A computational model for spatial working memory deficits in schizophrenia

Maria Cano-Colino, Albert Compte

Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS),
Rosselló 149, 08036 Barcelona, Spain

Keywords: delayed response, prefrontal cortex, NMDA receptor, inhibition, network, attractor

Corresponding author: Albert Compte
Rosselló 149
08036 Barcelona
Spain
acompte@clinic.ub.es
+34 93 2275400 ext. 4151

Acknowledgments: We are indebted to Francesc Artigas and Rita Almeida for fruitful discussions. Funded by Ministry for Science and Innovation of Spain and European Regional Development Fund (Ref: BFU2009_09537). The work was carried out at the Esther Koplowitz Centre, Barcelona

Disclosure: The authors declare that there are no conflicts of interest.
Abstract

Cognitive deficits in schizophrenia have been hypothesized to be caused by altered synaptic transmission in circuits of the prefrontal cortex. Two main hypotheses have been put forward: reduced inhibition and hypofunctional NMDA receptors. Recently, Lee et al. (2008) found that spatial working memory deficits in schizophrenic patients include a disproportionately high incidence of high-confidence error responses. Here, we studied what synaptic dysfunction can generate this specific behavioral deficit using a computational network model of spatial working memory. We developed quantitative behavioral readout from our network simulations, which reflected the qualitative properties of underlying neural dynamics. We then analyzed the behavioral effect of the GABAergic and glutamatergic hypotheses on our network simulations. We found that reduction in inhibitory transmission in the network caused a reduction in performance through an increase of high-confidence errors, as in the experimental data. In contrast, a concerted reduction in NMDA-receptor-dependent transmission reduced performance via increased low-confidence errors. Only when NMDA receptors were specifically depleted in interneurons did the behavioral read-out of our network mimic the behavioral results for schizophrenic patients. Thus, dynamics in our model network supports a role of both global inhibition reduction and hypofunctional NMDA receptors in interneurons in generating the behavioral deficits of simple spatial working memory tasks in schizophrenia.
Introduction

Working memory, the capability of temporarily storing and manipulating information as part of the performance of complex cognitive tasks, is impaired in schizophrenic patients [1-3]. Schizophrenic patients have deficits in all components of working memory, from maintenance to manipulation, inhibition or updating [4]. These deficits are permanent and remain when clinical symptoms remit [5], and they have been found to mark genetic predisposition toward schizophrenia [6,7]. Hence, dysfunctions in working memory are now thought to be a core deficit of this disease [4,8,9].

The cognitive deficits of schizophrenia are thought to reflect impairments in the function of the dorsolateral prefrontal cortex (DLPFC) and its interactions with other brain regions, such as the parietal cortex, the hippocampus, the thalamus and the striatum, possibly caused by imbalances in neurotransmitter systems [4,10,11]. In healthy individuals, a similar network of areas, pivoting around the DLPFC, are identified by neuroimaging studies as supporting working memory function [12]. Also the DLPFC has been singled out in neurophysiological experiments in nonhuman primates as the locus of persistent activity, the substrate of working memory maintenance in delayed response tasks [13,14].

The convergent evidence from neuroimaging and neurophysiological studies that links circuit dynamics in the DLPFC to working memory has motivated computational efforts to formulate explicit biological models of the mechanisms that support working memory in DLPFC. In the last two decades, several biophysical computational models have been proposed to explain persistent prefrontal neural activity in monkeys performing a working memory task [15-18]. One particularly influential family of such models is based on attractor dynamics [19-22]. Attractor models are characterized by internal dynamics such that the system evolves with time approaching and stabilizing in one of several possible pre-specified states, or attractors. In attractor models of working memory, a network of recurrently coupled biological neurons maintains stably, through synaptic reverberation, persistent and selective firing-rate elevations, which represent a stimulus long after it disappears.

Recently, synaptic imbalances in these computational models have been associated with various general symptoms of schizophrenia [23-25]. It was observed that alterations of synaptic transmission that are related with schizophrenia (reduction in GABA synthesis...
and/or hypofunctional NMDA receptor transmission [11,26,27]) induce changes in the stability of mnemonic network attractors to perturbative noisy inputs. These destabilized network states can give rise to diverse behaviors reminiscent of the very heterogeneous symptoms of schizophrenia [24,25]. We intend here to apply this general framework to a spatial working memory task and generate concrete predictions for this parametric task that can guide future experimental investigations on the mechanisms of cognitive dysfunction in schizophrenia.

