Prefrontal oscillations modulate the propagation of neuronal activity required for working memory

Jason S. Sherfey[1,2,3*§], **Salva Ardid**[3], **Earl K. Miller**[2], **Michael E. Hasselmo**[1], **Nancy J. Kopell**[3*]

***For correspondence:** sherfey@bu.com (JSS); nk@bu.com (NK)

Present address: [§]610 Commonwealth Ave., Boston, MA, 02215

[1]Center for Systems Neuroscience, Department of Psychological and Brain Sciences, Boston University, MA 02215; [2]The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139; [3]Department of Mathematics and Statistics, Boston University, Boston, MA 02215

Abstract

Cognition involves using attended information (e.g., stimuli, rules, responses), maintained in working memory (WM), to guide action. During a cognitive task, a correct response requires flexible, selective gating so that only the appropriate information flows at the proper time from WM to downstream effectors that carry out the response. Much evidence suggests that WM information is encoded in the firing rates of populations of neurons in prefrontal cortex (PFC). At the same time, many experiments have demonstrated separate, task-related modulation of oscillatory dynamics in PFC networks. In this work, we used biophysically-detailed modeling to explore the hypothesis that network oscillations, leveraging lateral inhibition, can independently gate responses to rate-coded items in working memory. Consistent with recent data, we modeled the superficial layers of PFC as a WM buffer that stores task-relevant information and the deep layers of PFC as an output gate that flexibly governs which information in the WM buffer is propagated downstream to guide action. Items in WM were stored in populations of neurons that had either asynchronous spiking or correlated spiking that was modulated at fast beta/gamma oscillation frequencies. We found that whichever WM item induced a response in the output gate with the shortest period between spike volleys would be most reliably propagated through the output gate. Furthermore, the output gate exhibited network resonance capable of selectively propagating items with resonant oscillatory modulation. We found that network resonance of the deep layer gate can be flexibly tuned by varying the excitability of deep layer principal cells. Our results demonstrate that the propagation of WM-associated neuronal activity can be modulated by tuning either the oscillatory properties of populations encoding WM items, themselves, or the resonant properties of the output gate through which item-encoding activity must propagate to reach downstream effectors. In our PFC model, these dynamics reveal how population rate-coded items embedded in superficial beta and gamma oscillations can be alternately selected by tuning network resonance in the deep layers of PFC depending on task demands. Thus, our model predicts that the experimentally-observed modulation of PFC beta and gamma oscillations could leverage network resonance and lateral inhibition to govern the flexible routing of signals in service of cognitive processes like gating outputs from working memory and the selection of rule-based actions.

Introduction

Many cognitive tasks require flexible gating of information from multiple sources to update internal rule representations and guide responses (***Melrose et al., 2007***; ***Badre and Frank, 2012***; ***Buschman et al., 2012***; ***Hasselmo and Stern, 2018***; ***Bhandari and Badre, 2018***; ***Zhu et al., 2018***). Cognition involves gating of attended information to guide correct responses based on information that has been buffered in working memory (WM) (***Baddeley***

and Hitch, 1974; *Miller, 2000*). The gating must regulate the flow of information about sensory stimuli (*Siegel et al., 2009*), about rules to be applied (*Buschman et al., 2012*; *Bhandari and Badre, 2018*) and feedback about responses (*Funahashi, 2013*).

Many experiments demonstrate that all facets of cognition are associated with the modulation of oscillatory dynamics in networks of neurons (*Cho et al., 2006*; *Tzur and Berger, 2009*; *Siegel et al., 2009*; *Buschman et al., 2012*; *Brincat and Miller, 2016*; *Lundqvist et al., 2016*, *2018*). Much discussion has addressed the potential role of oscillatory dynamics versus neuron spike rates in the gating of cognitively-relevant representations (*Ardid et al., 2010*; *Mante et al., 2013*; *Ardid et al., 2018*). In this work, we used biophysically-detailed modeling to explore the hypothesis that network oscillations, leveraging lateral inhibition, can independently gate responses to rate-coded items in working memory. The oscillatory gating mechanism we identified suggests how beta- and gamma-frequency oscillations could govern the flow of information maintained in WM to govern the selection of actions.

In the brain, WM tasks have implicated the prefrontal cortex (PFC) in all aspects of WM and cognitive control (*Fuster, 1973*, *2015*; *Goldman-Rakic, 1995*; *Miller, 2000*). Furthermore, the maintenance of WM item information has been recently localized to the superficial layers of PFC (*Bastos et al., 2018*). While the maintenance mechanism is debated ((*Miller et al., 2018*); see Discussion), active WM items are believed to be encoded at the population-level in the spike rates of PFC neurons (*Mante et al., 2013*). We can infer based on prefrontal anatomy (*Barbas, 2013*), that WM items encoded in superficial PFC often activate deep layer projection neurons (i.e., principal cells, PCs) to exert top-down influence over sensory and motor systems. Given these observations, we hypothesized that the deep layers of PFC function as an output gate for rate-coded WM items represented in a superficial buffer. Activation of select populations in the output gate would determine when WM items influence downstream circuits and which downstream circuits (i.e., systems and processes) are influenced (*Ardid et al., 2007*; *Kriete and Noelle, 2011*; *Badre and Frank, 2012*; *Kriete et al., 2013*; *Ardid and Wang, 2013*; *Hasselmo and Stern, 2018*; *Zhu et al., 2018*).

Encoding populations in PFC increase synchrony when the encoded information is relevant to the task at hand (*Buschman et al., 2012*); presumably, when the information should be communicated through the output gate. Specifically, *Buschman et al.* (*2012*) showed that the timing of spikes in an encoding population of PFC neurons becomes more or less phase-locked (i.e., synchronous) with fast network oscillations, depending on whether the encoded representation is relevant to rule-based action in the present moment. Furthermore, modulation of beta- and gamma-frequency oscillations has been observed in PFC during WM tasks when items were loaded, maintained, and retrieved from WM (*Siegel et al., 2009*; *Lundqvist et al., 2016*, *2018*; *Bastos et al., 2018*). These findings motivated us to further hypothesize that locking to fast network oscillations determines when and which rate-coded WM representations are communicated through the output gate to downstream effectors.

We have previously shown through detailed modeling that a PFC deep layer with strong feedback inhibition exhibits resonance (i.e., larger responses to preferred input frequencies), and that it prefers fast oscillatory inputs (*Sherfey et al., 2018a*). Resonance in the inhibitory interneurons (INs) endowed the network with a maximum output frequency (i.e., there was a lower bound on the period between spike volleys that the deep layer could generate), and it depended on the synchrony and strength of the activity driving the network.

In this work, we sought to build on those results to explore how a PFC deep layer with strong feedback and lateral inhibition can function as an output gate for rate-coded WM items. Specifically, we were interested in how responses in the output gate depend on the dynamical states of encoding populations (i.e., how the gate output depends on whether WM item-encoding spikes are asynchronous across a population or phase-locked to fast oscillations with varying degrees of synchrony).

We found that, given multiple PC populations in an output gate, whichever population has the shortest period between spike volleys most reliably engages local inhibition which, in turn, suppresses responses in all opposing populations. Therefore, given multiple items in the WM buffer, whichever item induces an oscillatory output with the shortest period between spike volleys (i.e., the highest oscillation frequency) will be the dominant signal that gets through the output gate. A consequence of this is that a WM item with oscillatory modulation at the maximum frequency of the output gate will be most reliably relayed through the gate for downstream readout. In our PFC model, these dynamics yield a mechanism for selective gating of WM items based on shifts of resonance between beta and gamma frequency modulation.

This paper will begin with a simulation study to investigate the mechanisms that underlie oscillatory output gating, their generality, and the ability to flexibly tune the gate to select WM items in different dynamical states. Gating rules and control mechanisms will be summarized in the Discussion, and the paper will close with a detailed look at how oscillatory gating could serve working memory and cognitive control.

Results

This section begins with a review of relevant work on PFC network dynamics and an investigation of the mechanisms that underlie oscillatory output gating. After that we will address how the gate can be tuned to select working memory (WM) items that are asynchronous or oscillatory with different frequencies. Finally, we will demonstrate the utility of our gating mechanism for winner-take-all selection of rate-coded WM items delivered to the output gate along either parallel or convergent (interlaminar) projections.

Background: inhibition-based oscillations and resonance in cortical networks

As was mentioned above, beta2 (20-40Hz, which we will call beta for simplicity) and gamma (40-100Hz) oscillations are modulated in PFC during *in vivo* WM tasks (***Cho et al., 2006***; ***Tzur and Berger, 2009***; ***Siegel et al., 2009***; ***Lundqvist et al., 2016***, ***2018***; ***Bastos et al., 2018***). Neuronal networks will generate oscillations at these frequencies given tonic drive (e.g., steady asynchronous spiking, current injection, or a glutamate receptor agonist) and strong feedback inhibition with inhibitory time constants comparable to those measured from the IPSCs produced by fast spiking, parvalbumin-positive interneurons in PFC (***Adams et al., 2017***; ***Sherfey et al., 2018a***). Such inhibition-based oscillations (***Whittington et al., 2000***) were first studied in networks with only interneurons (INs); see ***Whittington et al.*** (***1995***) for *in vitro* experiments and ***Wang and Buzsáki*** (***1996***) and ***Cannon and Kopell*** (***2015***) for models. They were subsequently observed and studied in networks of interconnected PCs and INs (***Figure 1***A); see ***Adams et al.*** (***2017***) for *in vitro* experiments and ***Sherfey et al.*** (***2018a***) for models. They output periodic volleys of spikes (***Figure 1***A, raster plot) and generate an input-strength dependent population frequency f_{pop} (i.e., the inverse of the period between spike volleys). See Methods for how we compute f_{pop} and ***Table 1*** for a summary of all symbols used throughout the paper. The population frequency in response to a tonic drive will be called the "natural frequency", f_N, of the network; in this case, f_N = 21 Hz. We will demonstrate the importance of the natural frequency for determining responses in an output gate below.

Since we are interested in gating the read-out of oscillatory WM items, we will also review the impact of oscillatory inputs on PC/IN networks (***Figure 1***B) before examining the output gate model. A rhythmically-driven PC population (***Figure 1***Bi) will produce spike volleys with a period matched to the input at low frequencies. At a sufficiently high input frequency, the depolarizing conductances in a fraction of PCs will fail to reach threshold on each cycle of the input; thus, the number of spikes per volley (i.e., the time-averaged firing rate) will decrease (***Figure 1***Bii). For further increasing frequencies, the output frequency of the PC/IN network will remain equal to the input frequency as long as a sufficient fraction of PCs are able to activate INs on every cycle of the input (***Figure 1***Biii). In this case, the output frequency peaks at $f_{max} \approx 28$ Hz; however, ***Sherfey et al.*** (***2018a***) shows that f_{max} increases with the synchrony and strength of the oscillatory input. For higher input frequencies >> f_{max}, the oscillatory input acts like a tonic input, yielding an output with spike volleys occurring at the natural frequency of the network. Therefore, there is only a finite range of input frequencies for which the PC/IN network generates an output faster than the natural frequency (i.e., $f_{pop} > f_N$) (***Figure 1***Biii, shaded region). This intermediate range of fast oscillation frequencies will be shown to impart to oscillatory WM items a competitive advantage in the output gate.

Competition in the output gate

The minimal output gate model consists of two PC populations connected to a common set of inhibitory INs that provide feedback and lateral inhibition (***Figure 1***C). Each PC population is driven by a distinct WM item; spiking in a given output population represents the relaying of information from its corresponding item in the WM buffer. We are interested in how the difference in spiking between the output populations depends on the oscillatory state of the WM items driving them; that is, how the dynamical state of items in the WM buffer affects the relative relay of information from the buffer. In some cases, all items will be relayed; in others, only a subset will be relayed. To assist with exposition, when two WM items have different dynamical states, the item that induces

Table 1. Meaning of symbols used in the study of output gating for working memory.

Symbol	Description
Parameters used to define working memory (WM) items	
$\lambda^T = \lambda^T(t)$	Instantaneous rate of Poisson process for a Target item (kHz)
r^T	Target item strength, $\langle \lambda^T \rangle_t$ (kHz)
f^T	Target item oscillation frequency of rate-modulation (Hz)
δ^T	Target item synchrony control parameter; pulse width of square wave rate-modulation (ms)
$\lambda^D = \lambda^D(t)$	Instantaneous rate of Poisson process for a Distractor item (kHz)
r^D	Distractor item strength, $\langle \lambda^D \rangle_t$ (kHz)
f^D	Distractor item oscillation frequency of rate-modulation (oscillatory distractors only) (Hz)
Output gate	
PC^T	Principal cell population driven by a Target item
PC^D	Principal cell population driven by a Distractor item
IN	Inhibitory interneurons providing strong feedback and lateral inhibition
$\bar{r}^T_{PC}$	Target output strength; time-averaged population firing rate of PC^T (sp/s)
$\bar{r}^D_{PC}$	Distractor output strength; time-averaged population firing rate of PC^D (sp/s)
$\bar{r}_{IN}$	Time-averaged population firing rate of interneurons (sp/s)
f^T_{pop}	Target output oscillation frequency (inverse of period between PC^T spike volleys) (Hz)
f^D_{pop}	Distractor output oscillation frequency (inverse of period between PC^D spike volleys) (Hz)
f_N	Natural oscillation frequency of asynchronously-driven output gate (Hz)
f_{max}	Maximum oscillation frequency of rhythmically-driven output gate (Hz)
f^{pop}_R	f_{pop}-resonant frequency of output gate (Hz) (i.e., item f^T that yields $f^T_{pop} = f_{max}$)

more spike output in the control model (***Figure 1***Ci) will be called the "target" item, and the item that induces less will be called the "distractor" item. To begin, in this section, items in an asynchronous state will be called distractors because equal-strength oscillatory items (i.e., with correlated spiking) always induce at least as many spikes (***Figure 1***Bii, compare solid and dashed lines).

