Title:

Structured connectivity exploits NMDA non-linearities to enable flexible encoding of multiple memoranda in a PFC circuit model

Authors:

Stefanou S. Stefanos[1,2], Athanasia Papoutsi[1], Panayiota Poirazi[1,3,*]

Affiliations:

[1]*Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion, Crete, 70013, Greece*
[2] *Department of Biology, University of Crete, Heraklion, Crete, 70013, Greece*

Author list footnotes:
[3]Lead Contact
*Correspondence: *poirazi@imbb.forth.gr*

Summary:

Prefrontal Cortex (PFC) exerts control on action selection and mediates behavioral flexibility during working memory (WM) task execution, when integration and retention of sensory information takes place. We used biophysical circuit modelling to investigate the dendritic, neuronal and circuit mechanisms that underlie these computations, aiming to causally link these three processing levels. Our simulations predict that dendritic NMDA non-linearities drive distinct activity dynamics of the same network, thus enabling adaptive coding in the absence of plasticity mechanisms. Specifically, we find that distinct assemblies of mixed-selectivity neurons emerge and fire in stable trajectories in a stimulus-dependent manner. Synaptic inputs that are spatio-temporally clustered, as provided by the structured connectivity of the PFC, facilitate these activity dynamics, thus further increasing the flexibility of the network. Our study suggests that behavioral flexibility may result from the formation of memoranda-specific assemblies in the PFC which are utilized dynamically in relation to the task at hand.

Keywords: prefrontal, working memory, dendritic integration, NMDA, behavioral flexibility, assemblies, dimensionality

Introduction

Working memory (WM) is the ability to maintain behaviorally relevant stimuli for a short period of time and to integrate them towards a choice/action. Seminal WM studies found that the activity of single neurons increases during the presentation of a to-be-maintained item (stimulus epoch) and persists throughout the blank 'post-stimulus' interval (post-stimulus epoch). This persistent activity constitutes a basic component of decision making by bridging the temporal gap between the item presentation and the subsequent contingent response (Funahashi et al., 1989; Fuster and Alexander, 1971) and is involved in the top-down modulation of behavior (Van Kerkoerle et al., 2017). The above findings form the backbone of the so-called fixed selectivity framework that has dominated the WM field. According to this framework, selective tuning of neurons or neuronal populations to the to-be-remembered information can be explained by the formulation of stable attractors: the fine-tuning of the respective network connections creates stable points, whereby individual neurons fire consistently to specific stimuli, both during the stimulus and the post-stimulus epoch (Renart et al., 2003; Wimmer et al., 2014).

Recent electrophysiological findings challenge this framework, based on the observation that the dynamics underlying WM in the PFC are more diverse than previously thought. First, individual neurons display time-varying discharge rates (Bernacchia et al., 2011; Brody et al., 2003; Jun et al., 2010; Shafi et al., 2007), leading to the hypothesis that the tuning properties of neurons change throughout the task duration. Second, individual PFC neurons can be selective for a broad range of task variables and stimuli, exhibiting thus a 'mixed selectivity' (Lindsay et al., 2017; Rigotti et al., 2013; Warden and Miller, 2010). Third, at the population level, the stimulus-induced network activity decays after the stimulus offset and emerges later with a different population activity pattern (Stokes et al., 2013), yet containing a stable representation of the to-

be-remembered item (Murray et al., 2017). These findings challenge the fixed-selectivity framework and call for a deeper mechanistic understanding of the cellular substrate of WM.

Single-neuron dynamic behavior and mixed selectivity has thus far been implemented using recurrent neural networks with random connectivity and non-biologically plausible training of either the weights of the readout neuron or of the entire network (Barak et al., 2013a, 2013b; Mante et al., 2013; Rigotti et al., 2010). Unlike the slow training required by these approaches and the resulting rigidity in responses, PFC circuits typically utilize short-term dynamic mechanisms that renders them highly flexible (Bouchacourt and Buschman, 2019; Fujisawa et al., 2008). In addition, the representational shift at the population level from the sensory towards a memory state has thus far been attributed to either the dynamic behavior of a randomly connected recurrent neuron network (Bouchacourt and Buschman, 2019) or to short-term facilitation mechanisms (Barak et al., 2010). Both approaches depend on the implementation of distinct populations of sensory and memory neurons. Yet, experiments show that selection and integration are implemented by the same PFC circuit, with both representations being present at the single neuron level (Mante et al., 2013; Warden and Miller, 2010). In this case, the shift in the population activity indicates integration of an incoming stimulus with the network's internal state, which depends on short-term dynamics (Buonomano and Maass, 2009).

At these short-term time scales, it remains an open question how a recurrently connected network utilizes a rapid mechanism that renders it able a) to perform both selection and integration b) to be highly dynamic at the single neuron level yet, c) stable and separable at the population level. Theoretical work suggests that the answer may lie in the highly non-random connectivity of pyramidal neurons, characterized by distance-dependent formation of neuronal clusters and over-represented structural 'motifs' (structured connectivity) (Perin et al., 2011;

Song et al., 2005), as this type of connectivity is optimal for storing a large number of attractor states in a robust fashion (Brunel, 2016). In the PFC in particular, layer 5 (L5) pyramidal neurons have almost double the reciprocal connections of other cortical regions, forming hyper-reciprocally connected circuits (Morishima and Kawaguchi, 2006; Morishima et al., 2011; Wang et al., 2006). This hyper-reciprocity could be a major player in the formation of multiple, transient, assemblies in the PFC. In addition to connectivity features, the NMDA receptor is considered a key determinant of stable attractor dynamics (Durstewitz et al., 2000; Wang, 1999). The non-linearity and slow kinetics of the NMDA receptor make it a complementary candidate to the short-term plasticity mechanisms, that can contribute to the activity-dependent selectivity and storage of the sensory input i.e. adaptive coding (Stokes et al., 2013). In support of this hypothesis, NMDA-dependent dendritic non-linearities have been shown to underlie the expression of memory states in computational models of small cell assemblies (Papoutsi et al., 2014, 2017).

In this work we investigated the role of NMDA-dependent dendritic nonlinearities and structured connectivity in the ability of a biophysical PFC network model to express multiple distinct memoranda. Towards this goal, we use a detailed biophysical network model of the PFC that is heavily constrained in terms of its connectivity and biophysical properties and show that it reproduces multiple types of experimental data. Overcoming many of the shortcomings of previous implementations, this model incorporates several recent findings of PFC function. Specifically, our model responds with single neuron variability comparable to what is experimentally reported (Bernacchia et al., 2011; Brody et al., 2003; Jun et al., 2010; Shafi et al., 2007). Moreover, at the population level, the model exhibits the representation shift between the stimulus and memory epochs, while maintaining robust memoranda over time (Murray et al.,

2017; Stokes et al., 2013). This is achieved without the need to artificially simulate different sensory and memory populations, rather utilizing a recurrent neural network demonstrating that stimulus selection and integration can occur in a single network (Mante et al., 2013; Warden and Miller, 2010). We predict that NMDA non-linearities serve as a complementary mechanism for adaptive coding in the PFC, supporting diverse activities of a single population of the PFC, ultimately to be utilized for different behaviors.

Results

A novel, biologically-constrained network model of the PFC

To investigate the cellular and connectivity mechanisms underlying the experimentally reported properties of WM, we developed a novel, biologically-constrained compartmental network model of the PFC. The network consisted of 250 biophysically detailed and morphologically simplified L5 PFC pyramidal neurons (PC) and 83 fast spiking interneurons (FS). Single neuron models were validated against experimental data with respect to both active and passive electrophysiological properties (Table S1). Basal dendrites of pyramidal neurons supported NMDA-dependent non-linear integration, as found experimentally (Figure 1A, Table S2). Non-random connectivity was implemented in congruence with available experimental data (Figure S1, Tables S3, 4). Specifically, the PC pair connection probability was set to be (a) distance-dependent and (b) local clustering dependent, based on the common neighbor rule (Perin et al., 2011; Song et al., 2005; Wang et al., 2006). PC-PC connections were created to reflect the high reciprocity ratio (47% of total PC-to-PC connections, given that there is a connection) found specifically in the PFC (Morishima et al., 2011; Wang et al., 2006) (Figure S1B-E). FS interneurons were connected to PCs in a sparse and distance-dependent manner (Packer and

Yuste, 2011), respecting the connectivity reciprocity ratio for FS to pyramidal projections in L5 of the frontal cortex (Otsuka and Kawaguchi, 2009) (Figure S1F-H). In addition, 66% of the FS interneurons were connected with each other as reported in (Galarreta and Hestrin, 1999). Concerning network creation, see Methods, sections "Network Connectivity" and "Synaptic connectivity properties".

The model reproduces single neuron and population dynamics.

To simulate WM in the network model, synaptic stimulation of 1-sec long, uncorrelated, Poisson-like spike trains was delivered to random subpopulations consisting of 50 PCs (20% of total, n=10 trials, Figure 1B). We focused on the emergence and dynamics of activity following the end of stimulus presentation (hereby termed post-stimulus epoch). Since both stimulated and non-stimulated neurons were recruited during the post-stimulus epoch with diverse spiking profiles (Figure 1C), we considered in our analysis all PCs and not just the stimulated ones. Trials with no PC activity during the last second of the simulation were not included in the analysis. Both inter-spike intervals (ISIs) and Coefficient of Variation (CV) were increased during the post-stimulus epoch ($ISI_{stimulus}$=40.3±53.8ms, $ISI_{post\text{-}stimulus}$= 52.9±115.3, p value=0.0004, Mann Whitney U test, $CV_{stimulus}$=0.42±0.25, $CV_{post\text{-}stimulus}$= 0.64±0.46, p value=0.0001, Mann Whitney U test, Figure 1D, 'control' cases), in line with experimental data (Compte et al., 2003; Stokes et al., 2013).

At the network level, activity was sparse (sparsity=0.89, values close to 1 correspond to sparser activity, see Methods, section "Sparsity", (Rolls, 1997)) and characterized by dynamic population responses; the firing of some neurons persisted, increased or fluctuated during the post-stimulus period (Figure 2A) (Bernacchia et al., 2011; Brody et al., 2003; Jun et al., 2010;

Shafi et al., 2007). To assess whether the model reproduces the experimentally reported shift from sensory to the memory representation(s) (Spaak et al., 2017), we measured the correlation coefficient of the binned network activity (bin size q=50ms) between different simulation time points. Figure 2B depicts the reduction in the correlation of the network activity during the stimulus and the post-stimulus epochs, denoting the representation shift. Within the stimulus epoch, population activity is consistently highly correlated, yet, during the post-stimulus epoch correlation is stronger over the diagonal time axis, indicating that population activity is not static. The population velocity (i.e. the speed with which the network activity changes, see Methods, section "Population multidimensional velocity") drops upon stimulus withdrawal and remains at stable values during the post-stimulus period (Figure 2C, top). Note that changes in the energy state (i.e. energy velocity, change in the average firing frequency of the network, see Methods, section "Population multidimensional velocity") are lower: they peak upon stimulus withdrawal and reach almost zero values during the post-stimulus period (Figure 2C, bottom). This indicates that the shift is not due to changes in the average firing rates per se, as seen also experimentally (Stokes et al., 2013). The above simulations replicate 1) the irregular spiking activity of single neurons reported in (Bernacchia et al., 2011; Brody et al., 2003; Jun et al., 2010; Meyers et al., 2008; Shafi et al., 2007) 2) the de-correlation of the network activity in the late post-stimulus epoch from the stimulus epoch reported in (Barak et al., 2010; Murray et al., 2017; Stokes et al., 2013) and 3) the change in the population activity trajectories over time (Stokes et al., 2013). Thus, our model network reproduces the complexity of single neuron and population responses during the post-stimulus epoch.

The PFC network model expresses multiple memoranda

The shift from the stimulus-driven network activity to a different, memory-related one is more prominently observed in PCA state space (Figure 3A). Network responses for different trials, whereby for each trial a randomly selected sub-population of PCs was stimulated with uncorrelated Poisson inputs (see Methods, section "Stimulation"), diverge from the stimulus epoch (green) and follow different trajectories in the PCA state space over time. Interestingly, during the late post-stimulus epochs (Figure 3A, darker colors, showing 9 trials with activity in post-stimulus epoch), activity for some trials appears to converge to specific trajectories in the PCA state space. We quantified the number of distinct memory trajectories (over the last second of the post-stimulus period) using k-means. Evaluation of the clustering results was done using the Bayesian Information Criterion (BIC, see Methods, section "Principal component analysis"). We found that for the specific instance (fixed connectivity, see Methods, section "Details of network connectivity") of the structured network shown in Figure 3, population activity converged to four different robust trajectories (BIC 213.87). These findings suggest that, additionally to the representational shift from the sensory to the memory epoch, a dynamic mechanism is recruited by specific stimuli in an otherwise fixed connectivity network, which allows its rapid re-configuration, in the absence of any learning mechanism, and amounts to the retention of different memoranda.

