test 2) are shown. At Test 1 beta bursting is globally suppressed, despite it being behaviorally identical to Cue 1. At the same time, information about Cue 1 stimulus (the same stimulus is provided at Cue 1 and Test 1 in this case) is more widespread in the network. Previously unselective sites are now selective, enabling a dissociation between Cue and Test information.*

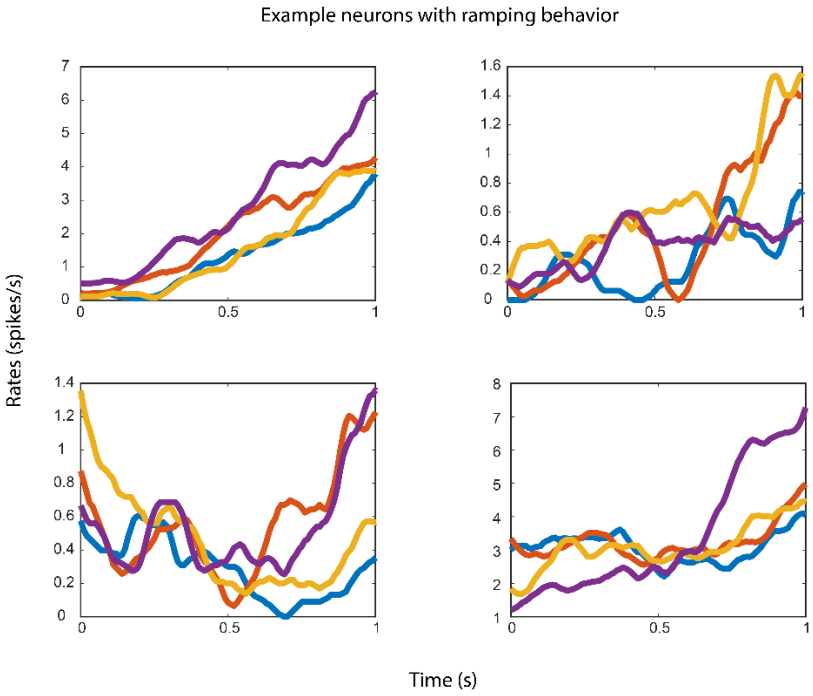


Figure S1. Single neuron examples of ramping activity. Shown are the activity grouped by cue identity (blue/red/yellow/purple) for all four neurons with ramping activity from two random sessions.

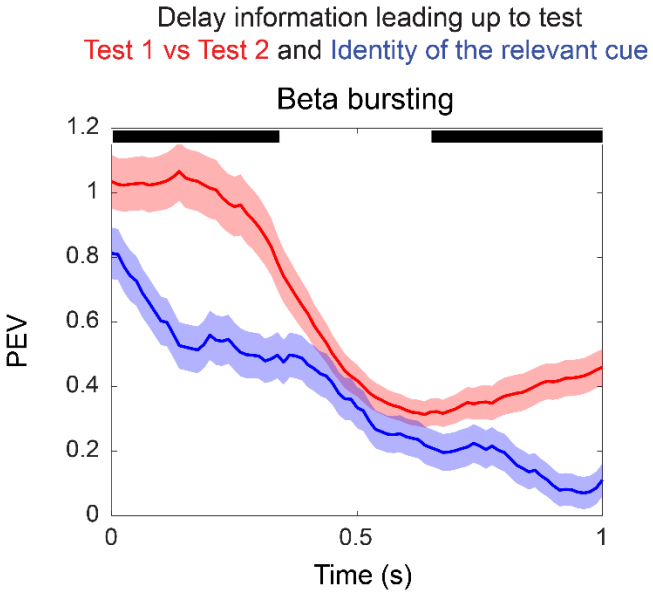


Figure S2. *Beta information about cue and upcoming test. Red plots show PEV accounting for test order effects estimated over two groups of epochs, preceding either Test 1 or Test 2. Blue curves reflect PEV wrt. cue identity (information about identity about Cue 1 prior to Test 1 and Cue 2 leading up to Test 2). Black bars demonstrate when blue and red plots differ, using cluster-based statistics (p<0.05).*