In the output gate model, the response to an asynchronous distractor item was suppressed for a finite range of target frequencies, f^T (***Figure 1***Cii, shaded region). In fact, it was precisely the intermediate range of target input frequencies that yielded an output frequency, f_{pop}, greater than the natural frequency (i.e., when target $f_{pop} > f_N$) (***Figure 1***Ciii, compare solid and dashed lines; ***Figure 1***Cii-iii, compare shaded regions). Maximum distractor suppression occurred at peak f_{pop} (***Figure 1***Cii-iii, vertical dashed lines) (i.e., when target $f_{pop} = f_{max}$) and not at peak target spiking (***Figure 1***Cii-iii, compare blue target curves). This implies that a f_{pop}-resonant item (i.e., an item that maximizes output population frequency) in a WM buffer will maximally suppress distractor responses in the output gate. ***Figure 2*** illustrates how the resonant bias enables different patterns held in working memory in different populations in superficial layers of neocortex to be gated by the resonance of circuits in deep layers.

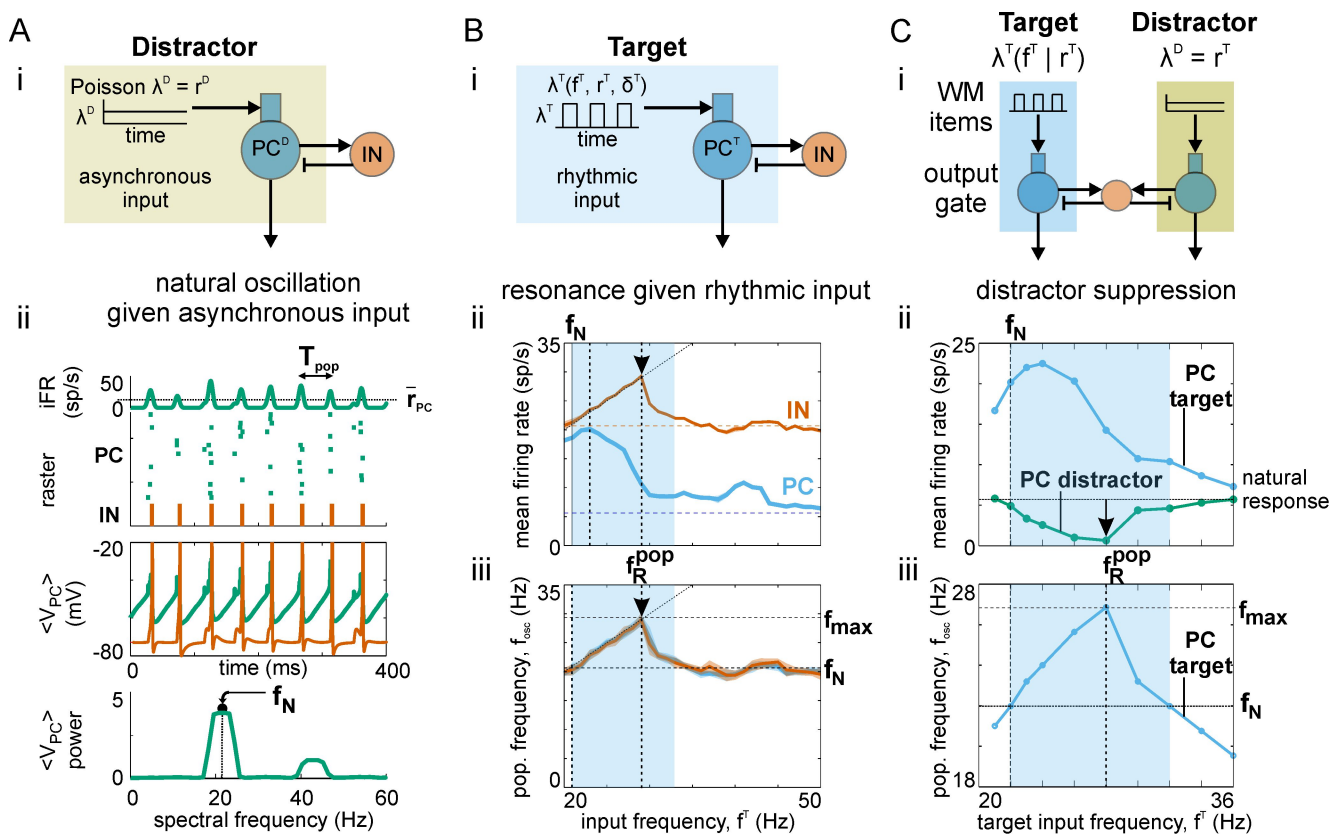


Figure 1. Dynamics of PC/IN network. (**A**) Natural oscillation. (i) Schematic showing asynchronous Poisson spike trains driving the dendrites of two-compartment principal cells (PCs) coupled to interneurons (INs) providing strong feedback inhibition. (ii) The response of the asynchronously-driven PC/IN network is an inhibition-paced, pulsatile oscillation of periodic spike volleys occurring at a natural frequency, f_N. (**B**) Resonant response. (i) Schematic showing rhythmically-modulated Poisson spike trains driving the PC/IN network. The input has square wave rate-modulation with a 10 ms pulse width; i.e., all input spikes occur within 10 ms on a given cycle. (ii) The impact of the frequency of input rate-modulation, f^T, on mean output firing rates (averaged over time, population, and 10 realizations) for PC (blue) and IN (red) populations. (iii) The impact of the input frequency on the frequency of output rate-modulation, f_{pop}, for PC (blue) and IN (red) populations. Vertical dashed lines mark the resonant input frequencies. Horizontal dashed lines indicate the natural response to asynchronous input. The interval with blue shading highlights the range of input frequencies that yield an output oscillation that is faster than the natural frequency. (**C**) Distractor suppression. (i) Schematic showing a rhythmically-driven target population of PCs (PC^T) competing with an asynchronously-driven distractor population (PC^D) through a shared pool of inhibitory INs. Strong feedback inhibition causes each output to be rhythmic; the shared INs induce competitive interactions between them. (ii) Mean firing rate outputs for target (blue) and distractor (black) as target input frequency is increased. Blue shading highlights the range of input frequencies yielding an output oscillation that is faster than the natural frequency; the range increases with input synchrony and strength. Parts of this figure were adapted from ***Sherfey et al.*** (***2018a***).

This can be understood most easily by considering the interaction between excitation and inhibition in the output gate over time (***Figure 3***). Essentially, the excitatory output gate population with the shortest period between spike volleys will be the dominant driver of local inhibition that suppresses spikes in all other populations connected to the same interneurons. The number of spikes per PC volley (i.e., the time-averaged PC firing rate) is less important than the period between volleys (i.e., the population frequency) as long as there are enough spikes in an excitatory volley to engage the inhibition.

Figure 3A shows that when a periodic item in the WM buffer oscillated with frequency $f_N < f^T < f_{max}$, its target output engaged interneurons in the output gate every $1/f^T$ seconds, which was shorter than the $1/f_N$ seconds required for the asynchronously-driven distractor output to reach threshold (***Figure 3***A; compare the periods marked by horizontal double arrows). We know from previous work (***Sherfey et al., 2018a***) that the natural frequency increases with the strength of an asynchronous drive. ***Figure 3***B shows that once an asynchronous item is strong enough for f_N to exceed f^T, then the distractor output becomes the dominant driver of local inhibition in the output gate. Consequently, the response to an oscillatory item can be suppressed by a stronger

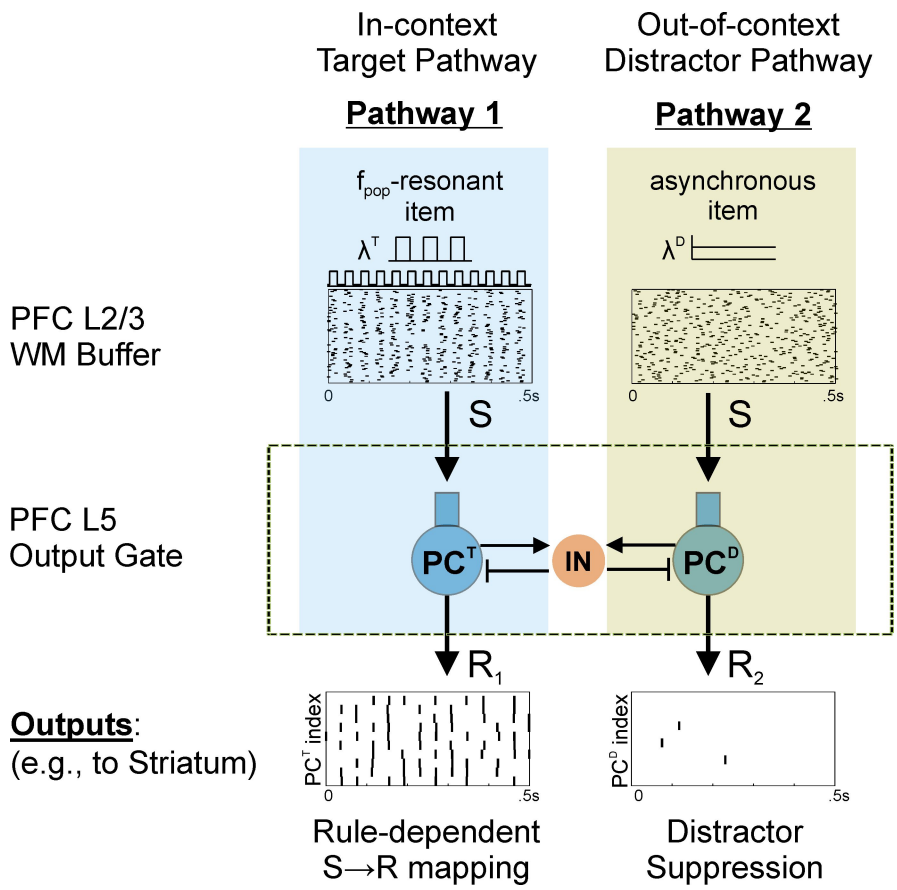


Figure 2. Resonant bias supports rule-based stimulus-response mapping. Example simulation showing two pathways (i.e., alternative stimulus-response mappings) with superficial layer inputs and deep layer outputs. One pathway has a resonant input that drives its output population while the other pathway has an asynchronous input and output that is suppressed by interneuron-mediated lateral inhibition. When both inputs represent different dimensions of the same stimulus (S) and outputs map onto alternative action plans (R1, R2), context-dependent rhythmicity in a select input population can provide adequate bias for rule-based selection of stimulus-response mapping.

asynchronous item when the latter induces a faster population frequency.

Constraints on the suppression of stronger distractors

The previous simulations have demonstrated that either the oscillatory or asynchronous WM item can produce a larger response in the output gate depending on the relative frequencies of oscillations that they induce. Next, we will explore how the output gate response depends on task-modulated properties of WM items when the asynchronous item is stronger.

Experiments have demonstrated that task-relevant items in WM show context-dependent synchrony (***Buschman et al., 2012***) and that the strength of items depends on stimulus familiarity (***Shen et al., 2018***). Suppose a more familiar, context-irrelevant item (i.e., a distractor) shows stronger, asynchronous activity while a less familiar, context-relevant item (i.e., a target) shows weaker, beta-synchronous spiking. We have shown above that rhythmic items in WM that are f_{pop}-resonant with the output gate can be exclusively read-out when the WM items have equal strength (***Figure 2***, ***Figure 3***A) and that the resonant item can be suppressed by stronger asynchronous items (***Figure 3***B). The question we asked next is at what point does a context-irrelevant, asynchronous distractor with increasing strength produce more output spikes than the context-relevant, resonant target?

We addressed this question by delivering spikes from a 28Hz resonant, target item to one population in the output gate and spikes from an asynchronous, distractor item to a competing population in the output gate (***Figure 4***A). The strength of the distractor item, r^D, was increased across simulations from 1.0 up to 2.0 times that

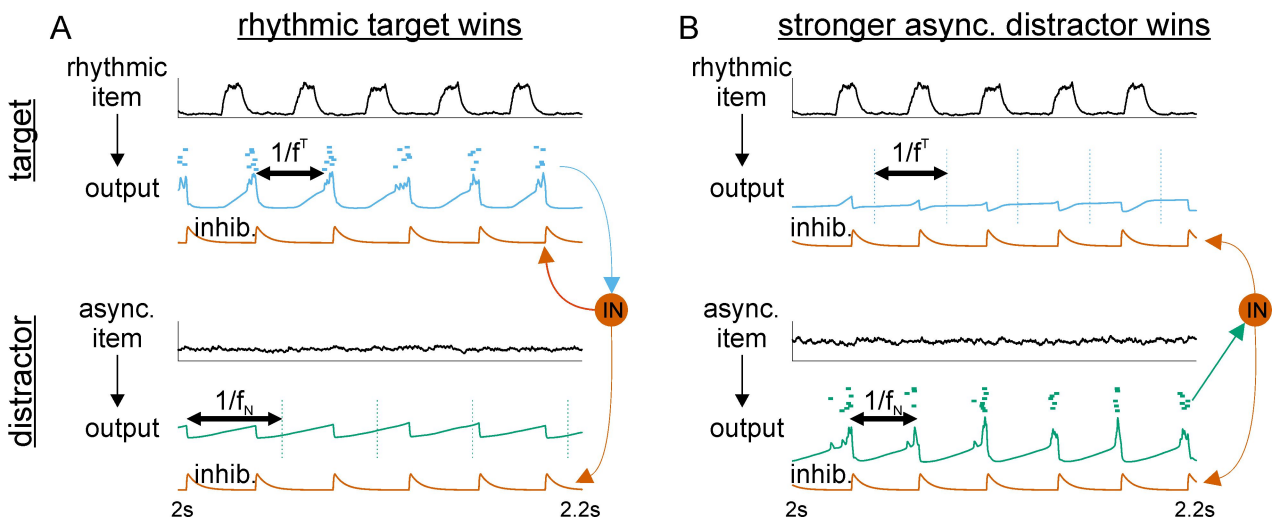


Figure 3. Oscillatory gating: dominance of the fastest output oscillation. (**A**) Rhythmic WM item dominates the output gate. The response to the oscillatory input has a shorter period than the natural response to the asynchronous item (compare horizontal double arrows). Consequently, the target engages local inhibition to suppress the distractor response before it can reach threshold. Vertical dashed lines on distractor output mark when spikes would have occurred in absence of lateral inhibition. (**B**) A stronger asynchronous WM item dominates the output gate. The response to a stronger asynchronous input produces a natural oscillation with a shorter period than that of the oscillatory input. Consequently, the distractor output engages local inhibition and suppresses the response to the oscillatory item. Vertical dashed lines on target output mark when spikes would have occurred in absence of lateral inhibition.

of the target item. We found that the target output produced more spikes than the distractor output until the distractor item was 50% stronger than the target item (***Figure 4***B, orange curve). Similar to above, the reason why the dominant output (i.e., the output producing the most spikes) switched at that point can be understood by considering the relative population frequencies of the two output populations. In the simulations here, the natural frequency of the distractor equals the resonant frequency of the target when the distractor item is 50% stronger. Across the set of simulations, whichever output has the fastest population rhythm produces more output spikes (***Figure 4***B, orange curve). In other words, as WM item strength is varied, it remains the case that whichever population in the output gate has the shortest period between volleys of spikes will dominate the output layer, even if each volley contains fewer spikes than the opposing population would otherwise.