Visuo-spatial delayed response tasks have been fundamental model tasks to study both the electrophysiology of working memory [28,29] and the cognitive deficits of schizophrenia [3,5,7,30]. In these tasks, participants have to remember the location of one or several visual stimuli in order to perform correctly at the end of a blank-screen delay period. During this task, recordings in the DLPFC of monkeys reveal that a large portion of neurons represent the cue location in a parametric way during the delay period, following a bell-shaped tuning curve of angular stimulus location [28]. These neural dynamics are quantitatively well captured by a line-attractor biophysical network model, which can relate synaptic and cellular properties to behavioral performance [19]. In a line-attractor model network dynamics converges towards one of a continuous family of attractors, as opposed to a discrete subset of attractors, and is therefore adapted to represent the continuous location of the stimulus in spatial working memory [31]. However, this model has not yet been examined in the context of schizophrenia, by studying its “behavioral” deficits following mechanistic alterations associated with this disease. This is remarkable given the important role that this task has played in establishing working memory deficits in schizophrenia [5,7,30]. In one such study, Lee et al. have recently shown that behavioral errors for which participants declare high confidence in their responses, of high incidence in schizophrenic patients, have different neural substrate than low-confidence errors [32]. This is consistent with persistent activity subserving working memory maintenance: erroneous responses may arise both from “false memories” (persistent activity of the wrong representation) or from forgetting the relevant information (failure of persistent activity), and this should be reflected in different image contrasts in neuroimaging studies. Here, we seek to give a neural and computational basis to this observation while we study performance changes in the spatial working memory model caused by synaptic alterations associated with schizophrenia.
Methods

We used a previously published network model of spatial working memory [19] to study performance changes in the model upon synaptic manipulations that have been associated with schizophrenia. Model parameters were taken from the “modulated parameter set” in [19], which was meant to represent the circuit in the optimally neuromodulated situation of an engaged working memory task. We provide here a brief description of this network model, for extended details please see [19].

The network model represents a local circuit of the monkey dorsolateral prefrontal cortex. To simulate the local recurrent cortical network we used two populations of leaky integrate-and-fire neurons [33]: pyramidal cells \((N_E = 1024)\) and interneurons \((N_I = 256)\). The membrane voltage \(V_m\) of each neuron obeys the following dynamical equation:

\[
C_m \frac{dV_m}{dt} = -I_L - I_{\text{syn},e} - I_{\text{syn},i} - I_{\text{ext}} + I_s
\]

where \(C_m\) represents the membrane capacitance of the neuron. When \(V_m\) reaches a threshold value \(V_{th}\), \(V_m\) is reset to \(V_{res}\) and stays there for an absolute refractory period \(\tau_{\text{ref}}\). \(I_{\text{ext}}\) represents random synaptic inputs from outside the network, simulated as uncorrelated Poisson spike trains activating AMPA channels (see below) of conductance \(g_{\text{ext}}\) at a rate \(v_{\text{ext}}\). \(I_s\) is the input current associated with stimulus presentation (see below). The leak current is \(I_L = g_L (V_m - E_L)\), with \(g_L\) and \(E_L\) being the conductance and reversal potential of leak channels. \(I_{\text{syn},e}\) and \(I_{\text{syn},i}\) are the recurrent synaptic inputs from presynaptic pyramidal cells and interneurons, respectively. Details of synaptic transmission are given below. The intrinsic parameters that characterize pyramidal cells are: \(C_m = 0.5 \, \text{nF}, \, g_L = 25 \, \text{nS}, \, E_L = -70 \, \text{mV}, \, V_{th} = -50 \, \text{mV}, \, V_{res} = -60 \, \text{mV}, \, v_{\text{ext}} = 1,650 \, \text{Hz}, \, g_{\text{ext}} = 4.99 \, \text{nS}, \, \tau_{\text{ref}} = 2 \, \text{ms}\). For interneurons \(C_m = 0.2 \, \text{nF}, \, g_L = 20 \, \text{nS}, \, E_L = -70 \, \text{mV}, \, V_{th} = -50 \, \text{mV}, \, V_{res} = -60 \, \text{mV}, \, v_{\text{ext}} = 1,800 \, \text{Hz}, \, g_{\text{ext}} = 1.789 \, \text{nS}, \, \tau_{\text{ref}} = 1 \, \text{ms}\).