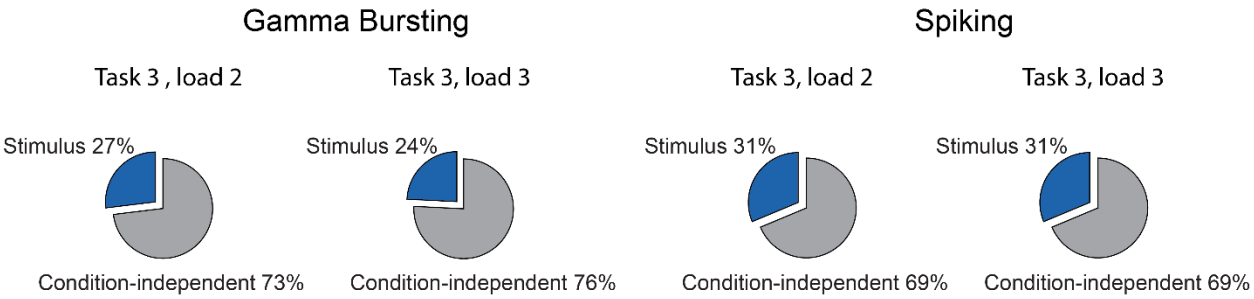


Figure S3. dPCA analysis of Task 3. Shown are the proportion of variance that can be attributed to stimulus (blue) and condition-independent (grey) components for gamma (left) and spiking (right). Due to their distinct task structures, load 2 and load 3 were analyzed separately.

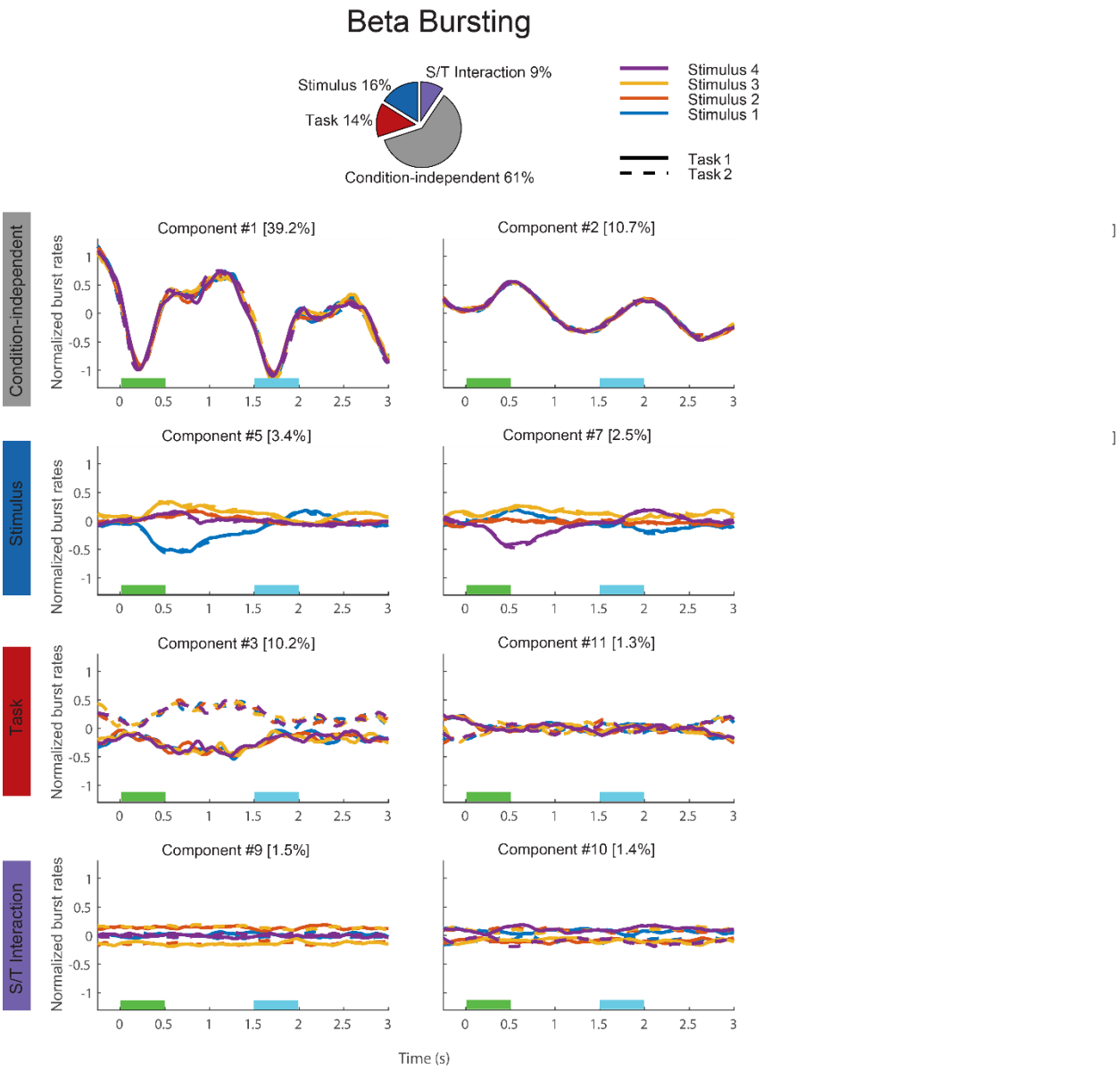


Figure S4. dPCA analysis of beta bursting. Same as in Figure 3 but for calculated over patterns of beta bursting.