Furthermore, we have shown previously that the f_{pop}-resonant frequency of the output layer depends on the synchrony of its input (***Sherfey et al., 2018a***). That implies that the maximum frequency of the output gate can be increased or decreased by varying spike synchrony of an oscillatory item in WM. In other words, the shortest period between volleys of spikes that can be achieved in a rhythmically-driven output population can be made shorter by increasing the synchrony of spiking in a superficial population that encodes the target item. This is the case because the ability of inhibitory interneurons in the output gate to spike on every cycle of the excitatory input increases with the synchrony of the feedforward spikes driving them (see ***Sherfey et al.*** (***2018a***) for more details). The consequence of this property for output gating is significant and important because synchrony, itself, is task-modulated during cognitive control tasks (***Buschman et al., 2012***).

Resonant WM items with different degrees of synchrony can induce maximal frequencies in the output gate with different distances from the natural frequency. In the high-synchrony case (i.e., all spikes on a given cycle occur within 1 ms), a 31Hz resonant item in WM can produce more spike output than a 70% stronger distractor (***Figure 4***A, blue curve). As target synchrony decreases, the maximal output frequency converges on that for an item with sine wave modulation of the instantaneous firing rate; in those cases, a resonant item can dominate a 30% stronger distractor (***Figure 4***B, black and red curves). Still, in all cases, the item producing output volleys with the shortest period (i.e., the fastest population rhythm) dominates the output gate (***Figure 4***B, all curves).

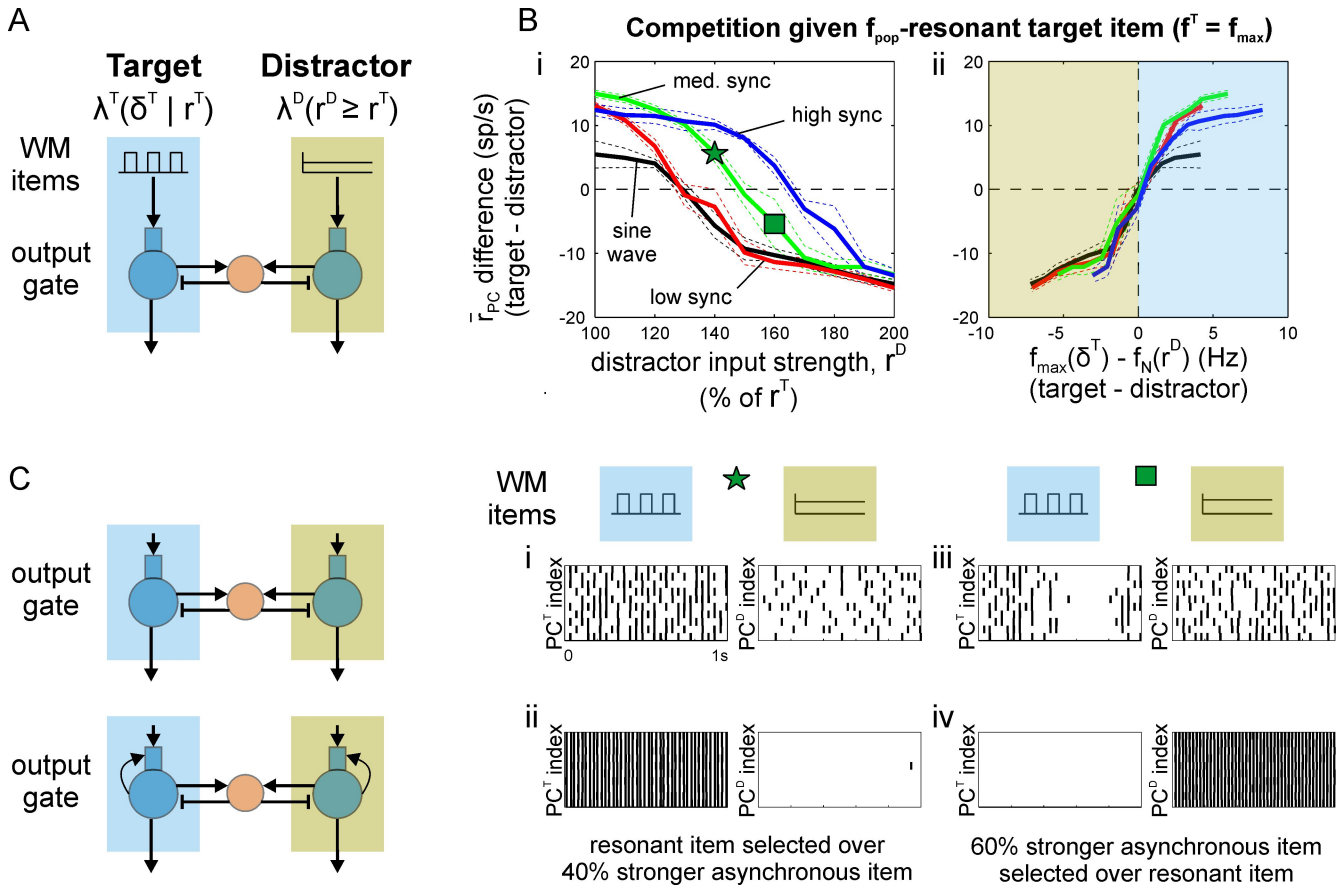


Figure 4. Resonant bias enables suppression of stronger distractor pathways. (**A**) Schematic showing target output (PC^T) receiving an oscillatory input with synchrony varied across simulations in competition with a distractor output (PC^D) receiving asynchronous input with strength varied 1-2x the strength of oscillatory input across simulations. Oscillatory inputs were either sinusoidal or square wave with high synchrony (1 ms pulse width), medium synchrony (10 ms pulse width), or low synchrony (19 ms pulse width). (**B**) Differential output firing rates (target-distractor) for target input frequencies maximizing output population frequency (i.e., for inputs at the f_{pop}-resonant frequency) and distractor inputs with increasing strength (i.e., asynchronous input rate, r^D). Differential output is plotted against (i) the strength of distractor input and (ii) the difference in population frequencies expected in the absence of competition (i.e., in an isolated PC/IN network). The blue shaded region highlights the range of responses where target output frequency exceeds the natural frequency expected for the distractor output; the green shaded region highlights the range where distractor output oscillates faster. The star and square in Bi mark distractor strengths that produce greater target and distractor outputs, respectively, and are used to investigate the effects of recurrent excitation. In all cases, the output with higher spike rate was the output with higher population frequency (i.e., a shorter period between spike volleys). (**C**) Recurrent excitation amplifies output differences for winner-take-all selection. (i) Without recurrent excitation, target output is greater despite the distractor receiving a 40% stronger input. This simulation corresponds to the point marked with a star in Bi. (ii) Recurrent excitation amplifies resonant bias producing winner-take-all dynamics that select the output driven by a weaker resonant input. (iii) Without recurrent excitation, distractor output is greater when it receives an asynchronous input that is 60% stronger than an opposing resonant input. This simulation corresponds to the point marked with a square in Bi. (iv) The response to an oscillatory target is suppressed when the stronger asynchronous input elicits a faster natural frequency.

Learning amplifies bias for winner-take-all output gating

We have seen that changing target f_{pop} (by varying f^T and synchrony of the target item) or distractor $f_{pop} = f_N$ (by varying r^D in the distractor item) creates a bias in the output gate that favors one item or the other based on the relative population frequencies induced in the respective outputs (***Figure 1***C, ***Figure 4***B). In the case of large differences between the output population frequencies (e.g., induced by a resonant or a strong asynchronous item), we have seen that this bias can produce a gated response (***Figure 2***, ***Figure 3***). Next, we investigated whether we could achieve winner-take-all (WTA) dynamics by incorporating a common motif in WTA networks.

Specifically, WTA dynamics are commonly observed in networks with both strong lateral inhibition and strong recurrent excitation (***Kaski and Kohonen, 1994***). Our control model of the output gate already contains strong lateral/feedback inhibition. When we added recurrent excitation among all cells within each output population,

the bias observed previously was, indeed, converted into a WTA response (***Figure 4***C). In the control network, the 28Hz resonant item produced more spiking in the output gate than a 40% stronger asynchronous item. With recurrent excitation, the resonant item was selected in the output gate while the asynchronous item was completely blocked. Conversely, when the asynchronous item was made 60% stronger, so that its natural frequency exceeded 28Hz, it was selected in the output gate while the resonant item was completely blocked. This suggests that learning input/output mappings (e.g., across trials of a task) can convert a rhythm-mediated bias into an exclusive, WTA output gate that selects the WM item that induces the fastest output oscillation (i.e., the output with the shortest period between spike volleys).

Multiple rhythmic items in competition

All gating simulations so far have examined the competitive interaction of responses driven by two items, one in an oscillatory state and one in an asynchronous state. However, experiments have shown that multiple rhythms often exist in the same region of cortex (***Lundqvist et al., 2016***; ***Adams et al., 2017***). During WM tasks, both gamma and beta-rhythmic activities have been observed in the same superficial layers of LPFC (***Bastos et al., 2018***). We next investigated the behavior of the output gate in the presence of multiple rhythmic items in WM.

To begin, we investigated whether relative population frequencies determine which output produces more spikes like they did in the case of rhythmic vs. asynchronous items. In the PC/IN network with a single PC population (***Figure 1***Bi), PC spiking peaked for $f^T \approx 24$ Hz (***Figure 1***Bii, blue curve) while IN spiking and f_{pop} peaked for $f^T \approx 28$ Hz (***Figure 1***Bii-iii) (for medium-synchrony inputs). Given two separate PC/IN networks (i.e., without competition) (***Figure 5***Ai) the relative spike output is determined only by the $\bar{r}^T_{PC}$ tuning curve in ***Figure 1***Bii. To show this as a reference case, we varied item frequencies over a grid from 16-36 Hz, arbitrarily labeled one the "target" and the other the "distractor", and then plotted whether target output exceeded distractor output over the (f^T, f^D) grid (***Figure 5***Aii). This plot shows that the dominant population flips as the input frequencies pass the peak FR at 24 Hz (***Figure 5***Aii, circle). The black zone to the lower-left of the circle in ***Figure 5***Aii indicates that while $f^D < f^T < 24$ Hz, the target produces more output; however, as f^T increases beyond 24 Hz and f^D remains closer to the 24 Hz input that maximizes spiking, the distractor produces more output. Similarly, the black region to the lower-right indicates that the (distractor) population driven by a higher frequency, f^D, produces fewer spikes when the target population is driven by a frequency, f^T, closer to 24 Hz. Essentially, as would be expected, whichever population in the output gate (without competition) receives an input closer to the frequency (24 Hz) that maximizes spiking in ***Figure 1***Bii will produce more spikes here; the only exceptions result from harmonics. The blue lines mark the higher frequency (28 Hz) at which f_{pop} peaks in the output (***Figure 1***Biii). If relative population frequencies determined which population produced more spike output, we would expect the flip to occur at 28 Hz (i.e., the maximal output frequency) and the target to dominate the region marked with an "x" as well (i.e., 24 Hz < f^D < f^T < 28 Hz).

Indeed, when we have the two outputs compete in the output gate (***Figure 5***Bi), the target most often produces more output as long as f^D < f^T < 28 Hz (***Figure 5***Bii). In this range, the population labeled "target" has the highest population frequency. Essentially, and in contrast to the scenario without competition, whichever population in the output gate (with competition) receives an oscillatory input closer to the maximum population frequency (28 Hz) in ***Figure 1***Biii will produce more spikes here. Furthermore, plotted a different way, we can see that the output with the shortest period between spike volleys usually produces more spikes (***Figure 5***C, lower-left and upper-right quadrants). In contrast to the analogous case for rhythmic vs. asynchronous items (***Figure 4***Bii), however, there are occasions when the lower-frequency population produces more output spikes (***Figure 5***C, upper-left and lower-right quadrants); these exceptions will be addressed in the Discussion.

Tuning the output gate

A useful gate is one that can be dynamically tuned to select different types of inputs. We have previously shown that the resonant frequency of PC/IN network can be adjusted by changing the excitability of the PC population via neuromodulation of intrinsic ion channels (***Sherfey et al., 2018a***). Here, we tested whether a similar adjustment of the network's excitability could be used to flexibly tune an output gate for selecting different frequencies of rate-modulation.