Neurons received their recurrent excitatory inputs through NMDAR-mediated transmission and their inhibitory inputs through GABA\(_A\)Rs. As mentioned above, inputs from neurons outside the model network \((I_{\text{ext}})\) activated AMPA receptors. These conductance-based synaptic responses were calibrated by the experimentally measured dynamics of synaptic currents (for references, please see [34]). Thus, postsynaptic currents were modeled according to \(I_{\text{syn}} = g_{\text{syn}} s (V_m - V_{\text{syn}})\), where \(g_{\text{syn}}\) is a synaptic conductance, \(s\) is a synaptic gating variable, and \(V_{\text{syn}}\) is the synaptic reversal potential.
(\(V_{\text{syn}} = 0\) for excitatory synapses, \(V_{\text{syn}} = -70\) mV for inhibitory synapses). AMPAR and GABA\(_A\)R synaptic gating variables were modeled as an instantaneous jump of magnitude 1 when a spike occurred in the presynaptic neuron followed by an exponential decay with time constant 2 ms for AMPA and 10 ms for GABA\(_A\). The NMDA conductance was voltage-dependent, with \(g_{\text{syn}}\) multiplied by \(1/(1 + [\text{Mg}^{2+}] \exp(-0.062V_m)/3.57)\), \([\text{Mg}^{2+}] = 1.0\) mM. The NMDA channel kinetics were modeled by the following equations:

\[
\frac{ds}{dt} = \frac{-1}{\tau_s} s + \alpha x (1 - s) \quad \frac{dx}{dt} = \frac{-1}{\tau_x} x + \sum_i \delta(t - t_i),
\]

where \(s\) is the gating variable, \(x\) is a synaptic variable proportional to the neurotransmitter concentration in the synapse, \(t_i\) are the presynaptic spike times, \(\tau_s = 100\) ms is the decay time of NMDA currents, \(\tau_x = 2\) ms controls the rise time of NMDAR channels, and \(\alpha = 0.5\) kHz controls the saturation properties of NMDAR channels at high presynaptic firing frequencies. Parameters for synaptic transmission were taken from [19].

The basis of the network model are neurons selective to the memorized location in working memory tasks [28,29]. Pyramidal cells and interneurons were spatially distributed on a ring simulating the cortical columnar organization, labeled by their preferred direction of motion (\(\theta_i\), from 0 to 360\(^\circ\)) (Fig. 1). Connections between pyramidal cells were spatially tuned (Fig. 1), such that nearby e-cells were strongly connected, whereas distant e-cells had relatively weaker connections [19]. The connection strength \(g_{\text{syn},ij}\) between pyramidal cells \(i\) and \(j\) depended on the difference in preferred angle between the cells and was described by the equation \(g_{\text{syn},ij} = W(\theta_i - \theta_j)G_{\text{syn}}\), where \(W(\theta_i - \theta_j)\) was the sum of a constant term plus a Gaussian: \(W(\theta_i - \theta_j) = J' + (J' + J) \exp[-(\theta_i - \theta_j)^2/\sigma^2]\). \(W(\theta_i - \theta_j)\) depended on two parameters, \(J' = 1.62\) and \(\sigma = 14.4^\circ\), while \(J\) was determined from a normalization condition [19]. All the other connections (from pyramidal to interneurons and from interneurons to other neurons) were unstructured, i.e. the strength \(g_{\text{syn}}\) of these connections did not depend on the preferred angle of pre- and post-synaptic neurons (\(W_{\text{EI}} = W_{\text{IE}} = W_{\text{II}} = 1\)). The strengths of local connections in the network were: \(G_{\text{EE}} = 0.457\) nS (pyramid to pyramid); \(G_{\text{EI}} = 0.352\) nS (pyramid to interneuron); \(G_{\text{IE}} = 1.870\) nS (interneuron to pyramid); \(G_{\text{II}} = 1.436\) nS (interneuron to interneuron).
Simulations

The chosen simulation protocol resembled other behavioral protocols used in working memory experiments [28,35]. In brief, monkeys fixate a central spot during a brief presentation of a peripheral cue and throughout a subsequent delay period. After this delay, they make a saccadic eye movement to where the cue had been presented in order to obtain a reward. To mimic this behavioral protocol in our simulations, simulation trials consist of four periods: fixation (3s), cue (0.25s), delay (3s) and response. In the fixation period there are no external inputs to the network so it stays in a spontaneous, unstructured firing state. In the cue period, a cue stimulus is applied at location \( \theta_s \). This was simulated as current injection to each excitatory neuron in the network (labeled by \( \theta_i \)) of intensity \( I_c(\theta_i) = I_1 \exp[\mu(\cos(\theta_i - \theta_s) - 1)] \). We typically use \( I_1 = 0.1 \) nA and \( \mu = 5 \). During the delay, no stimulus is presented so that the network maintains the cue position in a stable pattern of network activation (activity bump).