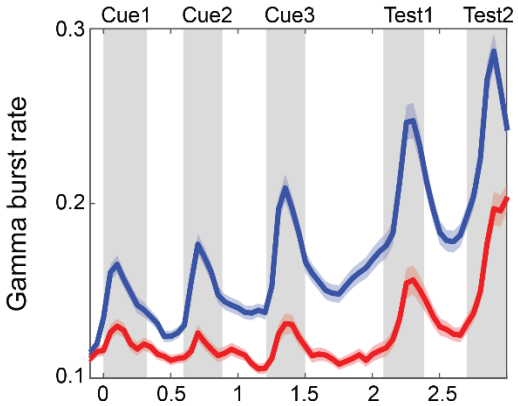


Figure S5. Same as Figure 4B, but for gamma bursting.

Methods

We analyzed data from two previous studies (Warden and Miller, 2010; Lundqvist et al., 2016). In total the two studies included three experimental tasks (task 1 & 2 from Warden and Miller, 2010; task 3 from Lundqvist et al., 2016). For details of training and data collection, please see those studies. Briefly, each task involved two Rhesus macaques that were trained until they performed well above chance. They were trained with positive reward (juice) only and maintained in accordance with the National Institutes of Health guidelines and the policies of the Massachusetts Institute of Technology Committee for Animal Care).

For each recording, a new set of acute electrode pairs (tungsten, epoxy-coated, FHC) was lowered through a grid. Between 8 and 20 prefrontal electrodes were recorded from simultaneously on each session (34 sessions for Task 1 and 2, 30 sessions for Task 3). Task 1 and Task 2 were recorded on the same sessions in a blocked design. Only electrodes containing isolatable units were kept for further analysis.

Signal Processing

Preprocessing Task 1 and 2: At first, all electrodes without any isolatable neurons were removed. Then, a notch filter with constant phase across a session was applied to remove 60-Hz line noise and its second harmonic. On some sessions there were high-power, broadband frequency artifacts; these sessions were discarded from further analysis.

Preprocessing Task 3: We first removed apparent noise sources from the signal. In particular, a notch filter was applied to remove 60-Hz line noise with constant phase across a session. In addition, we removed periodic deflections seen in the evoked potentials (every 47 ms, lasting 1 ms, on a subset of electrodes, phase locked to stimulus onset). The signal was filtered and downsampled to 1 kHz (from 30kHz).

For spectral analysis we applied multi-taper analysis (with a family of orthogonal tapers produced by Slepian functions; Slepian, 1978; Thomson, 1982; Jarvis and Mitra, 2001). The multi-taper approach was adopted with frequency-dependent window lengths corresponding to six to eight oscillatory cycles and frequency smoothing corresponding to 0.2–0.3 of the central frequency, f0, i.e., f0 ± 0.2f0, where f0 were sampled with the resolution of 1 Hz (this configuration implies that two to three tapers were used). The spectrograms were estimated with the temporal resolution of 1 ms. Typically we present total power of raw LFPs (after removal of noise) without subtracting any baseline or estimated evoked content.

Burst Extraction

To extract bursts of high power events on a single trial level we utilized a previously developed method (Lundqvist et al., 2016; 2018a). In the first step of the oscillatory burst identification, a temporal profile of the LFP spectral content within a frequency band of interest was estimated. We used two alternative methods of spectral quantification (see above). We either narrow-band-filtered LFP trials and extracted the analytic amplitudes (envelope) or we used single-trial spectrograms, obtained with the multi-taper approach, to calculate smooth estimates of time-varying band power (all presented results were obtained with the multi-taper approach; the results for the two methods were very similar). Next we defined

oscillatory bursts as intervals during individual trials when the respective measure of instantaneous spectral power exceeded the threshold set as two SDs above the trial mean value for that particular frequency, and with the duration of at least three cycles. Having the burst intervals extracted for the beta band (20–35 Hz) and three gamma sub-band oscillations (40–65, 55–90, and 70–100 Hz) from each trial, we defined a single-trial point process (binary state: no burst vs burst within a 10-ms window) with the resolution of 10 ms and trial-average measure, a so-called burst rate for each spectral band. This quantity corresponds to the chance of a burst occurrence on an individual electrode at a particular time in the trial (a proportion of trials where a given electrode displays burst-like oscillatory dynamics around the time point of interest sliding over the trial length).