NMDA nonlinearities underlie multiple memoranda

We hypothesized that a thresholding mechanism acting in short time scales may drive the network activity into the distinct robust trajectories. The most prominent such candidate, in the absence of any short- or long- term plasticity, is the NMDA receptor. To test this hypothesis, we repeated the above simulations while eliminating the NMDA current (conductance set to zero).

To compensate for the reduction in excitability, we increased the AMPA conductance such that the saturated response matched the one when both NMDA and AMPA conductances were included (Figure S2A top, inset in Figure 3B). We further adjusted the network inhibition (see Methods, section "Synaptic manipulations") to ensure the induction of post-stimulus activity. We found that the AMPA-only model network was unable to express more than one robust trajectory (Figure 3B, BIC=213.2). Activity in the AMPA-only network was also accompanied by a decrease in sparsity (mean of AMPA-only sparsity=0.49) and CV (AMPA-only $CV_{post\text{-}stimulus}$= 0.28±0.21, p value=0.0001, comparison with control case, Mann Whitney U test) and an increase in firing frequency (AMPA-only $ISI_{post\text{-}stimulus}$=16.6±23.6, p value=0.0001, comparison with control case, Mann Whitney U test) (Figure S2B top, middle). At the population level, there was a reduction in the correlation of the network activity during the stimulus and the post-stimulus epochs as before (Figure S2B bottom, also shown as change in the PCA space, Figure 2B), yet within the stimulus epoch, population activity was stable. These simulations suggest that the NMDA receptor is critical for the expression of multiple robust trajectories and the irregular dynamics in the PFC model network.

Next, we asked which component of the NMDA conductance could be responsible for this finding: its non-linear features or its slow kinetics? The NMDA conductance can create two distinct dendritic phenotypes with increasing synaptic magnitude: a) a non-linear jump in the depolarization, due to the Mg^{+2} mediated non-linearity of the NMDA activation function and b) a depolarization plateau potential following the jump, supported by its slow kinetics (Doron et al., 2017). To assess the contribution of the non-linear jump, we removed the voltage-dependence of the NMDA conductance associated with the Mg^{+2} block, while keeping the slow kinetics and the plateau potential intact (see Figure 3C, Fig S2A bottom). Again, the network failed to support

more than one robust trajectory (Figure 3C, BIC=212.9). Note that in this case sparsity is maintained at control levels (sparsity=0.91), while the ISI/CV change significantly (Mg^{+2} block $CV_{post\text{-}stimulus}$= 0.55±0.33, p value=0.0001, comparison with control case, Mann Whitney U test, Mg^{+2} block $ISI_{post\text{-}stimulus}$= 50.1±91.8, p value=0.0001, comparison with control case, Mann Whitney U test) (Figure S2C top, middle). At the population level, the results were similar to the AMPA-only case (Figure S2C bottom). Manipulations of the kinetics of the NMDA receptor, so as to match those of the AMPA receptor, resulted in identical responses to the AMPA-only simulations (data not shown). Overall, these simulations suggest that while the depolarizing plateau aids in maintaining a sparse network response, it is the non-linear component of the NMDA receptors that serves as a thresholding mechanism for the emergence of multiple activity states.

Synaptic arrangement along the dendrite regulates the network re-configurability

The spatial arrangement of synaptic inputs is known to affect NMDA receptor opening, resulting in non-linear dendritic integration (Losonczy and Magee, 2006; Poirazi et al., 2003a, 2003b; Polsky et al., 2004). We thus further assessed how synaptic arrangement within the basal dendrites of our model PCs may affect the NMDA-depended generation of multiple trajectories. Specifically, we evaluated how many distinct trajectories can be supported when varying the dendritic locations of recurrent synapses (between PC pairs) and dendritic inhibitory inputs (n=10 different synaptic arrangements). These manipulations resulted in a total of 100 simulated conditions: 10 synaptic arrangements for a given connectivity matrix x 10 trials (note that, for a given trial, when changing the synaptic arrangement, the same PCs are always stimulated). Three examples of the population activity trajectories for different synaptic arrangements, given a

specific connectivity matrix and stimulus subpopulation per trial, are shown in Figure 4A-C. In line with our hypothesis, responses in the PCA space vary considerably when changing the synaptic arrangement, giving rise to several distinct activity trajectories. Note that these different trajectories emerge in response to stimulation of the same neurons in a given network (fixed connectivity matrix) with the exact same stimuli. The only difference between trials is the dendritic localization of the incoming inhibitory and recurrent excitatory connections, indicating that dendritic integration is a key player in the observed phenotype.

The above findings do not discriminate between the following alternatives: given two synaptic arrangements that support the same number of trajectories, are these identical or distinct in the PCA state space? To answer this question, we pooled all 100 responses (Figure 4D), and found that different synaptic arrangements induce qualitatively different trajectories. These findings suggest that for a given connectivity matrix and input, the arrangement of recurrent synapses enables the network to engage in different subsets of PCA state space, effectively expanding the flexibility of the network.

We investigated the generalization of our results by identifying the number of trajectories found for the 10 different trials, given n=4 different connectivity matrices and n=10 synaptic arrangements (thus amounting to 40 cases). We found that the probability of a network with a given connectivity and synapse arrangement to express more than a single, activity trajectory was 0.625 (sum of probabilities for $K^* > 1$, Figure 4E, blue).

Structured connectivity exploits NMDA nonlinearities to increase memory

Given the above results, we sought to identify whether structural connectivity can selectively exploit the NMDA non-linearities to induce more than one post-stimulus trajectory. To assess the

role of connectivity (both in terms of recurrent connections and the clustering coefficient), we repeated the same simulations on a uniformly connected random network i.e. with each PC pair independently connected with a fixed probability (4 random connectivity matrices, for 10 recurrent synapse arrangements and 10 trials, see Methods, sections "Network Connectivity" and "Synaptic connectivity properties"). Both the structured and the random network types exhibited the same average connection probability (~0.12), to ensure a fair comparison. In the random case, single-neuron properties were also different compared to the structured one (random $ISI_{post\text{-}stimulus}$= 34.7±77.6, p value=0.0001, compared to control case, Mann Whitney U test, random $CV_{post\text{-}stimulus}$= 0.56±0.42, p value=0.0001, compared to control case, Mann Whitney U test, Figure S2D top, middle). At the population level, there was a reduction in the correlation of the network activity during the stimulus and the post-stimulus epochs as before (Figure S2D bottom), yet within the stimulus epoch, population activity was more stable compared to the control case. Most importantly, and in the presence of the NMDA conductance, the random network produced fewer distinct trajectories compared to the structured one (Figure 4E). The probability of the random network expressing more than one memory trajectory was 0.375 (sum of probabilities for $K^* > 1$, Figure 4E, orange). These simulations indicate that the structured, highly reciprocal connectivity of the PFC better exploits the NMDA non-linearities to produce additional trajectories compared to a uniform, randomly connected network.

Distinct trajectories correspond to different, dynamically recruited neuronal assemblies

We showed that the same network model can encode multiple memoranda during the post-stimulus period, utilizing a dynamic thresholding mechanism. How is this mechanism changing the effective connectivity between neurons to create distinct activity patterns? Are these activity

patterns associated with specific neuronal assemblies and, if yes, is there a meaningful way that these assemblies are recruited during the post-stimulus epoch? Regarding the first question, we utilized the non-negative matrix factorization (NNMF) approach (Lee and Seung, 1999), which allows us to approximate any neuronal activity pattern over time as a linear combination of a "basis" set. This "basis" set consists of neuronal activity profiles, as opposed to PCA components which do not directly map onto specific neurons (Figure 5A). We applied the NNMF on the network responses for the same network connectivity matrix and synaptic arrangement as used for Figure 3A, for the different trials. Population activity was decomposed to the activities of different neuronal assemblies, whereby the contribution of each assembly has a different weight (i.e. assembly activity) (Figure 5A). Note that assembly recruitment can also be transient during the post-stimulus epoch (Figure 5B), yet each trial can be identified by its assembly composition. Importantly, to assess whether different combinations of assemblies map to different trajectories, we visualized the activity for different trials, color-coded by the trajectory type, where same trajectories appear to be represented by the same assembly combinations (Figure 5C). The engagement of the assemblies in a trajectory can change over time, yet their combination appears distinct between the different trajectories.

Concerning the way that these assemblies are recruited during the post-stimulus epoch, correlation analysis of the number of trajectories and the optimal number of assemblies (Methods, section "Non-Negative Matrix Factorization") for the structured network connectivity, shows interesting results. First, there is a positive correlation ($r=0.37$, $p=0.017$) between the optimal number of assemblies and the optimal number of distinct trajectories (Figure 5D). This reveals that trajectories stored in the network tend to be represented by different populations, rather than different firing patterns of the same population. Second, there is no systematic one-to-

one correspondence between the number of distinct trajectories and the number of assemblies. This is shown by the identification of more assemblies than trajectories (Figure 5D) and can be interpreted as having multiple assemblies encoding a single trajectory. The latter suggests the presence of assemblies that are reusable across different memoranda and therefore, consisted of neurons that display mixed selectivity. Summarizing, the above results indicate that a recurrent network with NMDA nonlinearities and a structured, highly reciprocal connectivity provide the conditions to have mixed selectivity neurons, instead of the typical randomly connected network. This mixed selectivity is ideal for linearly separating these representations down-stream in complex cognitive tasks, because it stipulates that memoranda representations in network state space are not linearly dependent (can consist of different neurons), thus can more easily be separated by a linear readout neuron (Lindsay et al., 2017; Rigotti et al., 2013). Moreover, these results generate a testable prediction that neurons in PFC, after a task acquisition through learning, form assemblies with mixed selectivity neurons, that can be rapidly utilized, either solo or in combinations in response to the available stimulus and task at hand.

DISCUSSION

It is not well understood how information entering the PFC can be utilized for flexible decision making: how does a single network represent and support distinct tasks in a rapid, flexible manner (Mante et al., 2013; Stokes et al., 2013; Warden and Miller, 2010)? Experimental evidence indicates that network connectivity can sustain changes in synaptic efficacies faster than classical plasticity, enabling it to rapidly switch between different activity states, driven by input (Fujisawa et al., 2008; Raposo et al., 2014; Stokes et al., 2013) or internal network representations (Saez et al., 2015). Here, we tested the hypothesis that activity switches are the

result of NMDA-dependent nonlinear dendritic integration, which is a candidate complementary mechanism to dynamically reconfigure the network connectivity (Arnsten et al., 2010; Fujisawa et al., 2008; Lisman et al., 1998; Stokes, 2015).

We used a state-of-the-art modeling approach that enabled us to dissect the essential mechanisms that drive these highly dynamic phenomena. We built and validated a reduced, compartmental PFC network model that reproduced a large number of experimental observations related to dynamic memory encoding. More specifically, we showed that single neuron responses in the model are irregular, yet together create a coherent low dimensional trajectory. Our findings are consistent with a new experimental view, whereby individual neurons exhibit persistent activity in a highly dynamic manner, while retaining stability and discriminability of WM (Duncan, 2001; Duncan and Miller, 2002; Lindsay et al., 2017; Meyers et al., 2008; Miller et al., 1996; Murray et al., 2017; Rigotti et al., 2013; Saez et al., 2015; Shafi et al., 2007; Stokes et al., 2013). In addition, network activity shifted from encoding of the stimulus, towards a low energy state that represented the memorandum for the specific task at hand, a feature that -to the best of our knowledge- is replicated for the first time, in a single, biologically constrained (and thus recurrent), biophysical network model.