In order to extract behavioral responses from the network simulations we computed a population vector estimation [36,37] from the network activity at the end of the delay period. Thus, if \( \{n_i, i=1..N_E\} \) are the spike counts of all the excitatory neurons labelled by \( \{\theta_i, i=1..N_E\} \) in a 50-ms window at the end of the delay period, the population vector is computed as the normalized sum of each neuron’s selectivity vector \( e_i^{\theta_i} \) (we use complex notation to operate with vectors in a compact manner) weighted by its spike count: \( P = \left( \sum_{i=1}^{N_E} n_i e_i^{\theta_i} \right) \left( \sum_{i=1}^{N_E} n_i \right)^{-1} \). We then extract the modulus \( C \) and angle \( \hat{\theta} \) of the resultant population vector: \( P = C e^{i \hat{\theta}} \). For each individual simulation trial we take \( \hat{\theta} \) as the decoded location memorized in the network activity before response initiation, this is what we will take as the “behavioral response” in this simulation trial. Correct trials will be those trials for which \( |\hat{\theta} - \theta_s| < 22.5^\circ \), where 22.5\(^\circ\) is an arbitrarily defined window around \( \theta_s \) to define correct trials. In addition, we can view \( C \) as a measure of the confidence in the response. Indeed, \( C \) measures the signal-to-noise ratio of the population code contained in the network. Arbitrarily, we took a threshold of \( C > 0.5 \) as our criterion for a confident behavioral response.

Numerical integration

The integration method used was a second-order Runge-Kutta algorithm with a time step of \( \Delta t = 0.02 \) ms. The custom code for the simulations was written in C++.
Results

We tested the performance of a previously described computational network model for selective persistent neural activity in prefrontal cortex [19] in repeated realizations of a spatial working memory task. One trial of the task consisted in a waiting period of 3 sec after visual fixation and before the presentation of a visual stimulus (duration 250 ms), followed by a delay period of 3 sec after which the location of the cue stimulus has to be reported based on the neural activity in the network at the end of the delay period. We did not simulate explicitly the response period activity of the network. We assumed that the visual stimulus, typically a bright dot, appeared in a random location restricted to be on a circle of given eccentricity from the fixation point. Thus, the stimulus location was entirely described by an angle value $\theta_s$ ($-180^\circ \leq \theta_s < 180^\circ$). This task mimics behavioral tasks used in monkeys and humans to test spatial working memory in behavioral and neurophysiological studies [3,5,28].

The network model has been presented before [19]. In brief, it consisted of 1024 pyramidal neurons and 256 interneurons (integrate-and-fire model, see Methods) interconnected via synaptic interactions of realistic dynamics (mimicking NMDA and GABA$_\alpha$ receptor mediated conductance changes) and distributed on a ring structure such that nearby neurons shared similar selectivity for nearby locations (Fig. 1). Specifically, this means that when a stimulus was presented at location $\theta_s$, the $i$-th neuron (labelled by $\theta_i$, $-180^\circ \leq \theta_i < 180^\circ$) received external inputs whose strength was maximal for $\theta_s = \theta_i$ and decayed to zero as $|\theta_s - \theta_i|$ grew. In addition, the strength of recurrent connections between any two pyramidal neurons $\theta_i$ and $\theta_j$ was also modulated similarly and was strongest when $\theta_i = \theta_j$ and decayed to a baseline value as $|\theta_i - \theta_j|$ grew (see Methods). As a result of this spatial structure in the connectivity of the network, the model can sustain persistent activity in a stable bell-shaped attractor (bump attractor, Fig. 2A) by virtue of strong reverberatory recurrent excitation among neighboring excitatory cells and strong disynaptic inhibition between excitatory cells of dissimilar selectivity [19,22]. This network activity is consistent with single-cell responses in the prefrontal cortex of monkeys performing oculomotor delayed response tasks, as previously shown [19].

When we ran repeatedly trials with the task structure defined above but with different noise realization (neurons receive via AMPA channels independent-Poisson-distributed
action potentials from outside the network that differ from trial to trial, mimicking excitatory inputs from other brain areas), network activity over the course of the task varied very substantially from trial to trial. Thus, network simulations can yield trials where the localized network activity, or bump, triggered by the stimulus at $\theta_s$ is maintained by excitatory reverberation robustly through the delay. We call this type of trials memory bump trials (Fig. 2A). In other trials, before the cue stimulus is presented to the network a spontaneous bump may form in the network activity during the fixation period and remain stable for the duration of the delay period, despite the presentation of the stimulus at $\theta_s$. We will call these emergent bump trials (Fig. 2B). Alternatively, network activity formed earlier in the trial (whether or not in response to the stimulus at $\theta_s$) may fail to reverberate through the length of the delay period so that by the end of the delay network activity is unstructured and does not contain any robust signal. These trials are what we call decaying bump trials (Fig. 2C).