As we have seen above, in the control model the output gate has a maximum output frequency of 28 Hz

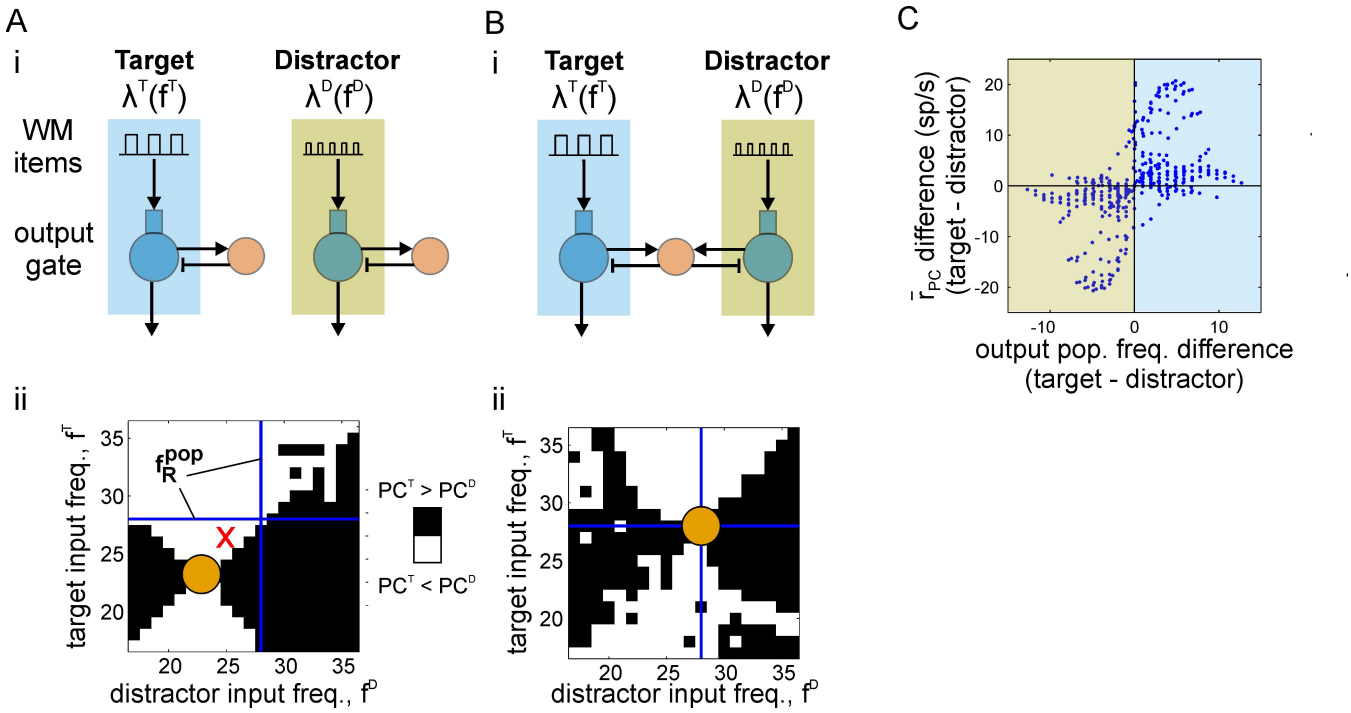


Figure 5. Competition between oscillatory items: the most f_{pop}-resonant item wins. (**A**) Without lateral inhibition, relative spike outputs are determined by $\bar{r}^T_{PC}$ tuning curves in ***Figure 1***Bii. (i) Circuit diagram showing two independent PC/IN networks without lateral inhibition. Both PC populations received oscillatory WM inputs with different modulation frequencies. The modulation frequency of items delivered to each PC population was varied from 18-36 Hz across simulations. Each pathway was arbitrarily labeled "Target" or "Distractor". (ii) A binary image indicating whether the Target or Distractor PC population outputs more spikes across the simulation given oscillatory inputs at different frequencies. A black pixel indicates the Target population produced more spikes, whereas a white pixel indicates the Distractor population output more spikes. The circle marks the intersection of inputs at the $\bar{r}^T_{PC}$-resonant frequency. The blue lines mark the f_{pop}-resonant frequency. Whichever output is closer to peak $\bar{r}^T_{PC}$ produces more spikes when the circuits are disconnected. (**B**) With lateral inhibition, relative spike outputs are determined by f_{pop} tuning curves in ***Figure 1***Biii. (i) Circuit diagram showing two PC populations competing through lateral inhibition. The inputs were the same as in (A). (ii) Same plot as (Aii). Whichever output is closer to peak f_{pop} produces more spikes when the output populations are connected through shared inhibitory interneurons. (**C**) Relative spike outputs with competition plotted against the relative population frequencies without competition. In most cases, the population with higher population frequency produces more spikes.

in response to a medium-synchrony rhythmic item. This causes a WM item that is modulated at a higher 40 Hz (gamma) frequency to induce a rhythmic response near the natural frequency, $f_N \approx 21$ Hz (***Figure 1***Biii). In contrast, an output can match a 25 Hz (beta) rhythmic item because that is below its maximum frequency. Thus, in the control case, the activity a beta-rhythmic item induces in the output gate has a shorter period between spike volleys and recruits lateral inhibition to block responses to a gamma-rhythmic item (***Figure 6***A). How can we selectively output the less-resonant gamma-rhythmic item in WM?

Building on our earlier work (***Sherfey et al., 2018a***), we hypothesized that increasing the excitability of the output gate would increase the gate's resonant frequency and lead to the selection of WM items with a higher modulation frequency. To test this, we delivered a nonspecific, subthreshold, asynchronous drive to all PC populations of the output gate, effectively increasing their resting membrane potential and, thus, their excitability. With sufficient strength, such a signal enabled the output gate to follow the gamma-rhythmic item. The response to the gamma-rhythmic item then exhibited a shorter period between volleys of inhibition-recruiting spikes and, consequently, suppressed the response to the beta-rhythmic item (***Figure 6***B). Thus, the gamma-rhythmic item was exclusively selected by making the output gate resonate to gamma frequencies. This represents a flexible mechanism by which a nonspecific modulatory signal can be adjusted to tune the maximum frequency of the output gate so that it selectively responds to WM items modulated at higher or lower frequencies.

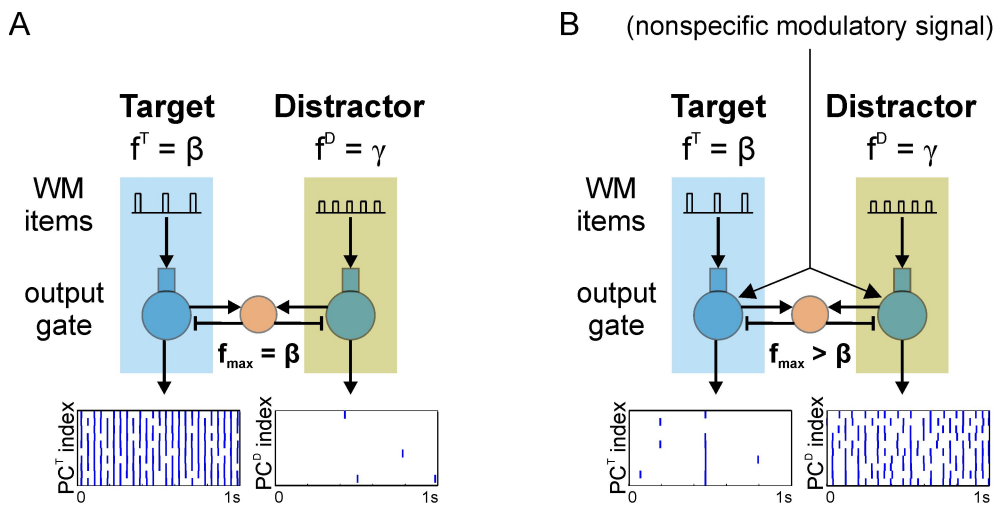


Figure 6. Nonspecific inputs can tune output resonance for switching between specific beta- and gamma-rhythmic pathways. (**A**) A resonant beta input suppresses the response to a less resonant gamma-frequency input. (**B**) A nonspecific asynchronous input to both output populations shifts their resonant frequency to the gamma-range, causing the output layer to select the gamma-rhythmic input and suppress response to the less resonant beta input.

Oscillatory gating for rate-coded items

Parallel pathways

The WM items we have simulated so far had time-averaged firing rates that were uniform across the encoding population. However, experiments have shown that PFC encodes items in the pattern of firing rates across PCs of an encoding population (***Mante et al., 2013***). For our oscillatory gating mechanism to be used to gate outputs from WM, it must be able to gate such rate-coded items. Next, we tested the ability of our oscillatory gating mechanism to select rate-coded items with partial Gaussian spatial profiles (***Figure 7***).

First, we simulated two items with resonant beta-frequency modulation of the instantaneous firing rate and different Gaussian spatial patterns of time-averaged firing rates. When delivered to two PC populations in the output gate, the spatial pattern was largely conserved in the corresponding PC populations (***Figure 7***Ai). This demonstrates that parallel, rate-coded items that are resonant with the output gate can be propagated simultaneously. In contrast, when the fast modulation of the instantaneous firing rate occurs at a non-resonant gamma-frequency, only the resonant beta-frequency item is conserved in its output population while the non-resonant item is blocked (***Figure 7***Aii). This represents a form of frequency-based output gating that selects for the most resonant rate-coded item in WM.

Convergent pathways

Anatomically, projections from superficial to deep layers of cortex involve some degree of overlap (***Levitt et al., 1993***; ***Kritzer and Goldman-Rakic, 1995***) and this, most likely, results in multiple WM items projecting to some of the same cells in the deep layer. We have shown throughout this paper that lateral inhibition driven by a resonant output gives rise to periodic inhibition that blocks responses in opposing pathways. Next we tested the hypothesis that feedback inhibition can similarly block responses to less resonant rate-coded signals when pathways converge on a single output population. Similar filtering has been demonstrated in simpler models with uniform input rates (***Cannon et al., 2014***).

We repeated the previous simulations demonstrating frequency-based gating of rate-coded signals, except this time we delivered both items to the same PC population in the output gate (***Figure 7***B). In this case, two rate-coded items with Gaussian profiles and resonant beta-frequency modulation were approximately conserved in the spatial pattern of output firing rates in the single PC population (***Figure 7***Bi). In contrast, when one item was modulated at a less resonant gamma frequency, the item with resonant modulation was conserved in the output while the response to the less resonant item was blocked by the feedback inhibition as hypothesized (***Figure 7***Bii).

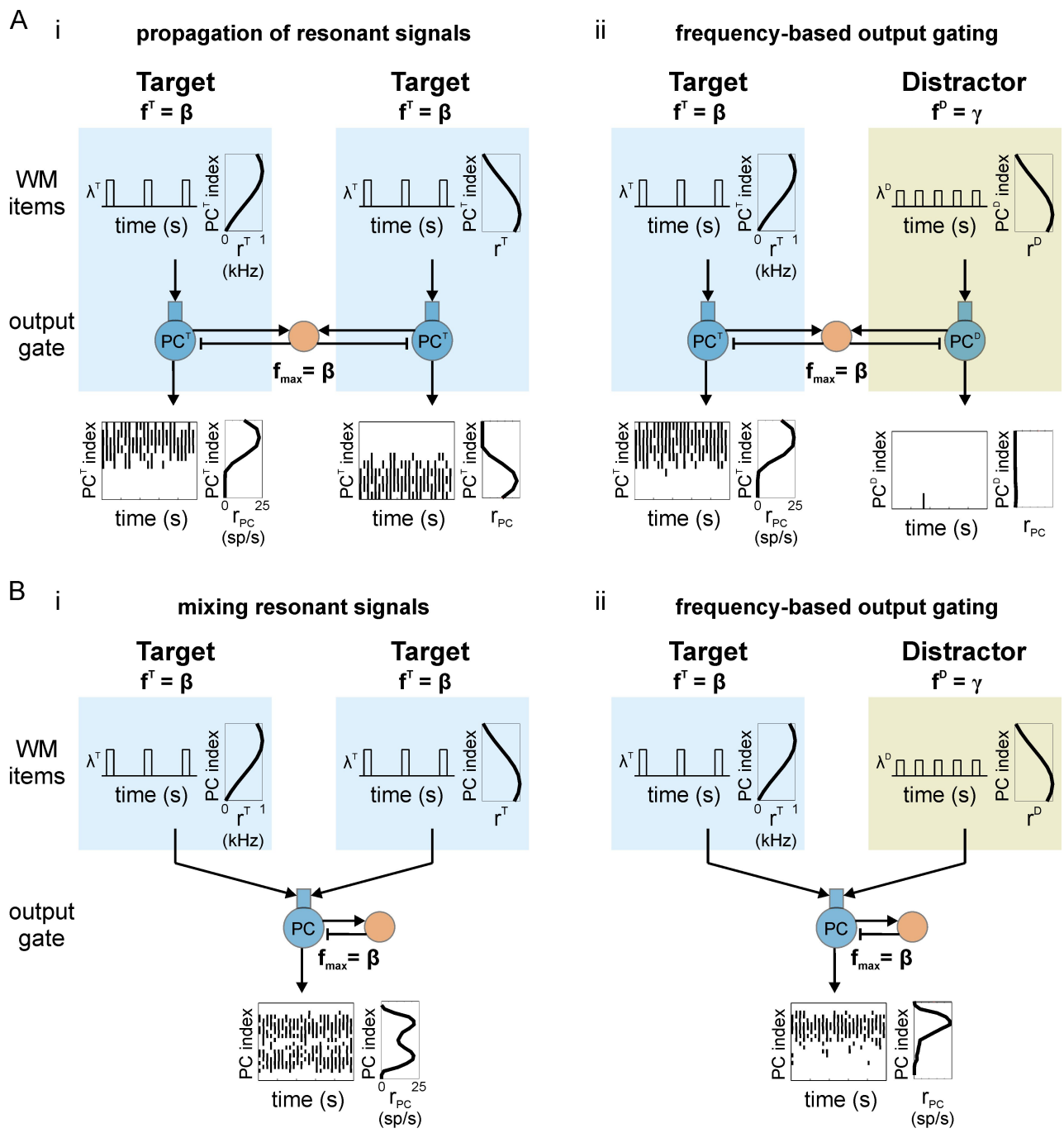


Figure 7. Oscillatory gating for rate-coded signals. (**A**) Resonant bias supports frequency-based gating of rate-coded signals among parallel pathways. (i) Outputs of two pathways reflect the spatial pattern of firing rates across their inputs when both inputs are embedded in resonant oscillations. PC index represents linear indices of cells in a given output PC population. The spatial pattern of time-averaged firing rates across cells of the population is assumed to encode item information. Individual cells have low firing rates while the collective population is modulated at a faster frequency. Both WM items have the same modulation frequency. (ii) Frequency-based output gating: More resonant rate-coded signals suppress less-resonant rate-coded signals. WM items have the same spatial pattern of time-averaged firing rates as in (Ai), but now the frequency of faster time scale modulation differs between the two items; only the Target item has a modulation frequency that is resonant with the output gate. (**B**) Resonant bias supports frequency-based gating of rate-coded signals among convergent pathways. (i) Similar to (A) except both inputs converge on a single output population that reflects the approximate sum of the input signals in its spatial pattern of firing rates. (ii) A less resonant gamma-frequency signal is blocked from the output population.