We predict that the key subcellular mechanism responsible for inter-trial changes in dynamic connectivity is the NMDA receptor. Dependent on the spatial arrangement of incoming inputs, the nonlinear integration properties of NMDARs can rapidly shift synaptic efficacies and enable the network to differentially integrate information for short periods of time (Buzsáki, 2010; Stokes, 2015). This is based on several lines of evidence indicating that NMDA receptors are important for PFC function (Lisman et al., 1998; Papoutsi et al., 2014; Wang, 1999; Wang et al., 2013). Furthermore, maximal exploitation of NMDA non-linearities is achieved by connecting

subsets of neurons in a clustered, hyper-reciprocal manner, as evidenced in the PFC anatomy (Wang et al., 2006), and previously hypothesized to be beneficial for memory (Brunel, 2016). Although different synaptic (Fiebig and Lansner, 2017; Mongillo et al., 2008) and network (Bouchacourt and Buschman, 2019; Fujisawa et al., 2008; Goldman, 2009; Lindsay et al., 2017; Lundqvist et al., 2016; Rigotti et al., 2010) mechanisms for flexible encoding in local PFC networks were previously hypothesized they fail to holistically summarize the multitude of PFC findings. The key PFC mechanisms we report can create a variable yet discriminable PFC output resembling stable attractor points in a sub-space domain, avoiding the limitations of a linear attractor model (fine-tuned network structure, limited capacity, no flexibility in synaptic weight changes (Goldman, 2009)). Lastly, we point out these mechanisms as a major player in the formation of multiple, transient, assemblies in the PFC. Although neuronal assemblies are not something newly observed (Buzsáki, 2010; Dejean et al., 2016; Fujisawa and Buzsáki, 2011; Peyrache et al., 2009; Schaub et al., 2015), we present a mechanism for rapid assembly creation based on stimulus, with mixed selectivity as a PFC adaptive coding mechanism. Our study provides a mechanistic explanation as to *how* dendritic integration, mediated by the NMDA receptor, can be exploited by the structured connectivity profile of the PFC to provide memoranda representation, encoded stably in time, while maintaining neural variability.

Finally, experimental observations (Saez et al., 2015; Warden and Miller, 2010) point away from the belief that: single neuron = single variable (Sreenivasan et al., 2014), thus increasing the capacity of the network (Barak et al., 2013b; Fusi et al., 2016). The distributed encoding over many neurons helps scale memoranda in a single network, without inducing completely redundant/overlapped responses. In accordance, different (specialized) neuronal ensembles of similarly ‘mixed/tuned’ single neurons can be utilized, as reusable memory units (neuronal

ensemble compositionality). This compositionality offers important time-savings when acquiring new tasks, as the network connectivity does not need to be modified by slow, synaptic plasticity processes. It is sufficient to simply activate a new combination of already learned components, thus rapidly augmenting the computational power of the PFC, a region of great behavioral importance (Wang et al., 2019). Our work contributes towards this direction, showing that such a configurability can indeed be supported by a biophysical network model of the PFC, and that such a clustered configuration is indeed optimal in terms of WM storage, contrasted with a non-specific, random projection between neurons, that compromises the compositionality of the network (Barak et al., 2013b).

This work focuses on modelling the key biophysical features of PFC networks, including the extensive connectivity patterns found in experimental PFC works, the dendritic nonlinearity features of pyramidal neurons and the spatial localization of excitatory/inhibitory inputs. Therefore, we refrained from expanding our exploration in many of the free parameters of the model. Specifically, the network size was chosen as to have a minimum of active pyramidals during the post-stimulus epoch, while maintaining the experimentally reported synaptic properties. Simulating networks of greater size is computationally expensive and in our view, does not offer additional insights to this particular problem. To maintain tractability, we modeled pyramidal neurons with just one basal and two apical (proximal/distal) dendrites, validated to reproduce the key synaptic/dendritic nonlinearities in these structures. This reduction is the simplest form of pyramidal neurons that enabled the expression of dynamic phenomena at the network level. We thus expect that more inputs in different dendritic compartments could only extend the model's adaptive coding complexity (Poirazi et al., 2003b). Moreover, our model included only one inhibitory interneuron type, modeled as a simplified three-compartment node:

the soma/proximal dendrite-targeting fast-spiking interneurons. This simplification is clearly not representative of the complex inhibitory networks found in the PFC (Ascoli et al., 2008; Markram et al., 2004; Povysheva and Zaitsev, 2008; Rao et al., 2000; Wang et al., 2004; Zaitsev et al., 2005; Zhang et al., 2017), which also includes dendrite-targeting interneurons and ignores any dendritic nonlinearities that may also contribute to PFC processing (Tzilivaki et al., 2019). Our reasoning for this choice was to focus on excitatory determinants of PFC flexible encoding. Given our own findings about the role of complex interneurons in associative memory (Tzilivaki et al., 2019), we expect that a carefully constrained network model with more realistic interneuronal types will exhibit augmented adaptive coding. Finally, our model did not incorporate any plasticity mechanisms, in order to dissect the contribution of the NMDA nonlinear effects on depolarization from its effects on synaptic plasticity. We believe that our findings regarding the role of NMDA receptors in rapidly configurating network responses are complementary to the slower, plasticity shaped connectivity changes induced by learning. Overall, we are confident that our model is both biologically relevant and tractable, offering significant advantages in probing PFC function.

Methods

DETAILS OF MODEL NEURONS

The Pyramidal Neuron Model (PC)

The model of the pyramidal cells was adapted from (Papoutsi et al., 2013) (Model DB accession number: 155057). In brief, it consisted of five compartments: the soma, the axon, the basal dendrite, the proximal apical dendrite and the distal apical dendrite (apical tuft). Membrane capacitance on the dendrites was scaled x2 to account for spine membrane area. The somatic,

proximal and distal apical dendritic compartments included Hodgkin–Huxley type transient and persistent Na^+ currents, voltage-dependent K^+ currents (delayed-rectifier, A-type and D-type) , a fast and slow Ca^{2+} -dependent K^+ current, a hyperpolarization-activated non-specific cation current (h current), and the N-, R-, L- and T- type voltage-dependent Ca^{2+}currents (I_{caN}; I_{caR}; I_{caL}; I_{CaT}). The basal dendrite included transient and persistent Na^+ currents, a delayed K^+ rectifier current, an A-type K^+ current, a D-type K^+ current, a N-type Ca^{2+} current and a h current. The axon included a transient Na^+ current and a delayed rectifier K^+ current. Indicative electrophysiological properties of the PC model are shown in Table S1. For detailed description see (Papoutsi et al., 2013, 2014).

Fast-Spiking interneuron model (FS)

The FS model was adapted from (Konstantoudaki et al., 2014) (Model DB accession number: 168310). In brief, the FS consisted of a somatic, a dendritic and an axonic compartment. The somatic compartment included mechanisms for the slow A-Type K^+ current, the N-type high-threshold activated Ca^{2+} current and h current. Indicative electrophysiological properties of the FS model are shown in Table S1. For detailed description see (Konstantoudaki et al., 2014).

Synaptic Properties

The AMPA and NMDA currents of PCs generated non-linear NMDA-dependent responses at the basal dendrites as in (Nevian et al., 2007), in response to increasing number of activated synapses (2 events at 50Hz, Figure 1A). The AMPA and NMDA EPSC kinetics of the FSs were constrained against data from the respective L5 PFC interneurons (Wang and Gao, 2009), as per (Konstantoudaki et al., 2014). The $GABA_A$ conductance was calculated by exponentially

interpolating the reported values in (Kubota et al., 2015) (Figure S3). Finally, we implemented a slow inhibitory synaptic current ($GABA_B$), as in (Papoutsi et al., 2014) with x2 the conductance value as the $GABA_A$ (Table S2). FSs projected to FSs with a single $GABA_A$ synapse of 1.7nS as in (Galarreta and Hestrin, 1999). The respective synaptic parameters are listed in Table S2 and detailed description of the synaptic models have been published in (Konstantoudaki et al., 2014; Papoutsi et al., 2013).

DETAILS OF NETWORK CONNECTIVITY

Network Connectivity

The network model consisted of 250 PCs (~75%) and 83 FSs (~25%) (Beaulieu, 1993; Dombrowski et al., 2001). We implemented an experimentally-constrained connectivity by treating the biophysical network as a directed graph and its connectivity as a square binary connectivity matrix $\mathbf{A} \in \mathbb{R}^{(M+F)}$, where M is the number of PCs and F is the number of FSs. In this matrix, the rows represented neurons' outgoing connections (efferents) and the columns their incoming connections (afferents), with PCs preceding FSs (for connectivity matrix index $i, i \leq M$ corresponds to PCs and $M < i \leq M + F$ corresponds to FSs). Although this matrix contains the connectivity for the whole network, for brevity we will refer to the first M by M part of $\mathbf{A}$ as its *excitatory* part, containing the PC-PC connectivity. We implemented two different connectivity configurations between the PCs: one highly non-random (termed 'structured') and one 'random'. In the structured configuration a) PC pair connection probability was distance dependent, b) almost half connections were reciprocals, as reported in the PFC (Wang et al., 2006) and c) the PC pair connection probability was re-adjusted to be local-clustering dependent, based on experimental data (Perin et al., 2013; Song et al., 2005). In the random configuration, each pair was connected independently with a fixed probability. Both types of configurations exhibited the

same average probability of connection ($P_{average}$ ~0.12), as described below.

In the structured configuration, to implement the distance dependent connectivity constraints, the somata of PCs were regarded as points randomly scattered in a 180x180x180 μm^3 cubic space (Figure S1A), resulting in a somatic density of $6x10^4$ neurons/mm^3 (Schuz and Palm, 1989). Distances between these points were calculated by wrapping the cube dimensions in order to eliminate edge effects (Packer and Yuste, 2011; Perin et al., 2011). Each unordered PC pair $\{i, j\}$, could be connected in two ways: either unidirectionally ($i \rightarrow j$ or $j \rightarrow i$) or reciprocally ($i \leftrightarrow j$). These disjoint events followed the experimentally reported probabilities as a function of distance in L5 (Perin et al., 2011). In addition, and to reproduce the hyper-reciprocity reported in the PFC, we implemented a similar probability density function (pdf) for both reciprocal and unidirectional connections. The respective pdfs were described as:

$$g^{\text{overall}}(x) = 0.22\, e^{-0.0052\, x} \tag{1}$$

$$g^{\text{reciprocal}}(x) = 0.12\, e^{-0.0085\, x} \tag{2}$$

where $g^{\text{overall}}(\cdot)$ was the pdf of having any connection and $g^{\text{reciprocal}}(\cdot)$ was the pdf of having a reciprocal connection, both as a function of the distance x (Figure S1B). We derived the unidirectional pdf from their difference:

$$g^{\text{unidirectional}}(x) = g^{\text{overall}}(x) - g^{\text{reciprocal}}(x) \tag{3}$$

where $g^{\text{unidirectional}}(\cdot)$ was the pdf of having non reciprocal connection, as a function of the distance x (Figure S1B). Following, for each pair its reciprocal or unidirectional probability of connection, $P^{\text{reciprocal}}_{\{i,j\}}$ and $P^{\text{unidirectional}}_{\{i,j\}}$ respectively was:

$$P^{\text{unidirectional}}_{\{i,j\}} \equiv P((i \rightarrow j) \cup (j \rightarrow i)) - P((i \rightarrow j) \cap (j \rightarrow i)) = g^{\text{unidirectional}}(\, d(i,j)\,) \tag{4}$$

$$P^{\text{reciprocal}}_{\{i,j\}} \equiv P((i \rightarrow j) \cap (j \rightarrow i)) = g^{\text{reciprocal}}(\, d(i,j)) \tag{5}$$

where $d(i,j)$ is the pair's intersomatic distance x.

The excitatory part of the connectivity matrix **A** was created by rejection sampling for each pair. Specifically, we sampled from a uniform random distribution once for each pair and the excitatory part of the connectivity matrix **A** was determined as:

$$\begin{gathered}
a_{i,j} = 1, a_{j,i} = 1, \text{if } r_{\{i,j\}} < P_{\{i,j\}}^{\text{reciprocal}} \\
a_{i,j} = 1, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{\text{reciprocal}} \leq r_{\{i,j\}} < P_{\{i,j\}}^{\text{reciprocal}} + \frac{P_{\{i,j\}}^{\text{unidirectional}}}{2} \\
a_{i,j} = 0, a_{j,i} = 1, \text{if } P_{\{i,j\}}^{\text{reciprocal}} + \frac{P_{\{i,j\}}^{\text{unidirectional}}}{2} \leq r_{\{i,j\}} < P_{\{i,j\}}^{\text{reciprocal}} + P_{\{i,j\}}^{\text{unidirectional}} \\
a_{i,j} = 0, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{\text{reciprocal}} + P_{\{i,j\}}^{\text{unidirectional}} \leq r_{\{i,j\}} < 1
\end{gathered} \tag{6}$$

where $r_{\{i,j\}}$ is a random number on the interval [0,1] and $a_{i,j}$ the elements of **A** matrix. The resulting $P_{average}$ was ~0.12. The $P_{average}$ was calculated as the percentage of connected pairs over all pairs,

$$\begin{gathered}
\mathbf{U_{i,j}} = \begin{cases} 1, & for\ a_{i,j} + a_{j,i} \geq 1 \\ 0, & for\ a_{i,j} + a_{j,i} = 0 \end{cases} \\
P_{average} = \frac{\sum_{i \in \{1,\ldots,M\}, j \in \{1,\ldots,M\}, i \neq j} U_{i,j}}{M(M-1)}
\end{gathered} \tag{7}$$

