UAB
Universitat Autònoma de Barcelona

Prefrontal circuits underlying working memory encoding and maintenance

Nicolás Pollán Hauer

Prefrontal circuits underlying working memory encoding and maintenance

PhD Thesis

Author: Nicolás Pollán Hauer

Supervisor: Klaus Wimmer

Academic tutor: Lluís Alsedà Soler

Universitat Autònoma de Barcelona (UAB)
Centre de Recerca Matemàtica (CRM)
Doctorate Programme in Mathematics
September 2023

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Mathematics

at the

Universitat Autònoma de Barcelona

September 2023

Für meine Großmutter

Acknowledgments

Quiero empezar agradeciendo a Pablo Varona, que accedió amablemente a supervisar mi tesis de máster. Fue quien me introdujo al mundo de la neurociencia y encendió mi entusiasmo. No creo que yo hubiera continuado por este camino si no hubiese sido por él. Gracias.

También gracias a José Carlos Martínez, padre de Ana, iniciador a la física, compañero de cuerda, tutor y amigo, al final sobre todo amigo. Aparte de la tutela científica, gracias por los chistes y canciones que compartimos, que sea por muchos años.

Gracias a Albert Compte y a Jaime De La Rocha, que me pusieron en contacto con Klaus, y con los que ha sido un privilegio y un placer poder coincidir todos estos años.

Gracias al CRM, como lugar cabe casi decir, tan luminoso y amplio. Gracias a las personas de casi todos los días, en particular a Nuria Hernández, Consol Roca y Lluís Alsedà, que han sido acogedores, amables y me han ayudado en este proceso. Gracias a Víctor Navas, que sonreía y tenía sentido del humor, todos los días! Gracias a Claudia Fanelli, compañera, hermana de despacho, también luz y sonrisa. Gracias a Federico Devalle, un sol, qué buenos recuerdos, Federico!

Gracias a José Mari Esnaola, Txema. Hubo una época en la que me salvaba la vida una vez al día. Así recuerdo su tutela durante la mayor parte del tiempo que coincidimos en el CRM. Además de ayudarme con múltiples vicisitudes informáticas, ha sido un amigo durante estos años y espero que lo siga siendo por muchos más. Muchas gracias.

Muchas gracias a Klaus Wimmer, mi director de tesis. Sin Klaus tampoco hubiera empezado este camino. Pero tampoco quizá hubiera continuado el doctorado si no hubiera contado con su apoyo en todo momento. Gracias por respetarme y confiar en mí. Gracias por introducirme al oficio, a la comunidad estupenda del Barccsyn y por muchos buenos momentos.

Agradezco mucho haber compartido momentos con Alex Roxin, con quien me encantaría seguir saliendo a correr una vez al año. También con Alex Hyafil, ¿Cuándo es el próximo concierto...? Estoy agradecido de haber coincidido con otras tantas personas en la comunidad de neurociencia de Barcelona (nunca me gustó mucho cómo suena Barccsyn). No esperaba tanta y tan buena interacción social dentro de la comunidad científica!

Gracias a mi madre, que siempre ha mostrado interés por lo que hago y en lo que me interesa. Me ha dado libertad para escoger, que tanto me cuesta a veces, y oportunidades de conocer muchos mundos. Así he acabado a aquí. Gracias por estar.

Gracias a mi padre, que también me acompañó estos años, en casi cada chiste y en muchos otros momentos.

Gracias a Tere y Enrique, también padres. Apoyándome siempre, preocupándose del nenín.

Quiero agradecer especialmente a mi compañera, Daria Stepanova, que siempre ha estado ahí, desde mi llegada al CRM. Desde que se fue Txema, antes quizá, ella también me salva la vida un par de veces también al día. Sabe que todas las palabras son pocas.

Abbreviations

List of abbreviations

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
EEG	electroencephalography
fMRI	functional magnetic resonance imaging
GABA	gamma-aminobutyric acid
LFP	local field potential
MEG	magnetoencephalography
MGS	memory-guided saccade task
NMDA	N-methyl-D-aspartate receptor
ODR	oculomotor delayed-response task
PCA	Principal Component Analysis
PFC	prefrontal cortex
PPC	parietal cortex
PSTH	peristimulus time histogram
V1	primary visual cortex
WM	working memory
LIF	leaky integrate and fire model
MGS	memory-guided saccade task, same as oculomotor delayed-response task
LC	Locus Coeruleus
BG	Basal Ganglia

Abstract

Working memory, the capacity to maintain and manipulate information in our minds when it is no longer available in the environment, is a central function of cognition. One of the most important neuronal correlates of this cognitive function are the so-called persistent neurons, which respond selectively to sensory stimulation and sustain their increased activity even after removing the stimulus. This phenomenon, most frequently observed in the prefrontal cortex, has been successfully described by neural network models with attractor dynamics. However, only a few of the neurons engaged in working memory tasks have persistent activity. Moreover, analysis of the experimental recordings at the population level reveals that the code undergoes a change between the stimulus presentation and the maintenance epochs, which is not compatible with a working memory code that would only rely on stably active persistent cells. The prevalence of this finding has motivated the proposal of alternative mechanisms, but current computational models that explain dynamics fail to include stable epochs or do not provide a clear mechanistic interpretation.