Discussion

We have presented a novel mechanism for gating outputs from a working memory (WM) buffer. The mechanism requires strong lateral and feedback inhibition in the output gate to induce periodic responses. Essentially, the output population with the shortest period between spike volleys will most reliably engage inhibitory interneurons to suppress responses in competing outputs. Resonant properties that result from the feedback inhibition support frequency-based output selection of items in the WM buffer. This enables the modulation frequency of WM items to govern output selection, possibly in opposition to the strengths (i.e., time-averaged firing rates) of items in WM. We showed that the resonant frequency can be tuned by flexible modulatory signals to select different frequencies. We also showed that with the addition of recurrent excitation (e.g., with learning), the gate can exhibit exclusive, winner-take-all responses. Finally, the same oscillatory gating mechanism could be used to gate flows, in general, for arbitrary, rate-coded signals along parallel and convergent pathways in any cortical region.

Summary of oscillatory gating rules

The results from this study can be organized into a set of rules that depend on dynamical states and determine which WM item(s) will dominate the output gate.

Asynchronous vs. asynchronous items

The strongest item in WM (i.e., the item with the highest time-averaged spike rate) will produce the strongest response in the output gate. This occurs in an output gate with strong feedback inhibition because the strongest asynchronous item will induce the highest natural frequency (i.e., the shortest period between spike volleys) (see ***Figure 1***A for the natural response to asynchronous items and ***Figure 3***, lower plots, for dependence of natural frequency on item strength).

Beta/gamma rhythmic item vs. asynchronous items

A (possibly weaker) rhythmic WM item will produce the strongest response in the output gate if its modulation frequency is within some bounds that yield a population response with the shortest period between spike volleys (***Figure 2***, ***Figure 3***A). That occurs when the modulation frequency of the WM item is between the strength-dependent natural frequency induced by the asynchronous items and the maximal frequency of the output gate as determined by the strength and synchrony of oscillatory items (***Figure 4***B, Ci-ii). Otherwise, the asynchronous item can produce an equally-strong or stronger response depending on how much the natural frequency that it induces exceeds the frequency induced by the rhythmic item (***Figure 3***B, ***Figure 4***Ciii-iv).

Rhythmic vs. rhythmic items

The item with modulation frequency that is most resonant (***Figure 5***) with the tunable output gate (***Figure 6***) will produce the strongest response.

Control mechanisms

According to the gating rules, responses in the output gate depend on the resonant frequency of the gate (i.e., its maximum frequency) as well as the synchrony and modulation frequency of rhythmic items in the WM buffer. Each of these properties can be dynamically regulated to guide the flow of information from working memory without requiring changes in the relative strengths of WM items (i.e., changes in the time-averaged firing rates).

Controlling the resonant frequency of the output gate

We have shown that the resonant frequency can be flexibly tuned by changing the level of background excitation present in the output gate using an asynchronous modulatory signal (***Figure 6***). A nonspecific, time-varying asynchronous signal to the deep layers of PFC could tune the resonant frequency of the WM output gate on the time scale of milliseconds needed to change the resting membrane potential. Based on the anatomy, such a signal could come from thalamus or the deep layers of another region of PFC (***Barbas, 2013***). In previous work (***Sherfey et al., 2018a***), we showed that the resonant frequency can be tuned on a slower time scale of seconds to minutes by neuromodulation of intrinsic ion channels. This could occur, for instance, via modulation of potassium conductances in PFC by dopamine (***Dong and White, 2003***) and acetylcholine (***Hasselmo, 2006***; ***Bloem et al., 2014***).

Controlling the rhythmicity and synchrony of WM items

The presence of inhibition-based oscillation and its synchrony in PC/IN networks are functions of the strength of feedback inhibition (***Sherfey et al., 2018a***). By having two competing interneuron populations that can alternately pace rhythms in the same PC population, differential drives to the interneuron populations could switch the same PC population between different degrees of synchrony. For instance, this could be achieved in cortical layers containing both parvalbumin-positive (PV+) and calbindin-positive (CB+) interneurons (e.g., L2/3 in DLPFC) which are known to be driven by distinct inputs (***Medalla and Barbas, 2009***). This mechanism supports rapid switching between dynamical states (e.g., oscillatory vs. asynchronous) and could be used to gate population signals. Precise control over which cells participate in the resonant oscillation may be achievable using this mechanism with PV+ and somatostatin-positive interneurons (***Khan et al., 2018***) or, more likely, experience-dependent plasticity.

Controlling the modulation frequency of WM items

The frequency of inhibition-based oscillation in PC/IN networks is a function of the time constant of inhibition (***Sherfey et al., 2018a***). Similar to the control of item synchrony, differential drives to distinct interneuron types with different time constants of inhibition in L2/3 could rapidly switch the same PC population between different modulation frequencies. For instance, PV+ fast spiking (FS) interneurons have been found with shorter time constants than CB+ interneurons (***Medalla and Barbas, 2009***), and co-expressing CB+/calretinin-positive interneurons have been found to inhibit PV+ FS cells (***DeFelipe, 1997***). Thus, we hypothesize that an interaction between these two classes in the superficial layers of PFC could switch WM items between beta and gamma frequencies. Furthermore, correspondence between item modulation and output gate resonant frequencies might be facilitated by oscillation-pacing interneuron types (***Sohal et al., 2009***) that are present in both input and output layers and that provide inhibition with the same time constants (e.g., PV+ fast spiking INs found in layers 3 and 5 of DLPFC (***DeFelipe, 1997***; ***Povysheva et al., 2008***)).

Oscillatory gating for cognitive function

Working memory

The idea that beta- and gamma-frequency oscillations in PFC govern output gating from WM is supported by an abundance of anatomical and physiological data. Task-relevant WM items, presumably engaged in ongoing processing, often exhibit synchrony at beta and gamma frequencies (***Siegel et al., 2009***; ***Lundqvist et al., 2016***, ***2018***; ***Bastos et al., 2018***). For instance, beta-rhythmic phase coding has been shown to segregate multiple items in a visual WM task (***Siegel et al., 2009***) while beta and gamma power increases have been observed in proportion to the number of items stored in WM (***Roux et al., 2012***) and with cognitive demand (***Basar-Eroglu et al., 2007***). By maintaining distinct item representations in a superficial buffer that targets PC populations in a deep output layer, resonant biases, like those reported here, utilizing observed beta- and gamma-frequency oscillations could gate outputs from WM. Our findings suggest that such resonant oscillations could support the selective output of WM items even when other less-resonant items are stored with stronger activation (e.g., items stored in asynchronous rate-codes with 70% higher mean firing rates) (***Figure 4***). The selected items would then be available in an oscillatory state for read-out in subcortical structures and participation in downstream processing.

Consistent with this hypothesized mechanism for dynamic routing of WM representations, beta-synchrony has been observed between PFC and higher-order thalamus during a WM task, and it was correlated with performance (***Parnaudeau et al., 2013***). Our results suggest that beta-rhythmic synchrony may serve thalamo-cortical processing by effectively selecting the relevant pathways and suppressing distractors. Suppression of responses to asynchronous inputs has the additional benefit of suppressing uncorrelated background activity, which increases the signal-to-noise ratio in the output layer. The loss of this noise suppression mechanism when beta-synchrony was disrupted could underlie the increase in errors observed by ***Parnaudeau et al.*** (***2013***). Without rhythmicity, outputs are rate-based with asynchronous background noise potentially contributing as much to output selection as asynchronous rate-coded signals.

Many researchers have proposed that WM items are maintained in the persistent spiking of item-encoding PCs (***Fuster, 1973***; ***Goldman-Rakic, 1995***; ***Wang, 1999***; ***Curtis and D'Esposito, 2003***). However, ***Lundqvist et al.*** (***2018***) have shown that item representations are more often expressed transiently in gamma bursts that last for only a

few cycles; such bursts are thought to reflect the transient reactivation of WM items stored in synapses (***Lundqvist et al., 2011***; ***Rose et al., 2016***). Since oscillatory gating in the present work arises within one cycle, our oscillatory gating mechanism could be used to selectively propagate any active WM representation, whether persistent or transient, as long as its modulation frequency is most resonant with the output gate. The resonant frequency of the output gate could be matched to the oscillation frequency of the target WM item (whether at a beta or gamma frequency) by tuning the strength of the WM item, itself, or using a time-locked modulatory "GO" signals (as in ***Figure 6***) that arrives when the target item should be read-out. At the same time, the asynchronous and less resonant population signals in the superficial WM buffer would be unaffected and held online for potential future read-out from WM.

Rule-based action and cognitive control

Dynamic regulation of the WM buffer and output gate, via the control mechanisms described above, allows us to cast oscillatory gating in the broader context of cognitive control. For instance, context-dependent changes in the oscillatory state of WM items or the resonant frequency of the output gate could mediate rule-based selection of input-output mappings.

For instance, experiments have revealed that populations coding task-specific rules exhibit increased beta-rhythmic synchrony (***Buschman et al., 2012***). In our model, the deep layer output gate exhibits resonance at a similar frequency and consequently supports the selective read-out of WM items that are synchronous with beta-frequency oscillations. Rule-dependent synchrony of WM items, then, could control which input-output mappings are engaged while suppressing responses of opposing outputs via lateral inhibition in the output gate (see ***Figure 2***). Rule updating via oscillatory state control mechanisms operating on the WM buffer in superficial PFC could potentially be directed by contextual inputs from hippocampus (***Komorowski et al., 2013***), error signals from ACC (***Kerns et al., 2004***; ***Amiez et al., 2005***), or basal ganglia gating mechanisms (***Frank et al., 2001***; ***Badre and Frank, 2012***). In particular, ACC is known to target superficial CB+ INs in PFC (***Medalla and Barbas, 2009***) that may be able to flexibly control the oscillatory state of WM items as described above.

Furthermore, a PFC L5 output gate could govern flows along parallel pathways from visual and auditory cortices (***Barbas, 2015***) through subregions of basal ganglia and thalamus (***O'Reilly and Frank, 2006***), effectively sculpting functional connectivity between disparate regions. For instance, the oscillatory gating mechanism could dynamically link conditions encoded in sensory cortices to subcortically-mediated responses that guide attention, action, and cognitive control, more broadly. Thus, the mechanism may contribute to dynamics engaged by many cognitive processes.

Generality of the oscillatory gating mechanism

Oscillatory gating is likely present throughout cortex since the necessary ingredients are ubiquitous. The primary requirement of the oscillatory gating mechanism is strong, fast inhibition in an output layer; this could occur in the presence of PV+ INs with FS physiology that are known to concentrate in the deep layers of all neocortical regions (***DeFelipe, 1997***). The output gating mechanism presented here does not require that the input items are maintained in a working memory buffer. It can gate stimulus-driven flows equally well. In that context, it can be seen as a general-purpose mechanism for shaping functional connectivity between upstream source regions and downstream receivers.

In the present model, the mechanism holds only in the beta/gamma frequency range because below beta the oscillation period is too long compared to the duration of inhibition. The mechanism may work at lower frequencies in the presence of INs with longer-lasting inhibition. This mechanism works for sine wave rhythms as well; however it is most effective when the rhythmic spikes are synchronized in periodic volleys (see ***Figure 4***Bi).

Comparison to other work

Theoretical work has shown that input gating mechanisms, controlling whether WM items are maintained or updated, can effectively manage a tradeoff between stability and flexibility of the WM (***Hochreiter and Schmidhuber, 1997***; ***Miyake and Shah, 1999***; ***O'Reilly and Frank, 2006***; ***Badre and Frank, 2012***) and that output gating mechanisms provide additional control over when maintained WM items effectively influence downstream targets (***Hochreiter and Schmidhuber, 1997***; ***Kriete and Noelle, 2011***; ***Kriete et al., 2013***). Experiments suggest

input and output gating of WM items are present in PFC (***Badre and Frank, 2012***), while models of input (***O'Reilly and Frank, 2006***) and output (***Kriete and Noelle, 2011***; ***Hasselmo and Stern, 2018***; ***Zhu et al., 2018***) gating of WM items in PFC have proposed activity-based mechanisms mediated by interactions between PFC and basal ganglia (BG). We have presented an alternative gating mechanism that is mediated by network oscillations within cortex, instead of subcortical activity levels. In contrast to activity-based mechanisms, oscillatory gating enables the flow of information to be governed by alterations in spike timing instead of activity levels (i.e., spike rates).

Frank and Badre (***2012***) have shown how the basal ganglia can gate outputs from a PFC WM buffer using an activity-based, disinhibitory mechanism. In their model, BG regulates a signal that increases or decreases thalamic activity to regulate whether PFC outputs can be sufficiently activated for read-out. It is possible that their activity-based mechanism operates in parallel with our oscillation-based mechanism or that the two work in conjunction. For instance, BG-regulated thalamocortical activity could provide an asynchronous modulatory signal that tunes the resonant frequency of an oscillatory output gate in PFC; given rule-specific BG activity, this conjunctive approach could represent a novel mechanism for selecting rule-based input-output mappings using a combination of activity- and oscillation-based gating.