Statistical Methods

The majority of tests performed in this study were nonparametric due to insufficient evidence for model data distributions. To address the multi-comparisons problem, we employed Kruskal-Wallis, Friedman's, and Wilcoxon's signed-rank tests where appropriate. In addition, for the comparison between temporal profiles of the normalized firing rates within versus outside oscillatory bursts, we resorted to a permutation test on the largest cluster-based statistics (Maris and Oostenveld, 2007), originally proposed to increase the test sensitivity based on the known properties of the data (here being temporal dependency). Finally, some attention should be given to the way we report correlations between the measures of time-varying spectral band content and burst rate statistics. The correlation analyses were performed on individual electrodes and only the summary statistics (mean and SE) for the electrode-wise significant effects ($p < 0.01$) are presented.

Estimation of information

The bias-corrected PEV (Olejnik and Algina, 2003) was estimated across trials with different conditions from firing rates averaged in 50-ms bins across trials within each trial. We performed two-way ANOVA where trials had multiple groupings (i.e. stimulus or delay/task). All correct trials were used, as the groups were well balanced each session. The bias correction was used as it avoids the problem of non-zero mean PEV for small sample sizes.

As a result, (bias-corrected) PEV allowed for the quantification of information carried by the modulation of firing rates or burst rates of individual units accounting for the stimulus, task or task epoch (delay 1 vs delay 2).

Demixed principal component analysis

To identify low-dimensional manifold for neural activity, we performed a demixed principal components analysis (dPCA) (Kobak et al., 2016). This approach allows not only for compressing the data, similarly to PCA, but also separates the underlying components with respect to the requested task parameters by demixing the dependencies of the population activity on the task parameters. In a nutshell, demixing is achieved by minimising the reconstruction error between the projections and the neural activity averaged over trials (unlike in PCA where the reconstruction error on single trials is minimised) and over the

requested task parameters. In addition, when compared to PCA the method used here benefits from greater flexibility offered by using two different linear mappings for encoding vs decoding. More technical as well as theoretical details of dPCA can be found in (Kobak et al., 2016).

In our analyses dPCA was applied to both spiking data (firing rates obtained by convolving the spike point process with 50-ms wide Gaussian kernel) and oscillatory bursts in beta and gamma bands (burst point process convolved with 50-ms wide Gaussian kernel). To achieve demixing effect we grouped trials into task (Task 1 vs Task 2)- and stimulus (Cue 1)-dependent sets, and analyzed trials in the interval from 100 ms prior to the first sample cue (Cue 1) until the first test cue (Test 1). Apart from task and stimulus-dependent components, dPCA also produced a condition independent component corresponding to low-dimensional time-dependent task activity.

Data availability.

All relevant data and code will be available from the corresponding author on reasonable request.

Acknowledgements

We would like to thank the Swedish Research Council Starting Grant 2018-04197, the 2017 Young Investigator Grant from the Brain & Behavior Research Foundation, the National Institutes of Mental Health Grant R37MH087027 and 5K99MH116100-02, Office of Naval Research Multidisciplinary University Research Initiatives Grant N00014-16-1-2832 for their support.

References

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Badre, D., & Wagner, A. D. (2004). Selection, integration, and conflict monitoring: assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron*, *41*(3), 473-487.

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2020). The dimensionality of neural representations for control. https://doi.org/10.31234/osf.io/asdq6

Bolkan, S. S., Stujenske, J. M., Parnaudeau, S., Spellman, T. J., Rauffenbart, C., Abbas, A. I., ... & Kellendonk, C. (2017). Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature neuroscience*, *20*(7), 987.

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930-942.

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current opinion in behavioral sciences*, *1*, 23-31.

Cueva, C. J., Saez, A., Marcos, E., Genovesio, A., Jazayeri, M., Romo, R., ... & Fusi, S. (2019). Low dimensional dynamics for working memory and time encoding. *bioRxiv*, 504936.

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X. L., Wang, M., & Arnsten, A. F. (2018). Persistent spiking activity underlies working memory. *Journal of Neuroscience*, *38*(32), 7020-7028.