In this thesis, we use statistical data analysis and neural network modeling to investigate whether specialized neuronal subpopulations underlie the stable and dynamic working memory codes.

First, we investigated the connection between the observed dynamics in the working memory code and the functional structure of the prefrontal circuits. We analyzed prefrontal recordings from behaving macaque monkeys and observed that feature selectivity is non-randomly distributed across the neurons. This non-random or structured feature selectivity distribution is related to functional distinct subpopulations whose contrasting activity explains the dynamic to stable transition in the working memory code.

Second, we developed a computational model that represents three functional subpopulations as attractor networks working on different dynamic regimes. The model illustrates how the population structure, which implies different neurons active at different task epochs, is directly related to the dynamic transition in the code. Furthermore, we show how the three-network architecture can be easily extended to account for additional features, such as ramping activity and variable maintenance periods.

Third, our subpopulation-based networks have the functional advantage of being robust against distracting stimuli. The model captures the experimentally observed vulnerability to distractors presented shortly after stimulus removal. Moreover, it predicts that top-down feedback enhances the overall network's robustness.

In summary, we show how the presence of functional subpopulations in the prefrontal cortex can be related to the dynamic to stable transition in the working memory code and to an enhanced capacity to filter out distracting stimuli. In conclusion, our work reconciles attractor dynamics with the observed dynamic changes in the code, still suggesting that attractor dynamics are essential for working memory maintenance.

Contents

Chapter 1

Introduction

1.1 Working memory, the basis of cognition

When we say that someone has a very good memory, we usually refer to his or her capacity to remember events in the past (you both met the day after Christmas in 2003, but you have forgotten), dates (Einstein won the Nobel prize in 1921) or facts (the capital of Thailand is Bangkok). In these cases, we are talking about long-term memory. For all that concerns the present work, working memory has little to do with long-term memory, even though these cognitive functions are related.

Working memory, sometimes referred to as short-term memory[1], is the capacity to hold and manipulate information in our minds. While long-term memory can be thought of as the information in the books we keep on our shelves, working memory refers to the chapter of the book we are currently reading or writing. In the inevitable analogy with computers, long-term memory corresponds to the hard disk (where we permanently store documents, pictures, and installed programs) and working memory to random-access memory (RAM), which provides a temporary workspace where information is actively processed and manipulated.

Situations where we rely on WM in life are countless. The examples span from the most boring to the most creative and illustrate some of its essential aspects. A simple example arises when you are required to input a password sent to your mobile phone to complete some transaction. You usually cannot open two screens simultaneously, so you memorize the numbers, close one page, and type them into the second page. You usually remember the password by reading it once if it contains up to four digits. However, for longer passwords, we either have to memorize them in chunks (going back several times to the page containing the password) or write them down. We find ourselves in similar situations when we are engaged in more complex tasks such as mathematical calculations or playing chess. More jolly examples are found in the listening and writing of poetry and music, where the traces of past sounds help to find the intonation and the ending of the following line. But among all the activities that justify the need for short-term storage and processing, the most important might be communication. In a normal conversation, remembering recent information is essential to make sense of the one that follows, whether talking or listening. Without this capacity, we risk losing track of the subject's identity even before a sentence concludes. Imagine then making sense of a joke or a story!

From these few examples, we can already understand how much working memory is involved in most distinctly human activities. For a good reason, this function is sometimes called the "cornerstone of cognition". Indeed, not even the smartest apes complete bank transactions, write poetry, or maintain conversations. However, this does not mean working memory is not at all developed in other species. On the contrary, the vast research carried out with animals assumes that many neural mechanisms, even the ones related to abstract operations, are conserved to some extent across species (Cisek, 2019).

In the following sections, we will summarize some of our knowledge about working memory based on behavioral, neurophysiological, and computational modeling studies.

[1] Although some authors highlight differences between these terms, we will use them interchangeably in this text.

1.2 Behavioral findings

Some of the properties that have been repeatedly reported and that can give cues about the underlying mechanisms of working memory are its shortcomings or limitations. For a more thorough account of working memory characteristics, see Chai *et al.* (2018); Fuster (2015); Oberauer *et al.* (2018).