In addition, these network simulations produced behavioral output in our simulated task and gave rise to correct and error trials that could be treated similarly as in a real psychophysics experiment. Indeed, different network trials differed in the accuracy and in the confidence of “behavioral” responses [32]. For each simulation trial we obtained a “behavioral response” by extracting a population vector read-out from the network activity in a window of 50 ms at the end of the delay period. The population vector computes an estimation of the angle $\hat{\theta}$ encoded in the network activity by averaging together vectors pointing in each neuron's preferred stimulus $\theta_i$, weighed by their respective firing rate (see Methods). Experimentally, the population vector has been shown to provide a read-out of neural activity in agreement with motor output [36,37]. In addition, we derived a measure of confidence $C$ on a trial by trial basis from the same population vector analysis computed from network activity at the end of the delay (see Methods). This measure quantifies the quality of the selectivity in the neural network at that point in that trial. Thus, for each simulation trial we obtained the full network dynamics over the course of the trial, and two different behavioral measures: the decoded stimulus location $\hat{\theta}$, and the confidence in the response $C$.

To evaluate quantitatively the performance of our network simulations in the control case (parameters as in the “modified parameter set” of Compte et al. [19]) we ran 200 trials with different noise realizations so that each trial had a different network activity at the end of the delay. We analyzed our simulation results in two different ways: by considering the neural network activity through each trial, to relate to neuroimaging and
neurophysiology studies; or by limiting our analysis to behavioral measures, as in psychophysical experiments. The model network performed 91.5% correct (see Methods). Of these correct responses, 98% were confident responses and 2% were low-confidence responses. Correct low-confidence responses all corresponded to decaying bump trials. Correct high-confidence responses were all memory bump trials. In these network simulations, all incorrect trials were declared of low confidence. Low-confidence errors corresponded to decaying bump trials. In our control-case simulations, no errors of high-confidence and no emergent bump trials were observed. However, these types of trials and behavioral outcomes can become very prevalent upon parameter modification, as we will see next. Thus, the classification based purely on behavioral data (\( \hat{\theta} \) and C) matches relatively well the classification based on neural activity at the network level, as suggested also experimentally [32]. Low-confidence responses, whether erroneous or correct, are linked to decaying bump trials, since in those trials the network activity before the response does not provide information about \( \theta \) and the response is consciously a mere guess. On the other hand, a well-defined activity bump is available at the end of the delay for memory and emergent bump trials, so that behavioral responses in these two types of trials are declared confident. Erroneous high-confidence responses might therefore come from either a spurious memory arising spontaneously in the circuit (emergent bump trial or “false memory” trials as described by Lee et al. [32]) or from a stimulus-triggered memory that did not maintain sufficient accuracy (memory bump trial).

It has been reported that within confident responses to visuo-spatial delayed response tasks, schizophrenic patients have a larger proportion of errors than controls [32], suggesting that alterations of the mechanisms in prefrontal cortical circuits in schizophrenic patients may give rise to either more emergent bump trials or reduced accuracy in memory bump trials. In order to test this quantitatively in our model simulation, we introduced manipulations of the model’s synaptic parameters that have previously been associated with the pathophysiology of schizophrenia. We changed parameters in two different scenarios, following the GABAergic and glutamatergic hypotheses of schizophrenia, respectively (Fig. 3).

According to the GABAergic hypothesis, reduced GABA synthesis in neocortical interneurons, and especially in the DLPFC, contributes to the clinical features of schizophrenia, in particular cognitive symptoms [27]. We asked whether concerted reductions in inhibitory connection strengths, applied both onto pyramidal neurons (G_{IE})
and interneurons ($G_{II}$), induced performance changes in our network model consistent with the increased high-confidence errors in schizophrenic patients [32]. We found that for a GABA$_A$ receptor conductance reduction in all inhibitory connections ($G_{IE}$ and $G_{II}$, Fig. 3B) of 12.5%, overall performance in the task decreased by 35% relative to the control network, reflecting primarily an increase in high-confidence errors (Fig. 4). We analyzed the network simulation trials and we found that these increased high-confidence errors were mostly due to a higher incidence of emergent bump trials, rather than to inaccurate memory bump trials. We further studied how these behavioral effects depended parametrically on the reduction in inhibitory conductances $G_{IE}$ and $G_{II}$. We found that the proportion of high-confidence errors increased monotonously as the strength of inhibitory conductances in the network was further reduced (Fig. 5A).