Both experiments and models have demonstrated that WM items in PFC may be represented in populations with spiking that is asynchronous (***Wang, 1999***; ***Tegnér et al., 2002***; ***Renart et al., 2010***) or synchronous with rhythmic modulation (***Compte et al., 2000***; ***Siegel et al., 2009***; ***Lundqvist et al., 2010***, ***2018***). Such asynchronous and oscillatory population activities are indicative of local networks in different dynamical states (***Tegnér et al., 2002***; ***Akam and Kullmann, 2010***; ***Lundqvist et al., 2018***). In the general context of routing oscillatory signals, ***Akam and Kullmann*** (***2010***) showed how a band-pass filter network based on resonant, feedforward inhibition supports the selective read-out of activity with gamma-frequency modulation. Their model was tuned to a balanced regime in which PCs were unresponsive to asynchronous inputs. However, it seems most likely that rhythmicity is not always necessary for WM items to activate populations driven by them. We focused instead on a network model with strong, feedback inhibition that exhibited resonance while remaining responsive to asynchronous signals. That way we could consider under which circumstances WM items in either an asynchronous or oscillatory state may be read-out for downstream processing.

Future work

We made two observations that need further investigation. First, when two oscillatory items induce output gate responses at different frequencies, the lower-frequency response can produce more spikes in exceptional cases (see ***Figure 5***C, lower-right and upper-left quadrants). This occurs when the slower output frequency is closer to the firing rate resonant frequency and the oscillatory input regularly arrives late enough during the inhibitory phase to elicit spikes; however, that is not the case when inputs are weak or low synchrony. More research is needed to examine these exceptional interactions between multiple oscillatory items in the strong item, medium/high synchrony regime. Second, information from low-activity item-encoding cells may be suppressed in the output gate due to the feedback inhibition. The tradeoff between the strength of feedback inhibition and the amount of information lost needs further investigation.

Conclusions

In this work, we introduced a new oscillatory gating mechanism that enables beta- and gamma-frequency oscillations to govern read-out from working memory. The output gate can by rapidly tuned to support routing for flexible cognitive processes, and it may be capable of leveraging fast network oscillations and local inhibition to shape functional connectivity throughout neocortex, more generally.

Methods

Cortical network models of the output gate

The network model represents a cortical output layer with 20 excitatory principal cells (PCs) connected reciprocally to 5 inhibitory interneurons (INs). Hodgkin-Huxley type PC and IN models were taken from a computational representation of a deep layer PFC network consisting of two-compartment PCs (soma and dendrite) with ion channels producing I_{NaF}, I_{KDR}, I_{NaP}, I_{Ks}, I_{Ca}, and I_{KCa} currents (μA/cm^2) and fast spiking INs with channels

producing I_{NaF} and I_{KDR} currents (**Durstewitz and Seamans, 2002**) (**Figure 8**A; see figure caption for channel definitions). IN cells had spike-generating I_{NaF} and I_{KDR} currents with more hyperpolarized kinetics and faster sodium inactivation than PC cells, resulting in a more excitable interneuron with fast spiking behavior. In the control case, PC and IN cell models were identical to those in the original published work while network connectivity was adjusted to produce natural oscillations (not in **Durstewitz and Seamans** (**2002**)), as described below, and the number of cells in the network was decreased to enable exploration of larger regions of parameter space while remaining large enough to capture the dynamics of interest for this study; however, the same resonant frequencies were obtained in simulations using the original network size (shown in **Sherfey et al.** (**2018a**)). All cells were modeled using a conductance-based framework with passive and active electrical properties of the soma and dendrite constrained by experimental considerations (**Durstewitz et al., 2000**). Membrane potential V (mV) was governed by:

$$C_m \frac{dV}{dt} = -I_{inp}(t, V) - \sum I_{int} - \sum I_{syn} \tag{1}$$

where t is time (ms), $C_m = 1$ μF/cm^2 is the membrane capacitance, I_{int} denotes the intrinsic membrane currents (μA/cm^2) listed above, $I_{inp}(t, V)$ is an excitatory current (μA/cm^2) reflecting inputs from external sources described below, and I_{syn} denotes synaptic currents (μA/cm^2) driven by PC and IN cells in the network. We chose to explore the prefrontal model as part of a larger project on prefrontal oscillations. We confirmed the generality of our qualitative results using a leaky integrate-and-fire model; see **Appendix 1** for details.

The output layer had either one or two populations of PC cells with each output population receiving inputs from one or two working memory (WM) items (defined below). Input frequency-dependent response profiles were characterized using a network with one input item and one output population (**Figure 8**A). Competition between outputs was investigated using a network with two homogeneous PC populations driven by one input item each while interacting through a shared population of inhibitory cells (**Figure 8**B).

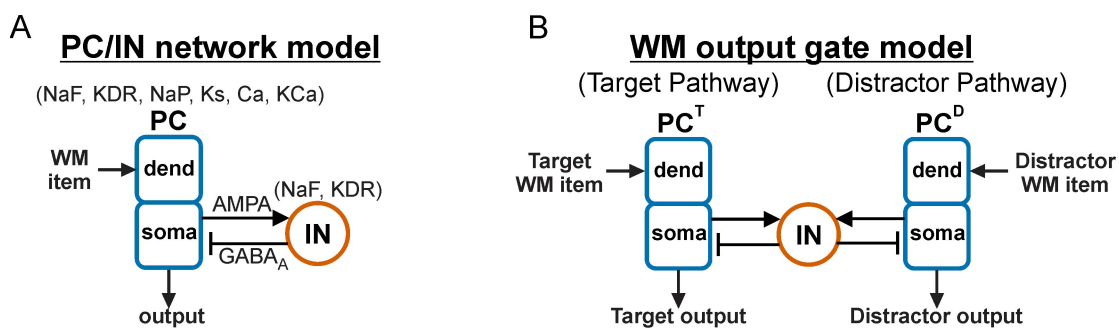


Figure 8. Architecture of output networks. (**A**) Diagram showing feedforward excitation from external independent Poisson spike trains to the dendrites of 20 two-compartment (soma, dend) principal cells (PCs) receiving feedback inhibition from a population of 5 fast spiking interneurons (INs). All PC and IN cells have biophysics based on rat prelimbic cortex (Ion channel key: NaF = fast sodium channel; KDR = fast delayed rectifier potassium channel; NaP = persistent sodium channel; Ks = slow (M-type) potassium channel; Ca = high-threshold calcium channel; KCa = calcium-dependent potassium channel). (**B**) Diagram showing a rhythmically-driven target population of PC cells (PC^T) competing with an asynchronously-driven distractor population (PC^D) through a shared population of inhibitory IN cells. This figure was adapted from **Sherfey et al.** (**2018a**).

Network connectivity

PC cells provided excitation to all IN cells, mediated by α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) currents. IN cells in turn provided strong feedback inhibition mediated by γ-aminobutyric acid (GABA$_A$) currents to all PC cells. This combination of fast excitation and strong feedback inhibition is known to generate robust network oscillations in response to tonic drive (**Whittington et al., 2000**; **Börgers and Kopell, 2005**). AMPA currents were modelled by:

$$I_{AMPA} = g_{AMPA} s (V - E_{AMPA}) \tag{2}$$

where V is the postsynaptic membrane voltage, g_{AMPA} is the maximal synaptic conductance, s is a synaptic gating variable, and $E_{AMPA} = 0$ mV is the synaptic reversal potential. Synaptic gating was modeled using a first-order

kinetics scheme:

$$\frac{ds}{dt} = H(V_{pre})\frac{1-s}{\tau_r} - \frac{s}{\tau_d} \tag{3}$$

where V_{pre} is the presynaptic membrane voltage, $\tau_r = 0.4$ ms and $\tau_d = 2$ ms are time constants for neurotransmitter release and decay, respectively, and $H(V) = 1 + tanh(V/4)$ is a sigmoidal approximation to the Heaviside step function. $GABA_A$ currents are modeled in the same way with $E_{GABA} = -75$ mV and $\tau_d = 5$ ms. Maximum synaptic conductances for PC cells were (in mS/cm^2): $GABA_A$ (.1); for IN cells: AMPA (1). This simplified model excludes NMDA and $GABA_B$ currents.

Working memory items

Each PC cell in the output gate received independent Poisson spike trains (***Figure 9***) with time-varying instantaneous rate $\lambda(t)$ (sp/s) and time-averaged rate $r = \langle\lambda\rangle$; spikes were integrated in a synaptic gate s_{inp} with exponential AMPAergic decay contributing to an excitatory synaptic current $I_{inp} = g_{inp}s_{inp}(V - E_{AMPA})$ with maximal conductance g_{inp} (mS/cm^2). WM items were modeled by collections of spike trains with the same instantaneous rate-modulation. A given input item to a PC output population can be interpreted as conveying rate-coded information from a population of excitatory pyramidal cells in superficial PFC that are collectively in a particular dynamical state.

Population rate-coding was incorporated into a WM item using a spatial pattern of time-averaged firing rates $r(x)$ for spike trains driving an output population with cells indexed by x . Spatial profiles were either uniform or Gaussian bumps centered on different output neurons. Signals from WM items in different dynamical states were generated by modulating instantaneous rates $\lambda(t)$ according to the activity patterns exhibited by item-encoding populations in those states. Signals from WM items in an asynchronous state were modeled by Poisson spike trains with constant rate $\lambda(t) = r$ (***Figure 9***A) whereas signals from WM items in an oscillatory state were modeled using periodically-modulated instantaneous rates (***Figure 9***B). WM items with sine wave modulation had $\lambda(t) = r(1 + sin(2\pi f t))/2$ parameterized by r (sp/s) and rate modulation frequency f (Hz). Sinusoidal modulation causes spike synchrony (the interval over which spikes are spread within each cycle) to covary with frequency as the same number of spikes become spread over a larger period as frequency decreases. Thus, we also investigated oscillatory inputs with square wave modulation in order to differentiate the effects of synchrony and frequency while maintaining the ability to compare our results with other work. Square wave rate-modulation results in periodic trains of spikes with fixed synchrony (pulse packets) parameterized by r (sp/s), inter-pulse frequency f (Hz), and pulse width δ (ms). δ reflects the synchrony of spikes in the source population with smaller values implying greater synchrony; decreasing δ corresponds to decreasing the duty cycle of the square wave. For the square wave input, we chose to hold constant r so that across frequencies the only significant change is in the patterning of spikes and not the total number of spikes; this results in larger pulses being delivered to postsynaptic PCs at lower frequencies as would be expected if lower frequencies are produced by larger networks (***Nunez et al., 2006***). Throughout the paper, WM item parameters r, f, and δ are assigned superscripts T or D indicating whether they represent "target" or "distractor" items, respectively (see ***Table 1*** for notation details).

All principal cells in the output gate received additional asynchronous inputs representing uncorrelated background activity from 100 cells in other brain areas spiking at 1 sp/s. Notably feedforward inhibition was excluded from the present work so that asynchronous items were maximally effective at driving PC cells in the output gate. Control values for WM item parameters were $r = 1000$ sp/s (corresponding to an item-encoding population with 1000 projection neurons spiking at 1 sp/s); $\delta = 1$ ms (high synchrony), 10 ms (medium synchrony), or 19 ms (low synchrony), and $g_{inp} = .0015$ mS/cm^2. High synchrony inputs are similar to strong, periodic spikes while medium and low synchrony inputs distribute spikes uniformly over intervals comparable to sine waves at 100 Hz and 53 Hz, respectively.

In simulations probing resonant properties of the output gate, the item modulation frequency f was varied from 1 Hz to 50 Hz (in 1 Hz steps) across simulations. In simulations exploring output gating among parallel pathways, input items had the same mean strength (i.e., r); this ensures that any difference between the ability of WM items to drive their corresponding outputs resulted from differences in the dynamical states of the item-encoding populations and not differences in their activity levels.