To implement the local-clustering dependent constrains we used the common neighbor rule. According to this rule, pairs sharing increasing number of common neighbors are more probable to be connected (Perin et al., 2011, 2013). A common neighbor is defined as a third PC that receives from or projects to a PC pair. Starting from the above distance dependent excitatory part of connectivity matrix **A**, we iteratively recreated the connectivity to reproduce the common neighbor rule. Specifically, in each iteration we calculated the number of common neighbors for all pairs of the network, resulting in the symmetric matrix $\mathbf{C_N}$.

$$\mathbf{C_N} = \mathbf{E} \times \mathbf{E}^{T} \tag{8}$$

, where **E** was the undirected symmetric binary connectivity matrix of the graph of the excitatory part of **A**. Following, we normalized $\mathbf{C_N}$

$$\widehat{\mathbf{C_N}} = \frac{\mathbf{C_N}}{\|\mathbf{C_N}\|_{max}} \tag{9}$$

so that the pair of neurons sharing the maximal number of common neighbors is assigned a factor $\widehat{\mathbf{C_N}}_{i,j} = 1$ and pairs sharing no common neighbors a factor $\widehat{\mathbf{C_N}}_{i,j} = 0$ (Perin et al., 2013). The updated connection probability of a pair ($P^{cn}_{\{i,j\}}$) was its $P^{overall}_{\{i,j\}}$, multiplied with $\widehat{\mathbf{C_N}}_{i,j}$

$$P^{cn}_{\{i,j\}} = P^{overall}_{\{i,j\}}\, \widehat{\mathbf{C_N}}_{i,j} \tag{10}$$

Then, an updated overall connection probability is calculated from the $P^{cn}_{\{i,j\}}$ so that the $P_{average}$ remains constant:

$$P^{updated\ overall}_{\{i,j\}} = P^{cn}_{\{i,j\}} \frac{\sum_{i\in\{1,\dots,M\},j\in\{1,\dots,M\},i\neq j} P^{overall}_{\{i,j\}}}{\sum_{i\in\{1,\dots,M\},j\in\{1,\dots,M\},i\neq j} P^{cn}_{\{i,j\}}} \tag{11}$$

This updated overall connection probability was used to calculate the unidirectional and reciprocal connection probabilities respectively. Towards this, we calculated the distance estimate $d'(i,j)$, that corresponded to the $P^{updated\ overall}_{\{i,j\}}$ for each pair. The unidirectional and the reciprocal connection probabilities were then calculated using this $d'(i,j)$. The excitatory part of **A**, was recreated with rejection sampling as indicated above.

$$P^{unidirectional}_{\{i,j\}} \equiv P((i \to j) \cup (j \to i)) - P((i \to j) \cap (j \to i)) = g^{unidirectional}(\, d'(i,j)\,) \tag{12}$$

$$P^{reciprocal}_{\{i,j\}} \equiv P((i \to j) \cap (j \to i)) = g^{reciprocal}\Big(\, d'(i,j)\Big) \tag{13}$$

This algorithm ended after 1,000 iterations, where the average clustering coefficient plateaus as in (Perin et al., 2011) (Figure S1C). The clustering coefficient (Watts and Strogatz, 1998) for a single PC was defined as

$$C_m = \frac{k_m^{available}}{0.5k_m(k_m - 1)}, m = 1, \dots, M \tag{14}$$

where k_m is the number of neighbors (undirected edges to other neurons that neuron m has) and $k_m^{available}$ are the undirected edges that exist between these neighbors. The above procedure

resulted in the connections shown in Figure S1D,E.

In the case of the randomly connected network, pyramidal neurons were connected with the same independent probability for each pair, irrespective of distance, computed so that the two networks (structured and random) have the same $P_{average}$, as defined above. The respective connectivity matrix for the random configuration was created by rejection sampling as indicated above.

We simulated a total of n=4 different network instances (4 structured and 4 random), characterized by their respective connectivity matrix, retaining the interneuron connectivity unchanged (see below). Because this excitatory part of **A** contains only PC connectivity, we changed that without affecting the FS-PC/FS-FS connection probabilities and ratios. These connectivity instances were then translated into the synaptic connectivity of the detailed biophysical models.

For both the structured and the random cases, connectivity between FSs and PCs were distance dependent, as reported experimentally (Packer and Yuste, 2011). Due to the lack of data for PFC regarding the distance-dependent connectivity of L5 FSs with nearby PCs, we used the respective pdf between L5 somatosensory neurons (Packer and Yuste, 2011). Connectivity ratios are summarized in Table S3. FS-to-PC connections were included in $\mathbf{A_{i,j}}, \mathrm{M} < \mathrm{i} \leq \mathrm{M} + \mathrm{F}, \mathrm{j} \leq \mathrm{M}$ and PC-to-FS connections in $\mathbf{A_{i,j}}, \mathrm{i} \leq \mathrm{M}, \mathrm{M} < \mathrm{j} \leq \mathrm{M} + \mathrm{F}$. Because the overall connection probability reported was conflicting (0.25 in (Otsuka and Kawaguchi, 2009) and 0.67 in (Packer and Yuste, 2011)), we averaged the overall FS-PC connection probability to 0.46. The FS-PC overall, distance dependent connection pdf was:

$$\mathrm{g}^{\mathrm{FS-PCoverall}}(\mathrm{x}) = 0.46\ \mathrm{e}^{-\mathrm{x}/180} \tag{15}$$

similar to what is reported for somatosensory L5 (Packer and Yuste, 2011). The reciprocal and unidirectional pdfs were based on the FS-PC overall pdf:

$$g^{FS-PCreciprocal}(x) = 0.616\, g^{FS-PCoverall}(x) \tag{16}$$

$$g^{PC-FS}(x) = 0.076\, g^{FS-PCoverall}(x) \tag{17}$$

$$g^{FS-PC}(x) = 0.308\, g^{FS-PCoverall}(x) \tag{18}$$

, respecting the percentages reported in (Otsuka and Kawaguchi, 2009) (Figure S1F,G,H). The PC-FS connections were added to the connectivity matrix **A** by rejection sampling for each pair, as previously, where the pair $\{i, j\}$ now refers to FS-PC neurons ($M * F$ pairs in total). Specifically, we sampled from a uniform random distribution once for each pair and the respective elements of the connectivity matrix, $a_{i,j}$ were determined as:

$$\begin{gathered} a_{i,j} = 1, a_{j,i} = 1, \text{if } r_{\{i,j\}} < P_{\{i,j\}}^{FS-PCreciprocal} \\ a_{i,j} = 1, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{FS-PCreciprocal} \leq r_{\{i,j\}} < P_{\{i,j\}}^{FS-PCreciprocal} + P_{\{i,j\}}^{FS-PC} \\ a_{i,j} = 0, a_{j,i} = 1, \text{if } P_{\{i,j\}}^{FS-PCreciprocal} + P_{\{i,j\}}^{FS-PC} \leq r_{\{i,j\}} < P_{\{i,j\}}^{FS-PCreciprocal} + P_{\{i,j\}}^{FS-PC} + P_{\{i,j\}}^{PC-FS} \\ a_{i,j} = 0, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{FS-PCreciprocal} + P_{\{i,j\}}^{FS-PC} + P_{\{i,j\}}^{PC-FS} \leq r_{\{i,j\}} < 1 \end{gathered} \tag{19}$$

where $r_{\{i,j\}}$ is a random number on the interval [0,1] and $a_{i,j}$ the elements of **A** matrix (Figure S1F,G,H).

Finally, for the connectivity between FSs-FSs was not distance dependent. We set the connection probabilities $P_{\{i,j\}}^{FS-FSreciprocal} = 0.045$ and $P_{\{i,j\}}^{FS-FSunidirectional} = 0.135$, resulting in 0.18 overall connection probability as reported in (Galarreta and Hestrin, 1999) (Table S3). FSs-FSs connections were added to **A** by rejection sampling, where the pair $\{i, j\}$ now refers to FSs-FSs, with $0.5F(F-1)$ pairs in total:

$$
\begin{gathered}
a_{i,j} = 1, a_{j,i} = 1, \text{if } r_{\{i,j\}} < P_{\{i,j\}}^{FS-FSreciprocal} \\
a_{i,j} = 1, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{FS-FSreciprocal} \leq r_{\{i,j\}} < P_{\{i,j\}}^{FS-FSreciprocal} + \frac{P_{\{i,j\}}^{FS-FSunidirectional}}{2} \\
a_{i,j} = 0, a_{j,i} = 1, \text{if } P_{\{i,j\}}^{FS-FSreciprocal} + \frac{P_{\{i,j\}}^{FS-FSunidirectional}}{2} \leq r_{\{i,j\}} < P_{\{i,j\}}^{FS-FSreciprocal} + P_{\{i,j\}}^{FS-FSunidirectional} \\
a_{i,j} = 0, a_{j,i} = 0, \text{if } P_{\{i,j\}}^{FS-FSreciprocal} + P_{\{i,j\}}^{FS-FSunidirectional} \leq r_{\{i,j\}} < 1
\end{gathered}
\tag{20}
$$

where $r_{\{i,j\}}$ is a random number on the interval [0,1] and $a_{i,j}$ the elements of **A** matrix. FSs-FSs connectivity elements of **A**, lie on its last F-by-F part with index $i, M < i \leq M + F$ and $j, M < j \leq M + F$.

Synaptic connectivity properties

The number and percentage of synapses per type (inhibitory/excitatory) of connection are summarized in Table S4. Single neuron connectivity properties, including the location and the number of synaptic contacts, as well as the latencies between pairs of neurons, were based on anatomical and electrophysiological data. The recurrent connections between PCs were located at the basal dendrite via 5 synapses (Markram et al., 1997) and the synaptic latencies were drawn from a Gaussian distribution with $\mu = 1.7$ms and $\sigma = 0.9$ms (Thomson and Lamy, 2007). The precise location of the connecting synapses was random with uniform probability across the basal dendrite. These locations were constant across trials, but reshuffled when the network model underwent a synaptic arrangement. PCs connected to FSs via 2 synaptic contacts (Buhl et al., 1997), with the corresponding latencies drawn from a Gaussian distribution with $\mu = 0.6$ms and $\sigma = 0.2$ms (Angulo et al., 1999). FSs project to PCs both to the soma and the basal dendrites (Kubota et al., 2015). Their location was varied randomly with uniform probability between the synaptic arrangements. FSs projected to FSs with a single GABAa synapse as in (Galarreta and

Hestrin, 1999) with synaptic latencies drawn from a Gaussian distribution with $\mu = 1.76$ms and $\sigma = 0.07$ms (Bacci et al., 2003).

NETWORK STIMULATION AND MANIPULATION

Stimulation

In each trial we stimulated 50 randomly selected (uniform probability) PCs with uncorrelated Poisson inputs (λ=60Hz) for 5sec. The incoming stimulus (40 AMPA and 40 NMDA excitatory synapses), was provided to the proximal apical dendrites of the selected PCs (Kuroda et al., 1998). For the 4 different excitatory parts of the **A** (n=4), 10 synaptic arrangements (defined as randomly positioning the excitatory recurrent PC-PC connections and inhibitory connections along the basal dendrite) were implemented, thus resulting in a total of 40 cases. For each of these cases, we stimulated a total of T=10 times (trial, h), whereby for each trial the 50 stimulated PCs and the stimulus timings were varied. We did not include in the analysis the trials with no activity from any pyramidal neuron in the last second of the post-stimulus epoch.

Synaptic manipulations

In order for the network to produce post-stimulus epoch activity we scaled x1.75 the recurrent excitation (g_{NMDA} and g_{AMPA}) and x3 the feedback inhibition (g_{GABAA} and g_{GABAB}) This was done as the small size of the network limits the number of inputs a single PC received using realistic connectivity probabilities. To compensate for changes in excitability in simulations where g_{NMDA} conductance was set to zero, the g_{AMPA} was further augmented x50. This resulted in the same saturated depolarization when increasing the number of pair-pulse inputs at 50Hz (Figure S2A). Feedback inhibition (g_{GABAa} and g_{GABAb}) was scaled to x0.3 so that to allow for post-stimulus activity.

We further manipulated the properties of the NMDA receptors. NMDA current was simulated as in (Polsky et al., 2009):

$$I_{NMDA} = g_{NMDA}(V - E_{NMDA}) \tag{21}$$

$$I_{Ca^{+2}} = \frac{I_{NMDA}}{10} \tag{22}$$

$$g_{NMDA} = \frac{A - B}{1 + \eta e^{-\gamma V}} \tag{23}$$

$$\frac{dA}{dt} = \frac{-A}{\tau_a} \tag{24}$$

$$\frac{dB}{dt} = \frac{-B}{\tau_b} \tag{25}$$

, where $E_{NMDA} = 0\text{mV}$, $\tau_a = 90$ ms, $\tau_b = 5$ ms, $\gamma = 0.08\ \text{mV}^{-1}$, $\eta = 0.25\text{mM}^{-1}$.