The glutamatergic hypothesis posits that hypofunctional NMDA receptor transmission in the neocortex is partly responsible for schizophrenia symptoms [26]. We tried in our network model whether concerted reduction in NMDA receptor conductances ($G_{EE}$ and $G_{EI}$, Fig. 3C) resulted in performance changes consistent with the increase of high-confidence errors observed in schizophrenic patients by Lee et al. [32]. We found that equal percentage reduction (1%) in $G_{EE}$ and $G_{EI}$ systematically resulted in more behavioral errors (Fig. 4). However, most of this NMDA-dependent errors were low-confidence errors due to decaying bump trials (Fig. 5B), and not high-confidence errors as reported by Lee et al. [32]. Before discarding a possible role of hypofunctional NMDA receptors in explaining spatial working memory maintenance deficits in our network model, we tried different percentage changes for receptors onto pyramidal neurons and interneurons. We found that reducing NMDA receptor conductance more in the connections from pyramidal cells to interneurons ($G_{EI}$ 0.85% reduction) than in the connections from pyramidal cells to pyramidal cells ($G_{EE}$ no reduction) (Fig. 3C) could impair performance through the occurrence of more high-confidence errors, again caused primarily by more emergent bump trials and not so much by inaccurate memory bump trials (Fig. 5C). We therefore conclude that our network model for spatial working memory does support a possible causal role for hypofunctional NMDA in spatial working memory deficits in schizophrenia, provided the reduction in NMDA-dependent transmission affects primarily interneurons, rather than pyramidal neurons, in the prefrontal circuit responsible for spatial working memory maintenance.
Discussion

We have studied the effects of neurotransmitter system disruptions associated with schizophrenia in a computational network model developed to formulate circuit mechanisms behind persistent prefrontal neural activity in monkeys engaged in spatial working memory tasks. We have shown that “psychophysical” evaluation of network performance based on behavioral response and confidence in the response can distinguish on a trial-by-trial basis the nature of the underlying neural dynamics. This has started to be tested in neuroimaging studies [32]. We have applied synaptic transmission manipulations following the GABAergic or glutamateric hypotheses for schizophrenia and we constrained the resulting networks to mimic behavioral data from schizophrenic patients performing an analogous spatial working memory task [32]. Results were threefold: First, concerted reductions in GABA_A receptor conductances onto both excitatory and inhibitory neurons resulted in increased high-confidence errors, as in the experimental data, supporting a role of prefrontal GABAergic deficits in causing spatial working memory dysfunction in schizophrenia. Second, concerted reductions in NMDA receptor conductances onto both pyramidal neurons and interneurons caused an increase of low-confidence errors, contrary to experimental evidence. This indicates that a general decrease in NMDA receptor transmission cannot account for the cognitive deficits reported by Lee et al. [32]. Thirdly, an unbalanced reduction in NMDA receptor conductances primarily targeting NMDA channels on postsynaptic interneurons does generate network performance impairment through increased high-confidence errors. We conclude that both a specific reduction of NMDA receptor transmission in interneurons and a general reduction of GABA_A receptor activity converge in impairing spatial working memory function in our computational network model according to available behavioral data for schizophrenic patients [32]. This provides a specific mechanistic link between cellular and behavioral deficits associated with schizophrenia through the evaluation of both the accuracy and confidence of behavioral responses.

Variability in behavioral responses has usually been addressed by averaging over many realizations. However, there could be additional parameters that allow us to identify underlying causes of the variability that give us more power in analyzing the data. Here, we give computational support to a strategy advanced by Lee et al. [32] of categorizing behavioral trials according to the declared confidence in the response. Based on such model-derived behavioral data we could classify trials in 4 different classes: high-confidence correct, low-confidence correct, high-confidence error, and low-confidence
error. Neurally, however, we have access to the full network dynamics through each trial and we characterized the variability of neural network responses according to three major patterns: a stimulus-triggered memory that remains through the delay (memory bump trials), a spontaneous “false” memory that emerges before the stimulus and is not reset by it (emergent bump trials), and trials in which network activity fails to reverberate and decays by the end of the delay (decaying bump trials). We found a good correspondence between the former, behavioral classification and the latter, neural classification. All low-confidence trials were represented neurally by decaying bump trial properties. High-confidence correct trials were mostly sustained by a sustained memory bump. High-confidence error trials could be associated both to emergent bump trials (which was the major contributor in our network model) or to inaccurate memory bump trials, where the memory bump diffused away from the stimulus location by the end of the delay. Disambiguating these two types of neural dynamics is difficult based solely on this behavioral data. The unequivocal distinction would be the presence of strong circuit activation before the presentation of the cue stimulus in emergent bump trials, but not in memory bump trials. Alternatively, psychophysical testing using variable delay period lengths could determine if increased high-confidence errors in schizophrenic patients are mostly due to “false memories” (non-delay-dependent) or due to inaccurate memory representations (delay-dependent).