D'esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, *378*(6554), 279-281.

Dotson, N. M., Hoffman, S. J., Goodell, B., & Gray, C. M. (2018). Feature-based visual short-term memory is widely distributed and hierarchically organized. *Neuron*, *99*(1), 215-226.

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2020). Complementary roles of dimensionality and population structure in neural computations. *bioRxiv*.

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. Proceedings of the National Academy of Sciences, 93(24), 13473-13480.

Jarvis, M. R., & Mitra, P. P. (2001). Sampling properties of the spectrum and coherency of sequences of action potentials. *Neural computation*, *13*(4), 717-749.

Ketz, N. A., Jensen, O., & O'Reilly, R. C. (2015). Thalamic pathways underlying prefrontal cortex–medial temporal lobe oscillatory interactions. *Trends in neurosciences*, *38*(1), 3-12.

Ko, H., Cossell, L., Baragli, C., Antolik, J., Clopath, C., Hofer, S. B., & Mrsic-Flogel, T. D. (2013). The emergence of functional microcircuits in visual cortex. *Nature*, *496*(7443), 96-100.

Kisvárday, Z. F., Toth, E., Rausch, M., & Eysel, U. T. (1997). Orientation-specific relationship between populations of excitatory and inhibitory lateral connections in the visual cortex of the cat. *Cerebral cortex (New York, NY: 1991)*, *7*(7), 605-618.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., ... & Machens, C. K. (2016). Demixed principal component analysis of neural population data. *Elife*, *5*, e10989.

Lewis-Peacock, J. A., Drysdale, A. T., & Postle, B. R. (2015). Neural evidence for the flexible control of mental representations. *Cerebral Cortex*, *25*(10), 3303-3313.Lundqvist et al., 2006

Lundqvist, M., Compte, A., & Lansner, A. (2010). Bistable, irregular firing and population oscillations in a modular attractor memory network. PLoS Comput Biol, 6(6), e1000803.

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

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and beta bursts underlie working memory. *Neuron*, *90*(1), 152-164.

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018a). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature communications*, *9*(1), 394.

Lundqvist, M., Herman, P., & Miller, E. K. (2018b). Working memory: delay activity, yes! Persistent activity? Maybe not. *Journal of Neuroscience*, *38*(32), 7013-7019.

Lusk, N., Meck, W. H., & Yin, H. H. (2020). Mediodorsal thalamus contributes to the timing of instrumental actions. *Journal of Neuroscience*.

MacDowell, C. J., & Buschman, T. J. (2020). Low-Dimensional Spatiotemporal Dynamics Underlie Cortex-wide Neural Activity. *Current Biology*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177-190.

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, 1.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463-475.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543-1546.Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283-328.

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nature communications*, *10*(1), 1-11.

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585-590.

Sandberg, A., Tegnér, J., & Lansner, A. (2003). A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems*, *14*(4), 789-802.

Schmitt, L. I., Wimmer, R. D., Nakajima, M., Happ, M., Mofakham, S., & Halassa, M. M. (2017). Thalamic amplification of cortical connectivity sustains attentional control. *Nature*, *545*(7653), 219-223.

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. Science, 348(6241), 1352-1355.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis, and uncertainty—V: The discrete case. *Bell System Technical Journal*, *57*(5), 1371-1430.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. Science, 364(6437).

Sreenivasan, K. K., & D'Esposito, M. (2019). The what, where and how of delay activity. *Nature Reviews Neuroscience*, *20*(8), 466-481.

Tang, H., Qi, X. L., Riley, M. R., & Constantinidis, C. (2019). Working memory capacity is enhanced by distributed prefrontal activation and invariant temporal dynamics. *Proceedings of the National Academy of Sciences*, *116*(14), 7095-7100.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, *70*(9), 1055-1096.

van Ede, F., Niklaus, M., & Nobre, A. C. (2017). Temporal expectations guide dynamic prioritization in visual working memory through attenuated α oscillations. *Journal of Neuroscience*, *37*(2), 437-445.

Wang, Y., Markram, H., Goodman, P. H., Berger, T. K., Ma, J., & Goldman-Rakic, P. S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nature neuroscience, 9(4), 534-542.

Warden, M. R., & Miller, E. K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. *Journal of Neuroscience*, *30*(47), 15801-15810.

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, *17*(3), 431-439.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature neuroscience*, *20*(6), 864-871.

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. *PLoS biology*, *18*(3), e3000625.

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS biology*, *18*(6), e3000769.