In order to remove the non-linear 'jump' of the NMDA receptors, we removed the Mg^{2+} blockage in (24):

$$g_{NMDA} = (A - B) \tag{26}$$

resulting in the saturating depolarization response of Figure S2A (bottom). No other compensations of inhibition were needed in this case to observe post-stimulus epoch activity. Finally, in simulations with the random network connectivity we further adjusted the inhibition to x0.5, in order for the network to produce post-stimulus activity.

ANALYSIS

Network states

We binned PC activity (number of spikes) using discrete temporal windows of length q (q=50ms non-overlapping windows). The instantaneous network state (INS) was defined as the binned activity of each PC in given temporal window, $\mathbf{u} \in \mathbb{R}^M$. This resulted in $N = \frac{\text{simulation time (ms)}}{q}$

INSs. For a given trial h, we concatenated the N INSs, resulting in a matrix $\mathbf{B_h} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N], \in \mathbb{R}^{M\times N}$, that is the population activity of the trial. We analyzed the population activity across the trials $h = 1, \dots, T$, thus resulting in a total of $L = TxN$ INSs. The matrix contained the $\mathbf{B_h}$across trials, $\mathbf{D} = [\mathbf{B_1}, \mathbf{B_2}, \dots, \mathbf{B_T}], \in \mathbb{R}^{M\times L}$. The latter is referred to as **network state matrix** and characterized the activity of each network instance (same network connectivity and synaptic arrangement, different trials).

Cross correlation

Cross correlation between INSs was calculated using the Pearson correlation

$$\mathbf{R}_{ij} = \frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{S}_{ii}\mathbf{S}_{jj}}}, \forall i, j \in N \tag{27}$$

where $\mathbf{S}_{ij}$ is the covariance of $\mathbf{u_i}$with $\mathbf{u_j}$.

Sparsity

To assess the sparseness of the network response we used the metric proposed by (Rolls, 1997) subtracted from unity, so that higher values correspond to more sparse activity:

$$S_{TR} = 1 - \frac{\left(\sum_{m=1}^{M} \frac{\hat{r}_m}{M}\right)^2}{\sum_{m=1}^{M} \frac{\hat{r}_m^2}{M}} \tag{28}$$

, where $\hat{r}_m$ is the average firing rate of the *m-th* PC.

Population multidimensional velocity

Network multidimensional velocity $\mathbf{v_{MD}}$ was calculated as the gradient of the activity with respect to time:

$$\mathbf{v_{MD}} = \frac{d(\mathbf{u}_n, \mathbf{u}_{n+1})}{q}, n = 1, \dots, N-1 \tag{29}$$

Where $d(\cdot)$ is the Euclidean distance in $\mathbb{R}^M$ of the two adjacent in time INSs, converted in Hz. The per trial average was used. We further evaluated the change in the average firing frequency of the network using the energy velocity $\mathbf{v_E}$, i.e., the change in the overall energy level, across the duration of one trial.

$$\mathbf{v_E} = \frac{1}{M}\sum_{m=1}^{M} r_m^{n+1} - \frac{1}{M}\sum_{m=1}^{M} r_m^{n}, n = 1, \dots, N-1 \quad (30)$$

, where $\hat{r}_m$ is the firing rate of the *m-th* PC in q.

Principal component analysis

In order to examine the neuronal dynamics in a lower dimensional space, the network state matrix $\mathbf{D}$ was transformed to $\mathbf{F}$, $\mathbf{F} \in \mathbb{R}^{M \times L}$ using PCA. Following we split the L principal component loadings to the respective trials, representing the *network state space trajectories* per trial. The M_{PCA} principal components with eigenvalues above unity were used for analysis. For visualization, we used the first three principal components.

We evaluated the optimal number of discrete state space trajectories by applying the mean-shift algoritm and k-means, and evaluating the output using the Bayesian Information Criterion (BIC) (Wit et al., 2012). Towards this, from the network state space trajectories (matrix $\mathbf{F}$) we kept the last second of each trial ($G = \frac{1\text{sec}}{q}$) in the matrix $\mathbf{L_h} \in \mathbb{R}^{M_{PCA} \times G}, h = 1, \dots, T$. Each trajectory was smoothed prior to clustering with a Savitzky–Golay filter with window size 31 and of polynomial order 3.

We initialized the centroids of k-means using the mean-shift algorithm over all trajectory data and across all trials. We used different Gaussian kernels (K_{RBF}) for each trial (thus amounting total to T) characterized by their mean and variance. We used as the mean the average across time of $\mathbf{L_h}, \mathbf{v_h} \in \mathbb{R}^{M_{PCA}}$.

$$\mathbf{v_h} = \frac{1}{G}\sum_{j=1}^{G}(\mathbf{L_h})_{\mathbf{i,j}}, \forall i = 1, \dots, M_{PCA} \tag{31}$$

and the variance was set equal to the average variance of $\mathbf{L_h}$across all trials. Thus, the K_{RBF} was defined as:

$$K_{RBF}(\mathbf{l}_j, \mathbf{v_h}) = e^{\left(-\frac{\|\mathbf{l}_j - \mathbf{v_h}\|^2}{2\sigma^2}\right)}, j = 1, \dots, G \tag{32}$$

, with $\mathbf{l_j}$ being the *j-th* column (trajectory point) of $\mathbf{L_h}$**.** We defined the neighborhood of $\mathbf{v_h}$ as:

$$Z = \{l_j |\ K_{RBF}(\mathbf{l}_j, \mathbf{v_h}) \leq 0.1\} \tag{33}$$

The mean-shift evolves the initial trajectory point $\mathbf{v_h}$iteratively and for each iteration the weighted mean was determined as:

$$\bar{\boldsymbol{\mu}}(\mathbf{v_h}) = \frac{\sum_{\mathbf{l_j} \in Z} K_{RBF}(\mathbf{l}_j, \mathbf{v_h})\mathbf{l}_j}{\sum_{\mathbf{l_j}_j \in Z} K_{RBF}(\mathbf{l}_j, \mathbf{v_h})} \tag{34}$$

The resulting centroids $\bar{\boldsymbol{\mu}}(\mathbf{v_h})$ were used as the new mean of the K_{RBF} $\mathbf{v_h} \coloneqq \bar{\boldsymbol{\mu}}(\mathbf{v_h})$ in the next iteration. The iteration continued until centers converge to non-moving positions (translation less than $1/10$ of the variance). This process was repeated for all trials h. Following, we assumed K as the optimal cluster number. The T mean-shift centroids were taken in pairs merging those with minimal Euclidean distance, until only K centroids remained ($\boldsymbol{\mu_k}, k = 1, \dots, K,$ with $1 \leq K \leq T$). We finally labeled the average trajectories across time of the trials ($\mathbf{x_h}$) to the K clusters. Each average trajectory was assigned to a cluster using the minimum Euclidean distance between $\mathbf{x_h}$ and the centroids $\boldsymbol{\mu}_k$, as previously identified by the mean-shift algorithm. Following labels were re-arranged iteratively so that to minimize the intra-cluster Mahalanobis distance:

$$D_M(\mathbf{x}_h) = \sqrt{(\mathbf{x_h} - \boldsymbol{\mu_k})^T\, \mathbf{S}_k^{-1}(\mathbf{x_h} - \boldsymbol{\mu_k})} \tag{35}$$

where $\mathbf{S}_k^{-1}$ is the inverse of the covariance matrix of the data of cluster k.

Clustering evaluation was done with the BIC:

$$\mathrm{BIC}(\mathrm{K}) = \ln(\mathrm{T})\mathrm{K} - 2\hat{\mathcal{L}} \tag{36}$$

where $\hat{\mathcal{L}}$ is the value of the maximized log-likelihood function across all T data points (one for each trial), clustered in K clusters:

$$\mathcal{L}_{\mathrm{h}} = \frac{e^{\frac{-D_M(\mathbf{x_h})}{2}}}{\sqrt{(2\pi)^{M_{PCA}} \det(\mathbf{S})}}$$

$$\hat{\mathcal{L}} = \sum_{\mathrm{h}=1}^{\mathrm{T}} \ln\left(\mathcal{L}_{\mathrm{h}}\right) \tag{37}$$

where $D_M(\mathbf{x_h})$ is the Mahalanobis distance of the average trajectory $\mathbf{x_h}$ from its cluster centroid and $\mathbf{S}$ the cluster centroid covariance matrix. We chose the optimal model the one with the smaller BIC value.

$$\mathrm{K}^* = \arg\min_{\mathrm{K}=1\ldots\mathrm{T}} \mathrm{BIC}(\mathrm{K}) \tag{38}$$

In order to avoid over fitting issues (each trial activity, clustered differently) we set an upper bound on K^*, $\mathrm{K}^*_{\max}$ by setting a lower bound on the inter-cluster distance:

$$\mathrm{K}^*_{\max} = \arg\min_{\mathrm{K}=1\ldots\mathrm{T}} \left(d_{\mathrm{threshold}} < \min_{k,k' \in \{1\ldots\mathrm{K}\}} \left(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|_{D_M} \right) \right) \tag{39}$$

where we set the Mahalanobis distance threshold to $d_{\mathrm{threshold}} = 3$ between cluster centroids $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_{k'}$.

Non-Negative Matrix Factorization

To identify the change of effective connectivity between neurons, we used nonnegative matrix factorization (NNMF) as in (Lee and Seung, 1999) on the network instance state matrix **D**. This method utilizes an approximate factorization of the form:

$$\mathbf{D} \approx \mathbf{WH} \tag{40}$$

were $\mathbf{W} \in \mathbb{R}^{M \times K_{assemblies}}$ and $\mathbf{H} \in \mathbb{R}^{K_{assemblies} \times L}$ matrices. The matrix **W** corresponds to $K_{assemblies}$ basis patterns (columns of **W**, neuronal assemblies, Figure 5A) and the rows of the matrix **H** contain the coefficients (activity) of each of the assemblies over time.

Since the number of assemblies is dictated by the specific choice of $K_{assemblies}$, we evaluated the optimal number $K^*_{\boldsymbol{assemblies}}$ using K-fold Cross-Validation (for a total of 20 $K^*_{assemblies}$). **D** was randomly partitioned into $K_{CV} = 20$ equally sized subsamples. Of these, a single subsample was retained as the testing data set and the remaining $K_{CV} - 1$ subsamples were used as the training data set for each K-fold iteration. To partition the matrix, we defined a binary mask **O**, with each of its elements indicating a left out data point $\mathbf{O}^k_{i,j} = 0$ (testing data), or a data point in the training data set $\mathbf{O}^k_{i,j} = 1$, for K-fold iteration k=1,…, K_{CV}. This matrix changed for each K-fold Cross-Validation iteration, as predefined by the initial random partition, reflecting the different testing and training datasets. Following, we calculated the $\mathbf{W}, \mathbf{H}$ for each $K_{assemblies}$. We estimated the NNMF factor matrices $\mathbf{W}, \mathbf{H}$ with the multiplicative update rules below as in (Lee and Seung, 2001):

Algorithm 1: NNMF with missing elements

1: Initialize W, H randomly

2: **for** t = 1 to 1000 **do**

3: $W_{ij} \leftarrow W_{ij} \frac{((O \circ D)H^T)_{ij}}{((O \circ WH)H^T)_{ij}}$

4: $H_{ij} \leftarrow H_{ij} \frac{(W^T(O \circ D))_{ij}}{(W^T(O \circ WH))_{ij}}$

5: **end for**

where $\circ$ denotes the Hadamard product.

For each of the K-fold iteration, we calculated the error of the model using the testing data set in each K-fold iteration k

$$\mathrm{D}_{\mathrm{NNMF}}^{\mathrm{k}} = \frac{\|(\mathbf{1} - \mathbf{O_k}) \circ (\mathbf{WH} - \mathbf{D})\|_{\mathrm{F}}}{\mathrm{J}} \tag{41}$$

where $(\mathbf{1} - \mathbf{O_k})$ selects the testing data and $\mathbf{W}, \mathbf{H}$ are the NNMF factor matrices of the training data. The error is normalized with J that is the number of zero elements of $\mathbf{O_k}$ when calculated for the testing data. Calculating this error for different number of assemblies, creates a 'bias-variance' type of error curve, with its minimum indicating the optimal number of assemblies that **D** can be divided to. Due to the random partitioning, the cross-validation procedure was repeated 50 times for the different $\mathrm{K}_{\mathrm{assemblies}}$. We defined the optimal number of assemblies as the value of $\mathrm{K}_{\mathrm{assemblies}}^{*}$ where the average $\mathrm{D}_{\mathrm{NNMF}}^{\mathrm{k}}$ was minimum:

$$\mathrm{K}_{\mathrm{assemblies}}^{*} = \arg \min_{\mathrm{k}=1 \ldots \mathrm{K}_{\mathrm{assemblies}}} \left[\frac{1}{50} \sum_{n=1}^{50} \left[\frac{1}{\mathrm{K}_{\mathrm{CV}}} \sum_{k=1}^{\mathrm{K}_{\mathrm{CV}}} \mathrm{D}_{\mathrm{NNMF}}^{\mathrm{k}} \right] \right] \tag{42}$$

DATA AND SOFTWARE AVAILABILITY

The microcircuit model was implemented in the NEURON simulation environment (version 7.4) (Hines and Carnevale, 2001). The parallel simulations were executed on a high-performance computing cluster with 312 cores, under CentOS 6.8 environment. Analysis software was written in the Python (version 3.6) programming language. Implementations of algorithms used to compute quantities presented in this study are available at:

https://github.com/stamatiad/prefrontal_analysis.