Our study indicates that hypofunctional NMDA receptors could have a role in the experimentally observed spatial working memory impairments, provided this deficit targets primarily the interneurons. Interestingly, independent lines of experimental evidence point at interneurons as the main target of hypofunctional NMDA transmission in schizophrenia. Firstly, systemically administered NMDA antagonists enhance, and do not depress, activity in cortical pyramidal neurons [38]. Secondly, in vitro studies reveal that cortical and hippocampal interneurons are particularly sensitive to NMDA antagonists, which induce net network disinhibition [39,40]. Finally, decreasing NMDAR signaling selectively in cortical and hippocampal interneurons in a conditional knockout mouse produces cellular and behavioral phenotypes typically associated with schizophrenia [41]. What this last study in addition proved is that a specific reduction of NMDA receptor signaling in interneurons results in changes of GABAergic transmission consistent with those supporting the GABAergic hypothesis, emphasizing the integrative explanatory power of the glutamatergic hypothesis [11,42]. Our computational model does not include intracellular mechanisms that could relate causally reductions in
NMDAR signaling in interneurons with GABA transmission deficits (possibly through NMDA-dependent calcium influxes [43]) but we found that both types of synaptic efficacy modulations can generate analogous “behavioral” deficits in the model network.

In this study we focused on establishing a mechanistic link between synaptic alterations in prefrontal circuits and behavioral deficits observed in spatial working memory tasks in schizophrenia. There are, however, also abnormal physiological traits of this disease that could conceivably be investigated using our computational network model. In particular, schizophrenia is also characterized by anomalous brain oscillations in the gamma frequency range. It has been argued that the glutamatergic hypothesis could also underlie changes in gamma cortical oscillatory dynamics, through the alteration of the excitatory inhibitory loop that has been associated with neural network oscillations [11,42]. Modeling studies have shown that inhibitory deficits may explain changes in sensory entrained gamma-frequency responses in schizophrenia [44] and that reduced excitatory drive onto interneurons suppresses network-generated gamma rhythms [45].

It remains to be tested how these synaptic alterations interact in a working memory network model to affect delay-dependent gamma-range oscillatory dynamics and cause behavioral deficits. Although our network model can generate gamma oscillations during sustained delay activity (adding AMPA-receptor-mediated transmission in recurrent connections [19]), we did not attempt to address this problem here and we defer this study for a future focused investigation.

We have studied here a spatial working memory computational model of a prefrontal circuit affected by synaptic alterations that have been associated with schizophrenia. This network model falls within the broad family of attractor-based network models for working memory, which have been studied in the context of schizophrenic symptoms before [23,25]. Our network aims to bring the general principles outlined in these earlier studies to the analysis of a very specific task, spatial working memory, and its behavioral outputs, high-confidence and low-confidence error responses. As shown before [24,25], we also find that synaptic alterations in this circuit cause behavioral deficits compatible with schizophrenia symptoms. However, we find that hypofunctional NMDA receptors can have very different behavioral outcomes (low-confidence vs. high-confidence erroneous responses) depending on the specificity of this deficit in interneurons. This complements the finding by Loh et al. [24] that NMDA-dependent deficits could be responsible for decreased stability of working memory (i.e. more decaying bump trials and more low-confidence errors, Fig. 5B). Our model shows that this occurs only for a balanced
reduction of NMDA-receptor-dependent transmission onto both pyramidal neurons and interneurons. Instead, reducing NMDA function primarily in interneurons (as supported by experimental data [38–41]) causes the spontaneous emergence of “false memories” (i.e. more emergent bump trials and more high-confidence errors, Fig. 5C).

We have not addressed in this manuscript the important effects that neuromodulatory systems may have in the performance of this network model, and the possible therapeutic effects of drugs acting on these systems. Others have studied how dopamine receptor activity affects other attractor network models for working memory [20,21,25]. It was found that dopamine neuromodulation via D1 receptors could promote resistance to distracting stimuli [20,21], while the activation of D2 dopamine receptors caused memory instability [25]. We previously tested synaptic manipulations associated with dopamine activity (NMDA and GABA\(_A\) receptor conductance enhancements) in our network model for spatial working memory and found that memory activity was rendered more stable, and the network could resist intervening distractors [19]. An explicit implementation of neuromodulatory systems in this network for spatial working memory is however still pending, which could shed light onto possible therapeutic targets to improve cognitive symptoms associated with spatial computations in schizophrenia.