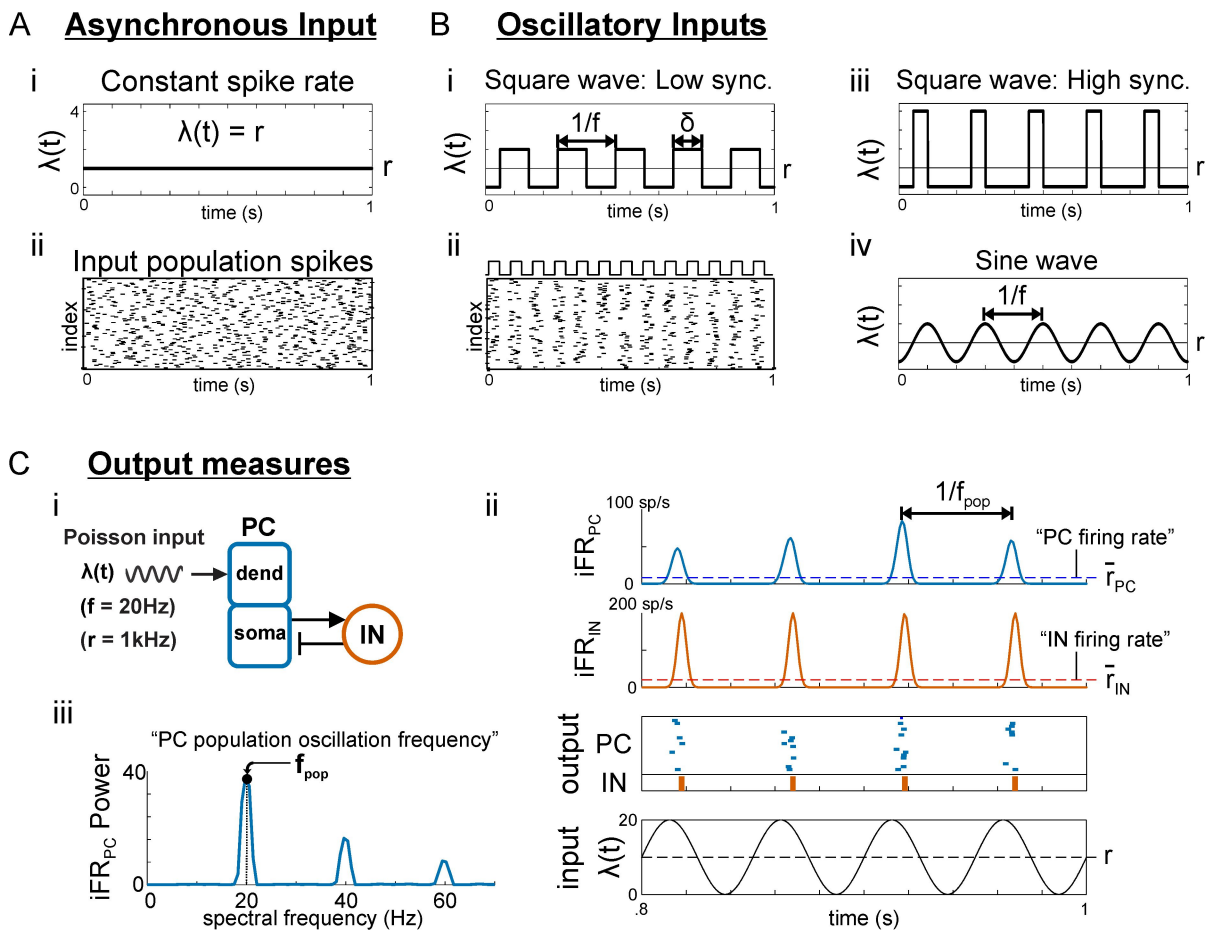


Figure 9. Input network activity. (**A**) Asynchronous Poisson input with (i) constant instantaneous rate r_{inp} and (ii) raster for 100 input cells with $r_{inp} = 10$ sp/s (equivalent to 1 input cell with $r_{inp} = 1000$ sp/s). (**B**) Poisson inputs with oscillatory instantaneous rate-modulation. (i) Instantaneous rate modulated by low synchrony square wave, parameterized by pulse width δ_{inp} and inter-pulse frequency f_{inp}. (ii) raster plot produced by square wave input. (iii) High synchrony, square wave rate-modulation. (iv) sine wave modulation, parameterized by frequency f_{inp}. (**C**) Output measures for the PC/IN network. (i) Diagram of a PC/IN network receiving an input from (B). (ii) Plots showing the instantaneous firing rate (iFR) computed for each population using Gaussian kernel regression on the spike raster. Time-averaged firing rates are defined by the mean iFR for each population. (iii) Power spectrum showing how population frequency, f_{osc}, is defined by the spectral frequency with peak power in the iFR. f_{osc} corresponds to the inverse of the period between spike volleys visible in the raster plot. This figure was adapted from ***Sherfey et al.*** (***2018a***).

Data analysis

For each simulation, instantaneous output firing rates, iFR, were computed with Gaussian kernel regression on population spike times using a kernel with 6 ms width for visualization and 2 ms for calculating the power spectrum. Mean population firing rates, $\bar{r}_{PC}$ and $\bar{r}_{IN}$, were computed by averaging iFR over time for PC and IN populations, respectively; they index overall activity levels by the average firing rate of the average cell in the population (***Figure 9***Ci-ii). The frequency of an output population oscillation, f_{pop}, is the dominant frequency of the iFR oscillation and was identified as the spectral frequency with peak power in Welch's spectrum of the iFR (***Figure 9***Ciii). In the beta/gamma range, this is equivalent to the inverse of the period between output spike volleys. The natural frequency f_N of the output network was identified as the population frequency f_{pop} produced in response to an asynchronous input.

Across simulations varying input frequencies, statistics were plotted as the mean ± standard deviation calculated across 10 realizations. Input frequency-dependent plots of mean firing rates and population frequencies will be called response profiles. The resonant input frequency, f_R^{pop}, maximizing output oscillation frequency

was found from peaks in f_{pop} population frequency profiles in the beta/gamma range. Our measure of spiking and oscillatory activity in the strongly-driven network differs from measures used in work on resonance in weakly-driven networks (***Richardson et al., 2003***; ***Akam and Kullmann, 2010***; ***Ledoux and Brunel, 2011***) and is more similar to measures used for strongly-driven single cells (***Rotstein, 2017***). See ***Sherfey et al.*** (***2018a***) for a comparison of perspectives on resonance in cells and networks.

Simulation tools

All models were implemented in Matlab using the DynaSim toolbox (***Sherfey et al., 2018b***) (http://dynasimtoolbox.org) and are publicly available online at: http://github.com/jsherfey/PFC_models. Numerical integration was performed using a 4th-order Runge-Kutta method with a fixed time step of 0.01 ms. Simulations were run for 2500ms and repeated 10 times. The network was allowed to settle to steady-state before external signals were delivered at 400 ms. Plots of instantaneous responses begin at signal onset. The first 500 ms of response was excluded from analysis, although including the transient did not alter our results significantly.

Acknowledgments

We would like to acknowledge Joachim Hass and Michelle McCarthy for early discussions of our modeling results, as well as Andre Bastos and Mikael Lundqvist for discussions relating our modeling work to their experiments.

This work was supported by the U. S. Army Research Office under award number ARO W911NF-12-R-0012-02 to N. K. and the U. S. Office of Naval Research under award number ONR MURI N00014-16-1-2832 to M. H.

References

Adams NE, Sherfey JS, Kopell NJ, Whittington MA, LeBeau FEN. Hetereogeneity in Neuronal Intrinsic Properties: A Possible Mechanism for Hub-Like Properties of the Rat Anterior Cingulate Cortex during Network Activity. eNeuro. 2017 Mar; 4(1). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5340498/, doi: 10.1523/ENEURO.0313-16.2017.

Akam T, Kullmann DM. Oscillations and Filtering Networks Support Flexible Routing of Information. Neuron. 2010 Jul; 67(2):308–320. http://www.sciencedirect.com/science/article/pii/S0896627310004770, doi: 10.1016/j.neuron.2010.06.019.

Amiez C, Joseph JP, Procyk E. Anterior cingulate error-related activity is modulated by predicted reward. The European Journal of Neuroscience. 2005 Jun; 21(12):3447–3452. doi: 10.1111/j.1460-9568.2005.04170.x.

Ardid S, Sherfey JS, McCarthy MM, Hass J, Pittman-Polletta BR, Kopell N. Biased competition in the absence of input bias: predictions from corticostriatal computation. bioRxiv. 2018 Jun; p. 258053. https://www.biorxiv.org/content/10.1101/258053v2, doi: 10.1101/258053.

Ardid S, Wang XJ. A Tweaking Principle for Executive Control: Neuronal Circuit Mechanism for Rule-Based Task Switching and Conflict Resolution. Journal of Neuroscience. 2013 Dec; 33(50):19504–19517. http://www.jneurosci.org/content/33/50/19504, doi: 10.1523/JNEUROSCI.1356-13.2013.

Ardid S, Wang XJ, Compte A. An Integrated Microcircuit Model of Attentional Processing in the Neocortex. Journal of Neuroscience. 2007 Aug; 27(32):8486–8495. http://www.jneurosci.org/content/27/32/8486, doi: 10.1523/JNEUROSCI.1145-07.2007.

Ardid S, Wang XJ, Gomez-Cabrero D, Compte A. Reconciling Coherent Oscillation with Modulationof Irregular Spiking Activity in Selective Attention:Gamma-Range Synchronization between Sensoryand Executive Cortical Areas. Journal of Neuroscience. 2010 Feb; 30(8):2856–2870. http://www.jneurosci.org/content/30/8/2856, doi: 10.1523/JNEUROSCI.4222-09.2010.

Baddeley AD, Hitch G. Working Memory. In: Bower GH, editor. *Psychology of Learning and Motivation*, vol. 8 Academic Press; 1974.p. 47–89. http://www.sciencedirect.com/science/article/pii/S0079742108604521, doi: 10.1016/S0079-7421(08)60452-1.

Badre D, Frank MJ. Mechanisms of Hierarchical Reinforcement Learning in Cortico–Striatal Circuits 2: Evidence from fMRI. Cerebral Cortex. 2012 Mar; 22(3):527–536. https://academic.oup.com/cercor/article/22/3/527/325808/Mechanisms-of-Hierarchical-Reinforcement-Learning, doi: 10.1093/cercor/bhr117.

Barbas H. General cortical and special prefrontal connections: principles from structure to function. Annual Review of Neuroscience. 2015 Jul; 38:269–289. doi: 10.1146/annurev-neuro-071714-033936.

Barbas PH. Frontal Cortex. In: Pfaff DW, editor. *Neuroscience in the 21st Century* Springer New York; 2013.p. 1289–1334. http://link.springer.com/referenceworkentry/10.1007/978-1-4614-1997-6_45.

Basar-Eroglu C, Brand A, Hildebrandt H, Karolina Kedzior K, Mathes B, Schmiedt C. Working memory related gamma oscillations in schizophrenia patients. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology. 2007 Apr; 64(1):39–45. doi: 10.1016/j.ijpsycho.2006.07.007.

Bastos AM, Loonis R, Kornblith S, Lundqvist M, Miller EK. Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. Proceedings of the National Academy of Sciences. 2018 Jan; p. 201710323. http://www.pnas.org/lookup/doi/10.1073/pnas.1710323115, doi: 10.1073/pnas.1710323115.

Bhandari A, Badre D. Learning and transfer of working memory gating policies. Cognition. 2018 Mar; 172:89–100. http://www.sciencedirect.com/science/article/pii/S0010027717303037, doi: 10.1016/j.cognition.2017.12.001.

Bloem B, Poorthuis RB, Mansvelder HD. Cholinergic modulation of the medial prefrontal cortex: the role of nicotinic receptors in attention and regulation of neuronal activity. Frontiers in Neural Circuits. 2014 Mar; 8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3949318/, doi: 10.3389/fncir.2014.00017.

Brincat SL, Miller EK. Prefrontal Cortex Networks Shift from External to Internal Modes during Learning. Journal of Neuroscience. 2016 Sep; 36(37):9739–9754. http://www.jneurosci.org/content/36/37/9739, doi: 10.1523/JNEUROSCI.0274-16.2016.

Buschman T, Denovellis E, Diogo C, Bullock D, Miller E. Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. Neuron. 2012 Nov; 76(4):838–846. http://www.sciencedirect.com/science/article/pii/S0896627312008823, doi: 10.1016/j.neuron.2012.09.029.

Börgers C, Kopell N. Effects of Noisy Drive on Rhythms in Networks of Excitatory and Inhibitory Neurons. Neural Computation. 2005 Mar; 17(3):557–608. https://doi.org/10.1162/0899766053019908, doi: 10.1162/0899766053019908.

Cannon J, Kopell N. The Leaky Oscillator: Properties of Inhibition-Based Rhythms Revealed through the Singular Phase Response Curve. SIAM Journal on Applied Dynamical Systems. 2015 Jan; 14(4):1930–1977. http://epubs.siam.org/doi/abs/10.1137/140977151, doi: 10.1137/140977151.

Cannon J, McCarthy MM, Lee S, Lee J, Börgers C, Whittington MA, Kopell N. Neurosystems: brain rhythms and cognitive processing. European Journal of Neuroscience. 2014 Mar; 39(5):705–719. http://onlinelibrary.wiley.com/doi/10.1111/ejn.12453/abstract, doi: 10.1111/ejn.12453.

Cho RY, Konecky RO, Carter CS. Impairments in frontal cortical gamma synchrony and cognitive control in schizophrenia. Proceedings of the National Academy of Sciences. 2006 Dec; 103(52):19878–19883. http://www.pnas.org/content/103/52/19878, doi: 10.1073/pnas.0609440103.

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. Cerebral Cortex. 2000 Sep; 10(9):910–923. https://academic.oup.com/cercor/article/10/9/910/289091, doi: 10.1093/cercor/10.9.910.

Curtis CE, D'Esposito M. Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences. 2003 Sep; 7(9):415–423.

DeFelipe J. Types of neurons, synaptic connections and chemical characteristics of cells immunoreactive for calbindin-D28K, parvalbumin and calretinin in the neocortex. Journal of Chemical Neuroanatomy. 1997 Dec; 14(1):1–19. http://www.sciencedirect.com/science/article/pii/S0891061897100138, doi: 10.1016/S0891-0618(97)10013-8.

Dong Y, White FJ. Dopamine D1-Class Receptors Selectively Modulate a Slowly Inactivating Potassium Current in Rat Medial Prefrontal Cortex Pyramidal Neurons. Journal of Neuroscience. 2003 Apr; 23(7):2686–2695. http://www.jneurosci.org/content/23/7/2686.

Durstewitz D, Seamans JK. The computational role of dopamine D1 receptors in working memory. Neural Networks. 2002 Jun; 15(4):561–572. http://www.sciencedirect.com/science/article/pii/S0893608002000497, doi: 10.1016/S0893-6080(02)00049-7.

Durstewitz D, Seamans JK, Sejnowski TJ. Dopamine-Mediated Stabilization of Delay-Period Activity in a Network Model of Prefrontal Cortex. Journal of Neurophysiology. 2000 Mar; 83(3):1733–1750. http://jn.physiology.org/content/83/3/1733.

Frank MJ, Badre D. Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. Cerebral Cortex. 2012 Mar; 22(3):509–526. https://academic.oup.com/cercor/article/22/3/509/325387, doi: 10.1093/cercor/bhr114.

Frank MJ, Loughry B, O'Reilly RC. Interactions between frontal cortex and basal ganglia in working memory: A computational model. Cognitive, Affective, & Behavioral Neuroscience. 2001 Jun; 1(2):137–160. http://link.springer.com/article/10.3758/CABN.1.2.137, doi: 10.3758/CABN.1.2.137.

Funahashi S. Thalamic mediodorsal nucleus and its participation in spatial working memory processes: comparison with the prefrontal cortex. Frontiers in Systems Neuroscience. 2013 Jul; 7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3728470/, doi: 10.3389/fnsys.2013.00036.

Fuster JM. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. Journal of Neurophysiology. 1973 Jan; 36(1):61–78. https://www.physiology.org/doi/abs/10.1152/jn.1973.36.1.61, doi: 10.1152/jn.1973.36.1.61.