Acknowledgements

The authors would like to thank Dr. Spyros Chavlis.

Author Contributions

Conceptualization, S.S.S., P.A., P.P.; Data curation, S.S.S., P.A.; Formal analysis, S.S.S.; Funding acquisition, P.P.; Investigation, S.S.S.; Methodology, S.S.S.; Project administration, P.P.; Resources, P.P.; Software, S.S.S; Supervision, P.A., P.P.; Validation, S.S.S.; Visualization, P.A., P.P.; Writing – original draft, S.S.S, P.A., P.P.

Declaration of Interests

The authors declare no competing interests.

References

Angulo, M.C., Rossier, J., and Audinat, E. (1999). Postsynaptic glutamate receptors and integrative properties of fast-spiking interneurons in the rat neocortex. J. Neurophysiol. *82*, 1295–1302.

Arnsten, A.F.T., Paspalas, C.D., Gamo, N.J., Yang, Y., and Wang, M. (2010). Dynamic Network Connectivity: A new form of neuroplasticity. Trends Cogn. Sci. *14*, 365–375.

Ascoli, G.A., Alonso-Nanclares, L., Anderson, S. a, Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., Buzsáki, G., Cauli, B., Defelipe, J., Fairén, A., et al. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat. Rev. Neurosci. *9*, 557–568.

Bacci, A., Huguenard, J.R., and Prince, D. a (2003). Functional autaptic neurotransmission in fast-spiking interneurons: a novel form of feedback inhibition in the neocortex. J. Neurosci. *23*, 859–866.

Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. J. Neurosci. *30*, 9424–9430.

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., and Abbott, L.F. (2013a). From fixed points to chaos: Three models of delayed discrimination. Prog. Neurobiol. *103*, 214–222.

Barak, O., Rigotti, M., and Fusi, S. (2013b). The Sparseness of Mixed Selectivity Neurons Controls the Generalization-Discrimination Trade-Off. J. Neurosci. *33*, 3844–3856.

Beaulieu, C. (1993). Numerical data on neocortical neurons in adult rat, with special reference to the GABA population. Brain Res. *609*, 284–292.

Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. *14*, 366–372.

Bouchacourt, F., and Buschman, T.J. (2019). A Flexible Model of Working Memory. Neuron *103*, 147-160.e8.

Brody, C.D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex *13*, 1196–1207.

Brunel, N. (2016). Is cortical connectivity optimized for storing information? Nat. Neurosci. *19*.

Buhl, E.H., Tamás, G., Szilágyi, T., Stricker, C., Paulsen, O., and Somogyi, P. (1997). Effect, number and location of synapses made by single pyramidal cells onto aspiny interneurones of cat visual cortex. J. Physiol. *500 (Pt 3*, 689–713.

Buonomano, D. V, and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. Nat. Rev. Neurosci. *10*, 113–125.

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. Neuron *68*, 362–385.

Compte, A., Constantinidis, C., Tegner, J., Raghavachari, S., Chafee, M.V. V, Goldman-Rakic, P.S., Wang, X.-J., and Tegnér, J. (2003). Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J. Neurophysiol. *90*, 3441–3454.

Dejean, C., Courtin, J., Karalis, N., Chaudun, F., Wurtz, H., Bienvenu, T.C.M., and Herry, C.

(2016). Prefrontal neuronal assemblies temporally control fear behaviour. Nature *535*, 420–424.

Dombrowski, S.M., Hilgetag, C.C., and Barbas, H. (2001). Quantitative architecture distinguishes prefrontal cortical systems in the rhesus monkey. Cereb. Cortex *11*, 975–988.

Doron, M., Chindemi, G., Muller, E., Markram, H., and Segev, I. (2017). Timed Synaptic Inhibition Shapes NMDA Spikes, Influencing Local Dendritic Processing and Global I/O Properties of Cortical Neurons. Cell Rep. *21*, 1550–1561.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. Nat. Rev. Neurosci. *2*, 820–829.

Duncan, J., and Miller, E.K. (2002). Cognitive Focus through Adaptive Neural Coding in the Primate Prefrontal Cortex. In Principles of Frontal Lobe Function, (Oxford University Press), pp. 278–291.

Durstewitz, D., Seamans, J.K., and Sejnowski, T.J.J. (2000). Neurocomputational models of working memory. Nat. Neurosci. *3 Suppl*, 1184–1191.

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

Fujisawa, S., and Buzsáki, G. (2011). A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. Neuron *72*, 153–165.

Fujisawa, S., Amarasingham, A., Harrison, M.T., and Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat. Neurosci. *11*, 823–833.

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

Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. Curr. Opin. Neurobiol. *37*, 66–74.

Fuster, J.M., and Alexander, G.E. (1971). Neuron Activity Related to Short-Term Memory. Science (80-.). *173*, 652–654.

Galarreta, M., and Hestrin, S. (1999). A network of fast-spiking cells in the neocortex connected by electrical synapses. Nature *402*, 72–75.

Goldman, M.S. (2009). Memory without feedback in a neural network. Neuron *61*, 621–634.

Hines, M.L., and Carnevale, N.T. (2001). NEURON: a tool for neuroscientists. Neuroscientist *7*, 123–135.

Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C.D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. J. Neurosci. *30*, 916–929.

Van Kerkoerle, T., Self, M.W., and Roelfsema, P.R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nat. Commun. *8*, 1–12.

Konstantoudaki, X., Papoutsi, A., Chalkiadaki, K., Poirazi, P., and Sidiropoulou, K. (2014). Modulatory effects of inhibition on persistent activity in a cortical microcircuit model. Front. Neural Circuits *8*, 7.

Kubota, Y., Kondo, S., Nomura, M., Hatada, S., Yamaguchi, N., Mohamed, A. a., Karube, F., Lübke, J., and Kawaguchi, Y. (2015). Functional effects of distinct innervation styles of pyramidal cells by fast spiking cortical interneurons. Elife 1–27.

Kuroda, M., Yokofujita, J., and Murakami, K. (1998). An ultrastructural study of the neural circuit between the prefrontal cortex and the mediodorsal nucleus of the thalamus. Prog. Neurobiol. *54*, 417–458.

Lee, D., and Seung, S. (2001). Algorithms for Non-negative Matrix Factorization. Adv. Neural Inf. Process. Syst. 13 556–562.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.

Lindsay, G.W., Rigotti, M., Warden, M.R., Miller, E.K., and Fusi, S. (2017). Hebbian Learning in a Random Network Captures Selectivity Properties of the Prefrontal Cortex. J. Neurosci. *37*, 11021–11036.

Lisman, J.E., Fellous, J.-M., and Wang, X.-J. (1998). A role for NMDA-receptor channels in working memory. Nat. Neurosci. *1*, 273–275.

Losonczy, A., and Magee, J.C. (2006). Integrative properties of radial oblique dendrites in hippocampal CA1 pyramidal neurons. Neuron *50*, 291–307.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L.L., Buschman, T.J., and Miller, E.K.K. (2016). Gamma and Beta Bursts Underlie Working Memory. Neuron *90*, 152–164.

Mante, V., Sussillo, D., Shenoy, K. V, and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Markram, H., Lübke, J., Frotscher, M., Roth, A., and Sakmann, B. (1997). Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J. Physiol. *500 (Pt 2*, 409–440.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. Nat. Rev. Neurosci. *5*, 793–807.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol. *100*, 1407–1419.

Miller, E.K., Erickson, C. a, and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci. *16*, 5154–5167.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science (80-.). *319*, 1543.

Morishima, M., and Kawaguchi, Y. (2006). Recurrent connection patterns of corticostriatal pyramidal cells in frontal cortex. J. Neurosci. *26*, 4394–4405.

Morishima, M., Morita, K., Kubota, Y., and Kawaguchi, Y. (2011). Highly differentiated projection-specific cortical subnetworks. J. Neurosci. *31*, 10380–10391.

Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., and Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. *114*, 394–399.

Nevian, T., Larkum, M.E., Polsky, A., and Schiller, J. (2007). Properties of basal dendrites of layer 5 pyramidal neurons: a direct patch-clamp recording study. Nat. Neurosci. *10*, 206–214.

Otsuka, T., and Kawaguchi, Y. (2009). Cortical inhibitory cell types differentially form intralaminar and interlaminar subnetworks with excitatory neurons. J. Neurosci. *29*, 10533–10540.

Packer, A.M., and Yuste, R. (2011). Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? J. Neurosci. *31*, 13260–13271.

Papoutsi, A., Sidiropoulou, K., Cutsuridis, V., and Poirazi, P. (2013). Induction and modulation of persistent activity in a layer V PFC microcircuit model. Front. Neural Circuits *7*, 161.

Papoutsi, A., Sidiropoulou, K., and Poirazi, P. (2014). Dendritic Nonlinearities Reduce Network Size Requirements and Mediate ON and OFF States of Persistent Activity in a PFC Microcircuit Model. PLoS Comput. Biol. *10*, e1003764.

Papoutsi, A., Kastellakis, G., and Poirazi, P. (2017). Basal tree complexity shapes functional pathways in the prefrontal cortex. J. Neurophysiol. jn.00099.2017.

Perin, R., Berger, T.K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. Proc. Natl. Acad. Sci. U. S. A. *108*, 5419–5424.

Perin, R., Telefont, M., and Markram, H. (2013). Computing the size and number of neuronal clusters in local circuits. Front. Neuroanat. *7*, 1.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., and Battaglia, F.P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat. Neurosci. *12*, 919–926.

Poirazi, P., Brannon, T., and Mel, B.W. (2003a). Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell. Neuron *37*, 977–987.

Poirazi, P., Brannon, T., and Mel, B.W. (2003b). Pyramidal neuron as two-layer neural network. Neuron *37*, 989–999.

Polsky, A., Mel, B.W., and Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. Nat. Neurosci. *7*, 621–627.

Polsky, A., Mel, B.W., and Schiller, J. (2009). Encoding and decoding bursts by NMDA spikes in basal dendrites of layer 5 pyramidal neurons. J. Neurosci. *29*, 11891–11903.

Povysheva, N. V, and Zaitsev, A. (2008). Parvalbumin-positive basket interneurons in monkey and rat prefrontal cortex. J. … 2348–2360.

Rao, S.G., Williams, G. V, and Goldman-Rakic, P.S. (2000). Destruction and creation of spatial tuning by disinhibition: GABA(A) blockade of prefrontal cortical neurons engaged by working memory. J. Neurosci. *20*, 485–494.

Raposo, D., Kaufman, M.T., and Churchland, A.K. (2014). A category-free neural population supports evolving demands during decision-making. Nat. Neurosci. *17*, 1784–1792.

Renart, A., Song, P., and Wang, X.-J. (2003). Robust spatial working memory through

homeostatic synaptic scaling in heterogeneous cortical networks. Neuron *38*, 473–485.

Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J., and Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. Front. Comput. Neurosci. *4*, 24.

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

Rolls, T.T. (1997). Neural Networks and Brain Function. 428.

Saez, A., Rigotti, M., Ostojic, S., Fusi, S., and Salzman, C.D. (2015). Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. Neuron *87*, 869–881.

Schaub, M.T., Billeh, Y.N., Anastassiou, C. a., Koch, C., and Barahona, M. (2015). Emergence of Slow-Switching Assemblies in Structured Neuronal Networks. PLOS Comput. Biol. *11*, e1004196.

Schuz, A., and Palm, G. (1989). Density of neurons and synapses in the cerebral cortex of the mouse. J. Comp. Neurol. *455*, 442–455.

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J.M., and Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience *146*, 1082–1108.

Song, S., Sjöström, P.J., Reigl, M., Nelson, S.B., and Chklovskii, D.B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol. *3*, e68.