**Figure Captions**

**Figure 1**

Schematic representation of the WM model.

Scheme of the ring structure of the network model with excitatory pyramidal neurons (black triangles) and inhibitory interneurons (gray circles) in a proportion of 4:1 interconnected within and between them. Nearby pyramidal neurons are strongly connected with each other (strength indicated by thickness of connections). Connections onto pyramidal neurons are indicated with a solid line and onto interneurons with a dashed line.

**Figure 2**
Different types of network behavior during the task.

Sample network activity (rastergram) during simulation trials of the visuospatial working memory task. Each dot represents an action potential from a pyramidal cell indexed by location y (labeled by the preferred cue $\theta_i$) at time x. Transient cue presentation of 250 ms (gray shaded area) at position $\theta_s = 0^\circ$ induces a tuned sustained memory state in the subsequent delay period. The task consists of an initial fixation period of 3s, the presentation of the stimulus (250 ms) and then a delay period of 3s. The population firing profile, averaged over the last 500 ms of the delay period is shown in the right panels.

A. Left, Example of a memory bump trial. The bump triggered by the stimulus is maintained until the end of the delay period close to the position where the cue was presented. Right, The example is a correct trial: the activity at the end of the delay (500 ms) is close to $\theta_s = 0^\circ$ ( $\hat{\theta} = 7^\circ$ , indicated by a black triangle), within the allowed range ($\theta_s \pm 22.5^\circ$, indicated by a gray bar). It is a high-confidence trial, C is above the limit (C = 0.82 > 0.5), indicating the presence of a well-defined activity bump.

B. Left, Example of an emergent bump trial. A spontaneous bump is formed before the cue presentation. Right, It is an error trial: the activity at the end of the trial is out of the allowed range ( $\hat{\theta} = 53^\circ$ ). C is above the limit (C = 0.76 > 0.5), so it is a high-confidence error trial.

C. Left, Example of a decaying bump trial. The bump triggered by the stimulus decays before the end of the delay period. Right, It is an error trial: the activity at the end of the trial is out of the allowed range ( $\hat{\theta} = -82^\circ$ ). And it is a low-confidence error trial, since C is below the limit (C = 0.30 < 0.5). This reflects the lack of tuning at the end of the delay.

Figure 3

Scheme of synaptic alterations used to model the pathophysiology of schizophrenia in the network model.

Synaptic connections between and within the pyramidal neuron population (black triangle, black lines are excitatory connections) and the interneuron population (gray circle, gray lines are inhibitory connections). Connections onto pyramidal neurons are indicated with a solid line and onto interneurons with a dashed line. Changes in synaptic
conductances are represented by a change in the width of connection lines in the scheme.

A. Initial synaptic conductances of the model network, simulating the healthy condition.

B. A “schizophrenic” network based on the GABAergic hypothesis is modeled by decreasing the strength in the $G_{\text{E}}$ and $G_{\text{II}}$ connections (thinner lines).

C. A “schizophrenic” network based on the Glutamatergic hypothesis is obtained by weakening connections $G_{\text{EE}}$ and $G_{\text{EI}}$, or $G_{\text{II}}$ alone.

**Figure 4**

Changes in the synapses following the hypotheses of schizophrenia increase the number of errors in the task.

Fraction of errors (in a total of 200 simulation trials) for four different conditions: healthy network; network with a decrease in GABA$_A$ receptor conductances (following the GABAergic hypothesis); network with a decrease in NMDA receptor conductances (same reduction in $G_{\text{EE}}$ and $G_{\text{EI}}$, glutamatergic hypothesis); network with a decrease in the NMDA receptor conductance onto interneurons $G_{\text{EI}}$ (following the glutamatergic hypothesis). The type of error is indicated by the color, the high-confidence error trials in black and in gray the low-confidence errors.

**Figure 5**

Parametric quantification of the fraction of correct and error trials for different amounts of synaptic changes.

Fraction of correct (left) and error (right) trials (over 200 simulation trials) for different values of the three synaptic modifications: reduction in GABA conductances (panels A), concerted reduction of NMDA conductances $G_{\text{EE}}$ and $G_{\text{EI}}$ (panels B), and reduction of $G_{\text{EI}}$ NMDA conductances alone (panels C). For error trials the type of error is specified, distinguishing high-confidence (black) and low-confidence errors (gray).

In all cases, correct trials decrease with increasing reduction in the corresponding conductances. For A and C, but not B, this is due to an increase in high-confidence errors as experimentally observed [32].
References


Figure 2
Figure 3
Figure 4
Figure 5