Fuster J. The Prefrontal Cortex. Academic Press; 2015.

Goldman-Rakic PS. Cellular basis of working memory. Neuron. 1995 Mar; 14(3):477–485. http://www.sciencedirect.com/science/article/pii/0896627395903046, doi: 10.1016/0896-6273(95)90304-6.

Hasselmo ME. The role of acetylcholine in learning and memory. Current Opinion in Neurobiology. 2006 Dec; 16(6):710–715. http://www.sciencedirect.com/science/article/pii/S095943880600122X, doi: 10.1016/j.conb.2006.09.002.

Hasselmo ME, Stern CE. A network model of behavioural performance in a rule learning task. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018 Apr; 373(1744):20170275. https://royalsocietypublishing.org/doi/full/10.1098/rstb.2017.0275, doi: 10.1098/rstb.2017.0275.

Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997 Nov; 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, doi: 10.1162/neco.1997.9.8.1735.

Kaski S, Kohonen T. Winner-take-all networks for physiological models of competitive learning. Neural Networks. 1994 Jan; 7(6):973–984. http://www.sciencedirect.com/science/article/pii/S0893608005801546, doi: 10.1016/S0893-6080(05)80154-6.

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. Anterior cingulate conflict monitoring and adjustments in control. Science (New York, NY). 2004 Feb; 303(5660):1023–1026. doi: 10.1126/science.1089910.

Khan AG, Poort J, Chadwick A, Blot A, Sahani M, Mrsic-Flogel TD, Hofer SB. Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. Nature Neuroscience. 2018 Jun; 21(6):851–859. doi: 10.1038/s41593-018-0143-z.

Komorowski RW, Garcia CG, Wilson A, Hattori S, Howard MW, Eichenbaum H. Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2013 May; 33(18):8079–8087. doi: 10.1523/JNEUROSCI.5458-12.2013.

Kriete T, Noelle DC. Generalisation benefits of output gating in a model of prefrontal cortex. Connection Science. 2011 Jun; 23(2):119–129. https://doi.org/10.1080/09540091.2011.569881, doi: 10.1080/09540091.2011.569881.

Kriete T, Noelle DC, Cohen JD, O'Reilly RC. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. Proceedings of the National Academy of Sciences. 2013 Sep; p. 201303547. https://www.pnas.org/content/early/2013/09/18/1303547110, doi: 10.1073/pnas.1303547110.

Kritzer MF, Goldman-Rakic PS. Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. The Journal of Comparative Neurology. 1995 Aug; 359(1):131–143. doi: 10.1002/cne.903590109.

Ledoux E, Brunel N. Dynamics of Networks of Excitatory and Inhibitory Neurons in Response to Time-Dependent Inputs. Frontiers in Computational Neuroscience. 2011 May; 5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103906/, doi: 10.3389/fncom.2011.00025.

Levitt JB, Lewis DA, Yoshioka T, Lund JS. Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46). The Journal of Comparative Neurology. 1993 Dec; 338(3):360–376. doi: 10.1002/cne.903380304.

Lundqvist M, Compte A, Lansner A. Bistable, Irregular Firing and Population Oscillations in a Modular Attractor Memory Network. PLOS Computational Biology. 2010 Jun; 6(6):e1000803. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000803, doi: 10.1371/journal.pcbi.1000803.

Lundqvist M, Herman P, Lansner A. Theta and Gamma Power Increases and Alpha/Beta Power Decreases with Memory Load in an Attractor Network Model. Journal of Cognitive Neuroscience. 2011 Mar; 23(10):3008–3020. https://doi.org/10.1162/jocn_a_00029, doi: 10.1162/jocn_a_00029.

Lundqvist M, Herman P, Warden MR, Brincat SL, Miller EK. Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nature Communications. 2018; 9(1):394. doi: 10.1038/s41467-017-02791-8.

Lundqvist M, Rose J, Herman P, Brincat S, Buschman T, Miller E. Gamma and Beta Bursts Underlie Working Memory. Neuron. 2016 Apr; 90(1):152–164. http://www.sciencedirect.com/science/article/pii/S0896627316001458, doi: 10.1016/j.neuron.2016.02.028.

Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature. 2013 Nov; 503(7474):78–84. http://www.nature.com/doifinder/10.1038/nature12742, doi: 10.1038/nature12742.

Medalla M, Barbas H. Synapses with inhibitory neurons differentiate anterior cingulate from dorsolateral prefrontal pathways associated with cognitive control. Neuron. 2009 Feb; 61(4):609–620. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2804928/, doi: 10.1016/j.neuron.2009.01.006.

Melrose RJ, Poulin RM, Stern CE. An fMRI investigation of the role of the basal ganglia in reasoning. Brain Research. 2007 Apr; 1142:146–158. http://www.sciencedirect.com/science/article/pii/S0006899307001321, doi: 10.1016/j.brainres.2007.01.060.

Miller EK. The prefontral cortex and cognitive control. Nature Reviews Neuroscience. 2000 Oct; 1(1):59–65. https://www.nature.com/articles/35036228, doi: 10.1038/35036228.

Miller EK, Lundqvist M, Bastos AM. Working Memory 2.0. Neuron. 2018 Oct; 100(2):463–475. http://www.sciencedirect.com/science/article/pii/S0896627318308250, doi: 10.1016/j.neuron.2018.09.023.

Miyake A, Shah P. Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. Cambridge University Press; 1999. Google-Books-ID: Dmu23_pfbb8C.

Nunez PL, Nunez EPoBEPL, Srinivasan R, Srinivasan APoCSR. Electric Fields of the Brain: The Neurophysics of EEG. Oxford University Press; 2006. Google-Books-ID: fUv54as56_8C.

O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation. 2006 Feb; 18(2):283–328. doi: 10.1162/089976606775093909.

Parnaudeau S, O'Neill PK, Bolkan SS, Ward RD, Abbas AI, Roth BL, Balsam PD, Gordon JA, Kellendonk C. Inhibition of mediodorsal thalamus disrupts thalamofrontal connectivity and cognition. Neuron. 2013 Mar; 77(6):1151–1162. doi: 10.1016/j.neuron.2013.01.038.

Povysheva NV, Zaitsev AV, Rotaru DC, Gonzalez-Burgos G, Lewis DA, Krimer LS. Parvalbumin-positive basket interneurons in monkey and rat prefrontal cortex. Journal of Neurophysiology. 2008 Oct; 100(4):2348–2360. doi: 10.1152/jn.90396.2008.

Renart A, Rocha Jdl, Bartho P, Hollender L, Parga N, Reyes A, Harris KD. The Asynchronous State in Cortical Circuits. Science. 2010 Jan; 327(5965):587–590. http://science.sciencemag.org/content/327/5965/587, doi: 10.1126/science.1179850.

Richardson MJE, Brunel N, Hakim V. From Subthreshold to Firing-Rate Resonance. Journal of Neurophysiology. 2003 May; 89(5):2538–2554. http://jn.physiology.org/content/89/5/2538, doi: 10.1152/jn.00955.2002.

Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR. Reactivation of latent working memories with transcranial magnetic stimulation. Science. 2016 Dec; 354(6316):1136–1139. http://www.sciencemag.org/cgi/doi/10.1126/science.aah7011, doi: 10.1126/science.aah7011.

Rotstein HG. Spiking Resonances In Models With The Same Slow Resonant And Fast Amplifying Currents But Different Subthreshold Dynamic Properties. bioRxiv. 2017 Apr; p. 128611. https://www.biorxiv.org/content/early/2017/04/19/128611, doi: 10.1101/128611.

Roux F, Wibral M, Mohr HM, Singer W, Uhlhaas PJ. Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2012 Sep; 32(36):12411–12420. doi: 10.1523/JNEUROSCI.0421-12.2012.

Shen Z, Popov V, Delahay AB, Reder LM. Item strength affects working memory capacity. Memory & Cognition. 2018 Feb; 46(2):204–215. https://doi.org/10.3758/s13421-017-0758-4, doi: 10.3758/s13421-017-0758-4.

Sherfey JS, Ardid S, Hass J, Hasselmo ME, Kopell NJ. Flexible resonance in prefrontal networks with strong feedback inhibition. PLoS computational biology. 2018; 14(8):e1006357. doi: 10.1371/journal.pcbi.1006357.

Sherfey JS, Soplata AE, Ardid S, Roberts EA, Stanley DA, Pittman-Polletta BR, Kopell NJ. DynaSim: A MATLAB Toolbox for Neural Modeling and Simulation. Frontiers in Neuroinformatics. 2018; 12. https://www.frontiersin.org/articles/10.3389/fninf.2018.00010/full, doi: 10.3389/fninf.2018.00010.

Siegel M, Warden MR, Miller EK. Phase-dependent neuronal coding of objects in short-term memory. Proceedings of the National Academy of Sciences. 2009 Dec; 106(50):21341–21346. http://www.pnas.org/content/106/50/21341, doi: 10.1073/pnas.0908193106.

Sohal VS, Zhang F, Yizhar O, Deisseroth K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. Nature. 2009 Jun; 459(7247):698–702. doi: 10.1038/nature07991.

Tegnér J, Compte A, Wang XJ. The dynamical stability of reverberatory neural circuits. Biological Cybernetics. 2002 Dec; 87(5-6):471–481. https://link.springer.com/article/10.1007/s00422-002-0363-9, doi: 10.1007/s00422-002-0363-9.

Tzur G, Berger A. Fast and slow brain rhythms in rule/expectation violation tasks: Focusing on evaluation processes by excluding motor action. Behavioural Brain Research. 2009 Mar; 198(2):420–428. http://www.sciencedirect.com/science/article/pii/S0166432808006359, doi: 10.1016/j.bbr.2008.11.041.

Wang XJ. Synaptic Basis of Cortical Persistent Activity: the Importance of NMDA Receptors to Working Memory. Journal of Neuroscience. 1999 Nov; 19(21):9587–9603. http://www.jneurosci.org/content/19/21/9587, doi: 10.1523/JNEUROSCI.19-21-09587.1999.

Wang XJ, Buzsáki G. Gamma Oscillation by Synaptic Inhibition in a Hippocampal Interneuronal Network Model. Journal of Neuroscience. 1996 Oct; 16(20):6402–6413. http://www.jneurosci.org/content/16/20/6402.

Whittington MA, Traub RD, Jefferys JG. Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. Nature. 1995 Feb; 373(6515):612–615. doi: 10.1038/373612a0.

Whittington MA, Traub RD, Kopell N, Ermentrout B, Buhl EH. Inhibition-based rhythms: experimental and mathematical observations on network dynamics. International Journal of Psychophysiology. 2000 Dec; 38(3):315–336. http://www.sciencedirect.com/science/article/pii/S0167876000001732, doi: 10.1016/S0167-8760(00)00173-2.

Zhu H, Paschalidis IC, Hasselmo ME. Neural circuits for learning context-dependent associations of stimuli. Neural Networks. 2018 Nov; 107:48–60. http://www.sciencedirect.com/science/article/pii/S089360801830217X, doi: 10.1016/j.neunet.2018.07.018.

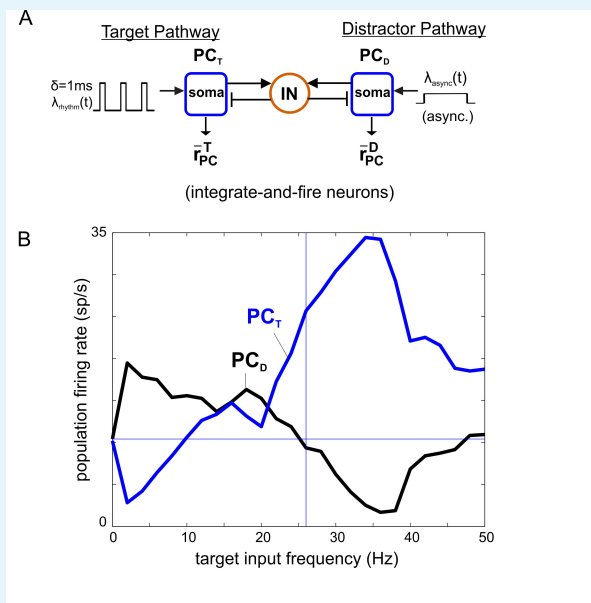
Appendix 1

Integrate-and-fire networks

Figure 1Cii for the output gate network of conductance-based PFC neurons was reproduced using a network of simpler integrate-and-fire neurons. Qualitative features of the prefrontal cortical output gate model were reproduced in a simpler model of leaky integrate-and-fire (LIF) neurons with membrane potential V (mV) governed by:

$$\frac{dV}{dt} = -I_{inp}(t, V) - g_l(V - E_l) - g_{syn}(V - E_{syn})$$

where t is time (ms), $g_l = 0.1$, $E_l = -65$ mV, $I_{inp}(t, V)$ is an excitatory current (μA/cm^2) reflecting inputs from external sources described in the Methods section, and I_{syn} denotes synaptic currents (μA/cm^2) with double exponential conductances driven by other populations. When the membrane potential reaches the threshold of 0 mV, the voltage is reset and held at -65 mV for a refractory period of 3 ms. There were 25 PCs and 5 INs. For PCs, synaptic inputs were inhibitory with $g_{syn} = 0.1$, $E_{syn} = -80$ mV, 2 ms decay and 0.4 rise time constants. For INs, synaptic inputs were excitatory with $g_{syn} = 0.03$, $E_{syn} = 0$ mV, 10 ms decay and 0.2 rise time constants. Inputs to the LIF network were the same as the more detailed PFC network described in the Methods section except that $g_{inp} = .00375$ mS/cm^2 and $g_{noise} = 0.0056$ mS/cm^2.



Appendix 1 Figure 1. Competition in the output gate with integrate-and-fire neurons. (**A**) Simplified output gate model with single-compartment integrate-and-fire neurons. (**B**) Mean firing rate outputs for target (blue) and distractor (black) as target input frequency is increased. The distractor is suppressed when the target oscillates faster than the natural frequency, as in the case of the more detailed output gate model with Hodgkin-Huxley neurons.