Spaak, E., Watanabe, K., Funahashi, S., and Stokes, M.G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. J. Neurosci. *37*, 6503–6516.

Sreenivasan, K.K., Curtis, C.E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. *18*, 82–89.

Stokes, M.G. (2015). ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. 1–12.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron *78*, 364–375.

Thomson, A.M., and Lamy, C. (2007). Functional maps of neocortical local circuitry. Front. Neurosci. *1*, 19–42.

Tzilivaki, A., Kastellakis, G., and Poirazi, P. (2019). Challenging the point neuron dogma: FS basket cells as 2-stage nonlinear integrators. BioRxiv 251314.

Wang, X.-J. (1999). Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. J. Neurosci. *19*, 9587–9603.

Wang, H.-X., and Gao, W.-J. (2009). Cell type-specific development of NMDA receptors in the interneurons of rat prefrontal cortex. Neuropsychopharmacology *34*, 2028–2040.

Wang, M., Yang, Y., Wang, C.-J., Gamo, N.J.J.J., Jin, L.E.E.E., Mazer, J.A.A. a, Morrison, J.H.H.H., Wang, X.-J., and Arnsten, A.F.T. (2013). NMDA Receptors Subserve Persistent Neuronal Firing during Working Memory in Dorsolateral Prefrontal Cortex. Neuron *77*, 736–749.

Wang, X.-J., Song, H.F., Yang, G.R., Joglekar, M.R., and Newsome, W.T. (2019). Task representations in neural networks trained to perform many cognitive tasks. Nat. Neurosci. *22*, 297–306.

Wang, X.-J.X.-J., Tegnér, J., Constantinidis, C., and Goldman-Rakic, P.S. (2004). Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. Proc. Natl. Acad. Sci. U. S. A. *101*, 1368–1373.

Wang, Y., Markram, H., Goodman, P.H., Berger, T.K., Ma, J., and Goldman-Rakic, P.S. (2006).

Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat. Neurosci. *9*, 534–542.

Warden, M.R., and Miller, E.K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. J. Neurosci. *30*, 15801–15810.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature *393*, 440–442.

Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci.

Wit, E., van den Heuvel, E., and Romeijn, J.W. (2012). 'All models are wrong. ': An introduction to model uncertainty. Stat. Neerl. *66*, 217–236.

Zaitsev, A., Gonzalez-Burgos, G., Povysheva, N. V, Kröner, S., Lewis, D. a, and Krimer, L.S. (2005). Localization of calcium-binding proteins in physiologically and morphologically characterized interneurons of monkey dorsolateral prefrontal cortex. Cereb. Cortex *15*, 1178–1186.

Zhang, X.J., Li, Z., Han, Z., Sultan, K.T., Huang, K., and Shi, S.H. (2017). Precise inhibitory microcircuit assembly of developmentally related neocortical interneurons in clusters. Nat. Commun. *8*, 1–16.

Figures

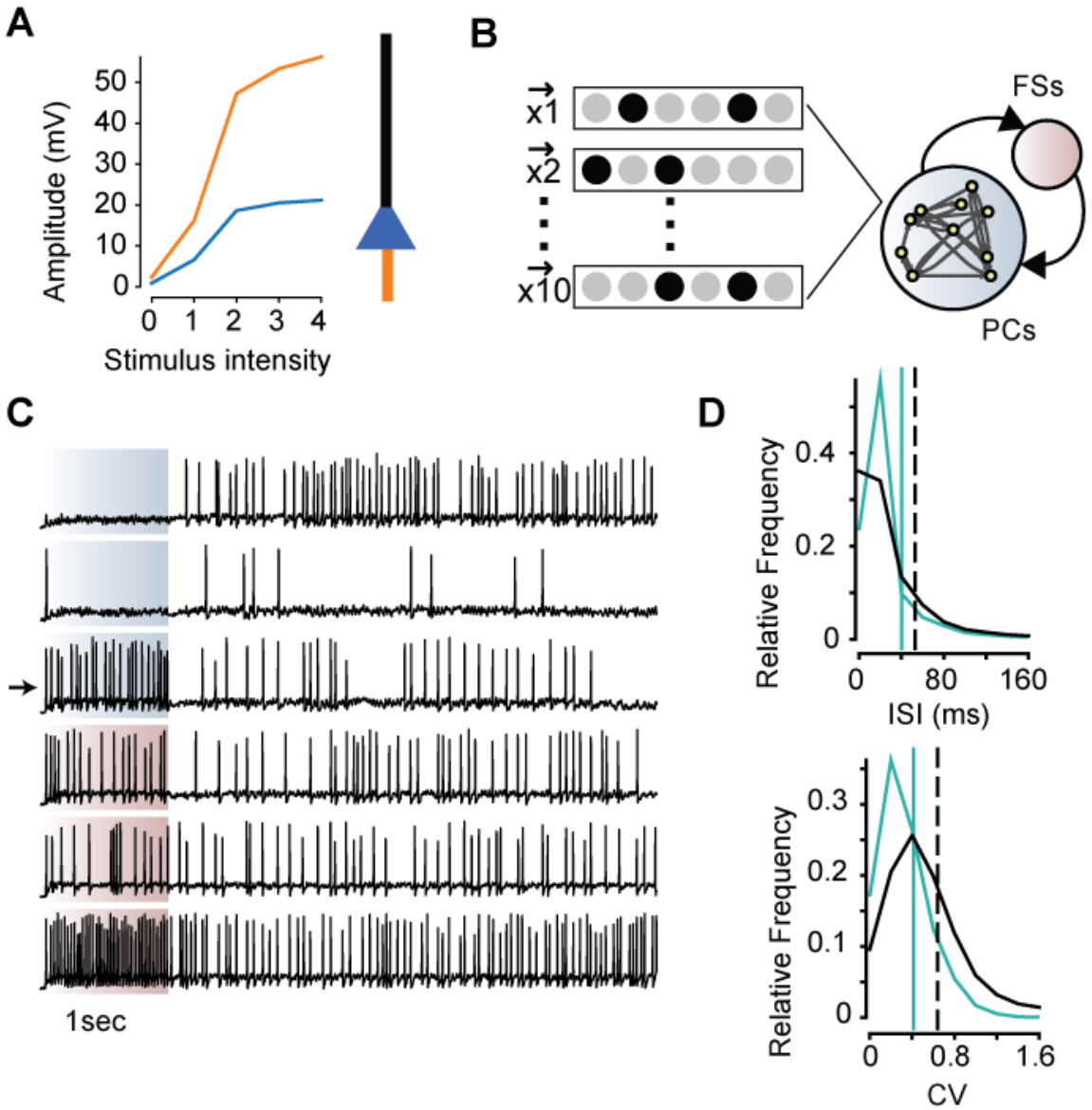


Figure 1 | PFC network model reproduces single neuron dynamics.

(A) Left: Non-linear NMDA responses are generated in the basal dendrites of the pyramidal neurons as in (Nevian et al., 2007). Orange: voltage response of the basal dendrite while increasing the number of activated synapses (2 events at 50Hz, activation of 1 to 21 synapses, with step 5). Blue: Corresponding somatic response. Right: Cartoon morphology of the

pyramidal neuron model. Orange: basal dendritic compartment, blue: somatic compartment, black: apical dendritic compartments (B) Schematic representation of the network model. Fifty randomly selected PCs were stimulated in each trial, as indicatively denoted by the vectors **x**. (C) Exemplar voltage responses of PCs (top three) and FSs (bottom three) in a single trial. The shaded region indicates stimulus presentation (1sec). Arrow: PC that received stimulation. (D) Histograms of inter-spike intervals (ISIs) (top) and Coefficient of Variation of ISIs (CV, bottom) for PCs. Data are pooled for different structured network instances and trials for the stimulus period (green) and post-stimulus period (black). Distribution means are indicated with vertical lines.

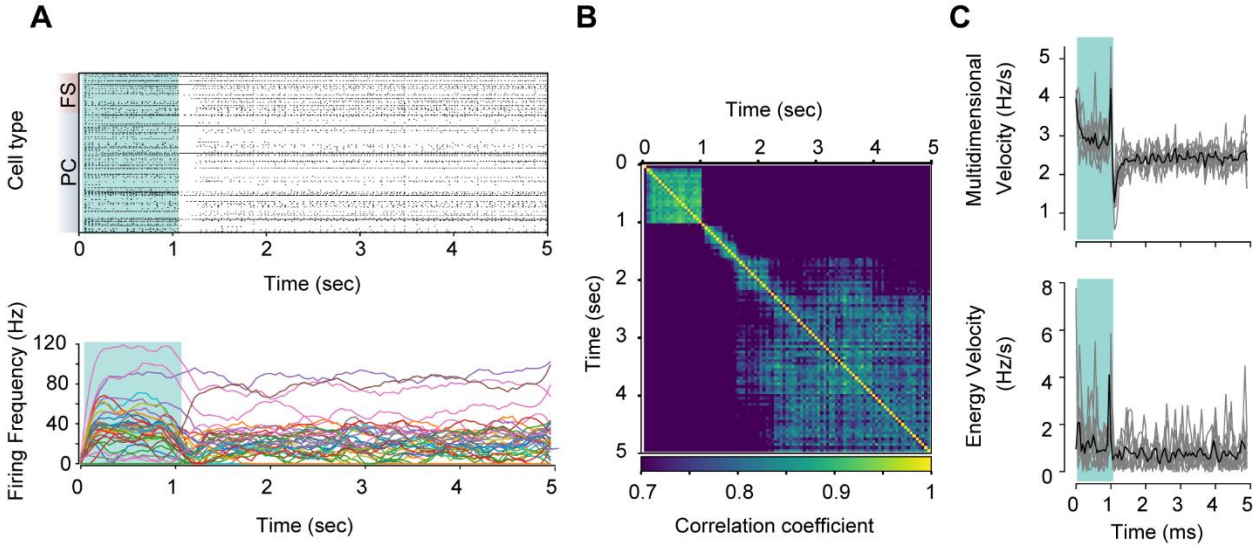
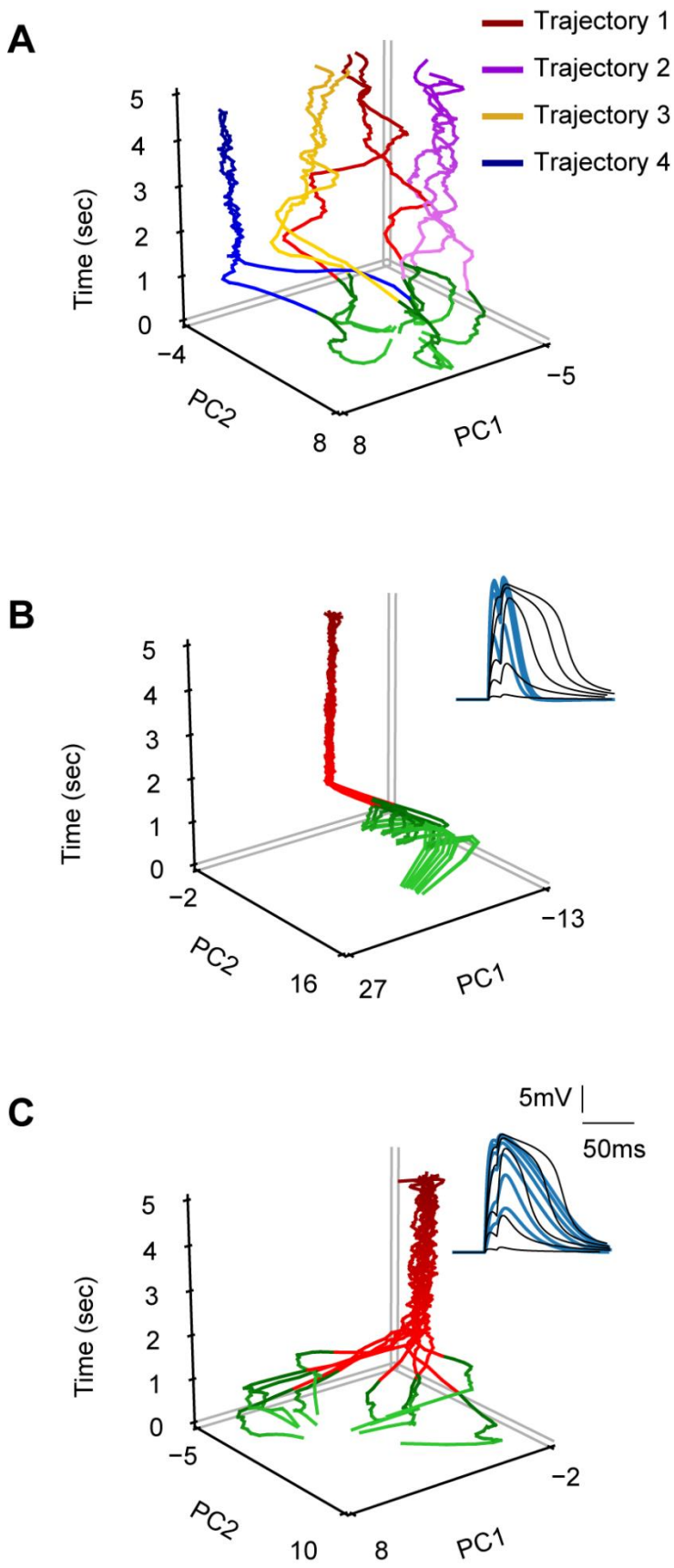


Figure 2 | Dynamic population activity following stimulation.

(A) Top: Raster plot of network activity in response to a 1-sec stimulus (shaded green area), for PCs and FSs. Bottom: Smoothed (temporal window: 11ms) firing frequencies of each PC, showing the different and highly dynamic responses over time (same trial as in A). Only PCs with average firing frequency across stimulus and post-stimulus epochs above 10Hz are shown here. (B) Cross correlation of network states (activity of PCs in a temporal window q=50ms) across the stimulus and post-stimulus epochs (aligned on stimulus onset, stimulus duration:

1sec). Color bar indicates the correlation coefficient, focused on the 0.7-1 range. (C) Top: Multidimensional velocity of network response, aligned on stimulus epoch onset. Bottom: Energy velocity (change in mean firing rate), aligned on stimulus epoch onset. Shaded region corresponds to stimulus epoch.


A
Trajectory 1
Trajectory 2
Trajectory 3
Trajectory 4
Time (sec)
5
4
3
2
1
0
−4
PC2
8
8
PC1
−5
B
Time (sec)
5
4
3
2
1
0
−2
PC2
16
27
PC1
−13
C
5mV
50ms
Time (sec)
5
4
3
2
1
0
−5
PC2
10
8
PC1
−2

1008

Figure 3 | NMDA nonlinearities create distinct stable trajectories.

(A) Population activity (n=9 trials with post-stimulus epoch activity) reduced in the first two principal components over time. Stimulus epoch is colored green. Different colors (yellow, red, magenta, blue) correspond to different clusters of post-stimulus activities, as dictated by k-means analysis. Color map luminosity (brighter to darker) denotes time progression. (B) As in A (n=8 trials), with the NMDA conductance set to zero, compensated by increased AMPA conductance. Simulating blockade of NMDARs results in the formation of one trajectory for different stimuli. Inset: voltage response at the soma for the control condition (black) and the NMDA blockade condition (blue). Both the non-linear jump and the depolarizing plateau are eliminated (blue), for increasing synaptic drive when NMDA conductance is zeroed (basal dendrite stimulation, 2 events at 50Hz, activation of 1 to 21 synapses). (C) As in A, after removal of the Mg^{2+} blockage (n=9 trials). This manipulation also results in the formation of a single trajectory. Inset: By removing the Mg^{2+} blockage, the non-linear jump is removed (black control, blue Mg^{2+} blockage), while the depolarizing plateau is retained (basal dendrite stimulation, 2 events at 50Hz, activation of 1 to 21 synapses).

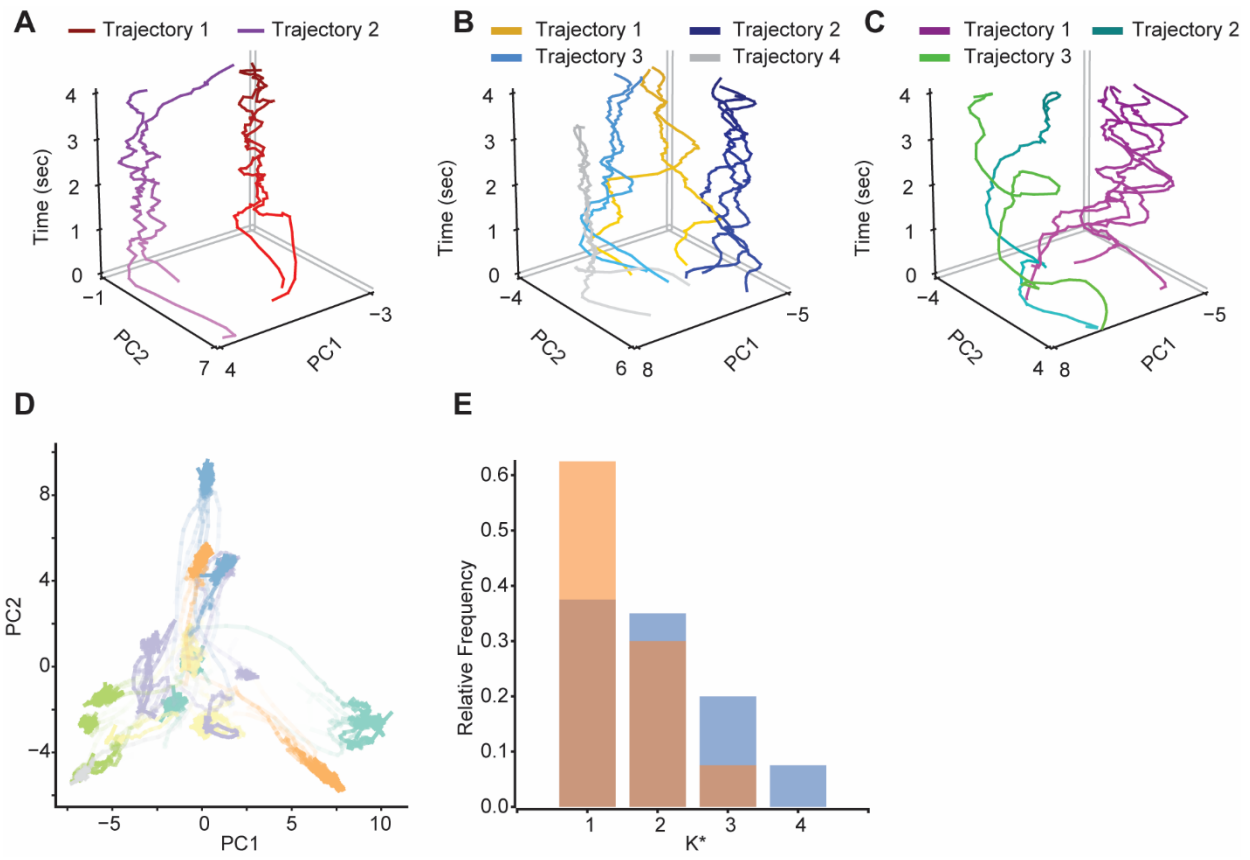


Figure 4 | Network activity states in response to same inputs for different synaptic arrangements. (A,B,C) Post-stimulus population activity in PCA space for the same structured network instance, but over three different synaptic arrangements. In the respective trials, the exact same neurons are stimulated. Clusters of trajectories are color coded. It is evident that the number of trajectories changes for each synaptic arrangement. (D) Pooled data for n=10 synaptic arrangements, showcasing that trajectories are mostly nonoverlapping (10 trials, 10 synaptic arrangements). Color code by synaptic arrangement. (E) Histogram of optimal number of clusters (K*) after application of k-means clustering, for all n=4 structured network instances of n=10 synaptic arrangements, for n=10 trials each (blue). In the presence of NMDA nonlinear synaptic integration, a randomly connected network produces fewer trajectories compared to the structured one (orange).

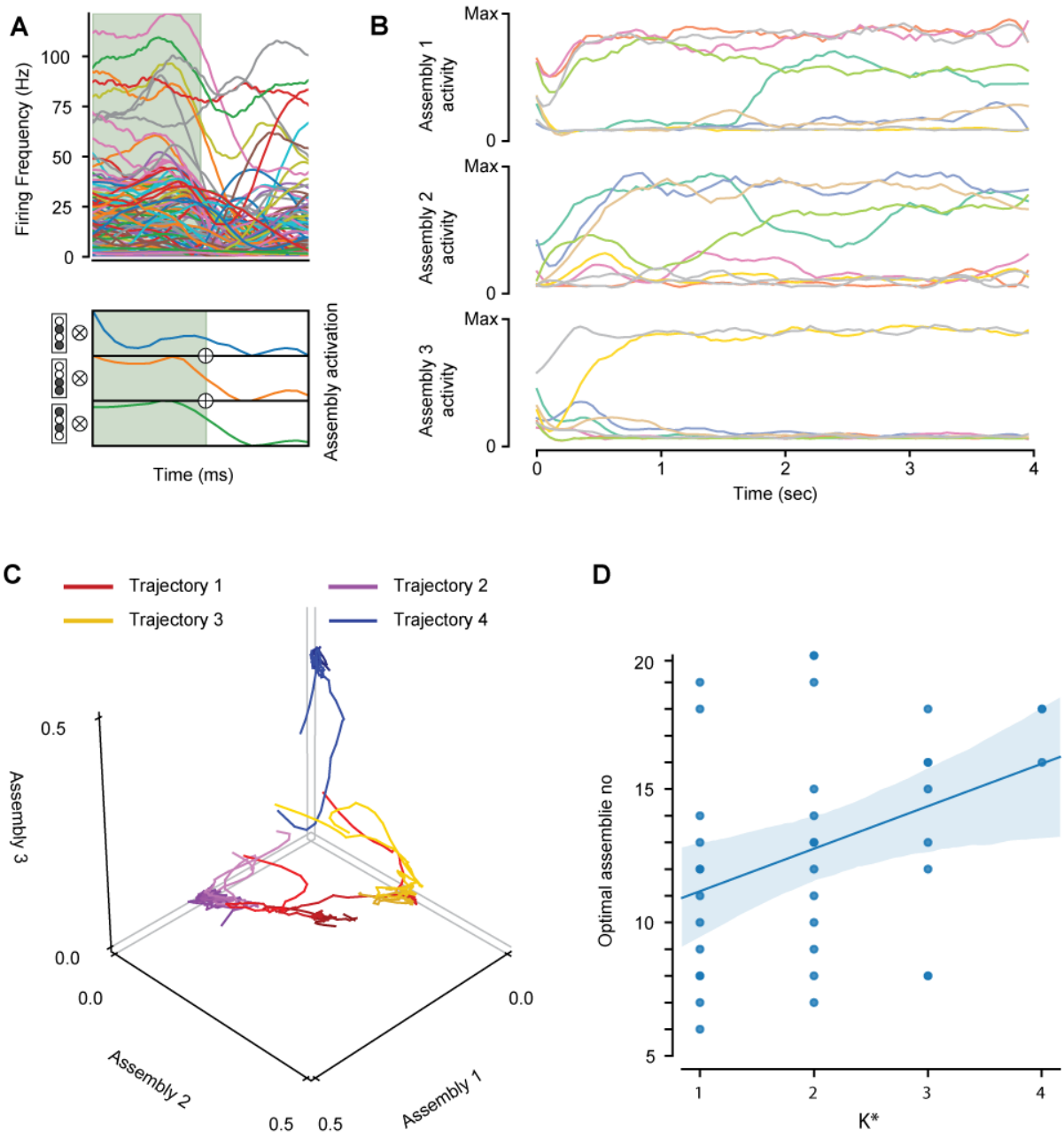


Figure 5 | Different neuronal assemblies are active for each population activity trajectory. (A) Top: Smoothed single neuron activities for the last 0.5 sec of stimulus (green area) and the initial 0.5 sec of the post-stimulus epoch of a given trial. Bottom: Same activity, but now reduced

with NNMF to 3 components (assemblies). Left: toy examples of assemblies created by 4 neurons. Each assembly is expressed as a vector of participating neurons and each neuron can participate in more than one assembly. Right: each of the assemblies is multiplied (outer product) with its activity vector over time. Population activity is reconstructed by the summation of the respective matrices across assemblies. The dynamic and highly irregular single neuron activity appears smooth in the assembly space and the shift in the representation from the stimulus to post-stimulus epoch is evident. (B) Post-stimulus epoch activity in 9 trials out of 10 with PC activity up until the last second of the simulation, reduced with NNMF to three assemblies and presented each time as the activity of each assembly. Note that assembly activity is dynamic across time. Colors correspond to the different trials. (C) Same data as in B, color coded per the trajectories identified in Figure 3A and plotted in the assembly space. It is evident that trajectories that cluster in PCA space (similar colors), appear to correspond to the activity of specific assemblies/assembly combinations. Note that assemblies are quasi-orthogonal (a neuron can be engaged in more than one assembly) and orthonormal axis are just for visualization purposes. This is expected due to the existence of neurons with mixed selectivity to memoranda. (D) Correlation analysis of optimal number of trajectories and assemblies shows that there is correlation ($r=0.37$ $p=0.017$), yet a single trajectory might not necessarily be created from only one assembly.