1.2.1 Working memory has limited capacity and fades with time

A working memory limitation that we are familiar with has to do with its capacity: the number of items we can hold in mind. An experimental paradigm that has been used to quantify memory capacity (Bays *et al.*, 2009) involves presenting human subjects with a different number of colored squares on each trial (Figure 1.1a, example array of size 6 is shown) and asking them to report the color of one of the items after a blank period (delay) where no information is displayed. As expected, the subjects' precision decreases as a function of the array size (Figure 1.1b). From the example of remembering passwords, we know that increasing the number of digits to be remembered eventually renders the task impossible. We also know how demanding it can be to do some calculations without the help of a blackboard or a piece of paper. So the experimental results in Bays *et al.* (2009) come with little surprise. However, behavioral measures can be used to test the validity of predictions made by different models of brain function (Bays *et al.*, 2009; Oberauer, 2009). Moreover, this ability to test mechanistic predictions is increased by complementing behavioral with electrophysiological measurements (Vogel & Machizawa, 2004; Vogel *et al.*, 2005; Wei *et al.*, 2012).

Another feature of working memory is that it fades with time. This decay can be observed as an increase in the memory error for longer delay times (Figure 1.1c). Using a color and orientation task with human subjects, Pertzov *et al.* (2017) quantified the combined impact of the array size and the delay length, illustrating their correlation (Figure 1.1b). They found that the time decay of memory fidelity was more pronounced the larger the number of presented items. Their results suggest that items encoded simultaneously compete for working memory resources (Bays *et al.*, 2009; Edin *et al.*, 2009; Oberauer *et al.*, 2018; Pertzov *et al.*, 2017; Shin *et al.*, 2017; Wei *et al.*, 2012).

A further example of working memory time decay comes from a classical task with non-human primates (Funahashi *et al.*, 1989). In this paradigm, monkeys learn to report the remembered location of a visual stimulus with a saccadic eye movement (Figure 1.1d). The animals learn to keep their gaze at the center of the screen (the fixation spot is indicated by a cross) until the moment of the motor response. In experiments where neuronal activity is recorded while the monkeys perform the task, fixation ensures that the retinal input to the frontal cortex has a constant reference point. In this way, only task-related events can trigger changes in cortical activity. During a trial, the monkey fixates at a fixation spot at the center of a screen, where a cue is presented for 300 ms in one out of eight possible locations. After a period with no stimulation (memory delay), the fixation spot is switched off, indicating the monkey to perform an eye movement toward the remembered cue location (Figure 1.1d). The distribution of eye positions after the saccades shows an increasing dispersion as a function of the delay length (Figure 1.1e), indicating the gradual loss in memory accuracy or memory drift.

Given that working memory assists us during behaviorally relevant time scales, and that it is a function that we differentiate from long-term memory storage, we would reasonably expect it to be reliable for a limited time. In this regard, the decay of working memory with time can be seen as a feature rather than a limitation. Intuitively speaking (although this can be easily quantified; Vogel & Machizawa, 2004; Vogel *et al.*, 2005) if we need our working memory for active interaction with the surrounding world, we do not want its resources to be fully allocated but rather available for processing new input. Working memory's fading fidelity prevents our minds from being saturated and allows us to interact meaningfully with the surrounding world.

1.2.2 Working memory resists distracting stimuli

The limited capacity of working memory makes it necessary to prioritize the behaviorally most relevant information. We usually refer to this feature as working memory's robustness against distracting stimuli (Lorenc *et al.*, 2021) If nothing would protect the information stored in working memory from distraction (of external or internal sources), our daily function would be dramatically impaired. A simple task such as keeping a sequence of three numbers in mind would only be possible in silent surroundings, we would not be able to communicate in crowded environments or enjoy polyphonic music ...

A result that has been repeatedly reproduced is that the interference or impact is larger when distractors have a high featural overlap with the memory target (Lorenc *et al.*, 2021). This effect is observed when comparing distractors of the same or different sensory modality as the target (Bae & Luck, 2019; Oberauer *et al.*, 2018) or for comparison between different classes within the same modality (such as faces and shoes, Figure 1.2a; Jha *et al.*, 2004). For distractors of the same type as the target stimulus, larger differences along the same low-level feature produce more significant memory biases (Nemes *et al.*, 2012; Rademaker *et al.*, 2015). For instance, this result can be observed when memory for grating orientation is tested in human subjects (Figure 1.2b).

These behavioral results give a first cue about the possible underlying mechanism. Even though converging evidence places the frontal regions of the cortex as the principal locus of working memory (Funahashi, 2017; Fuster, 2015), the role of sensory cortices during maintenance remains unclear (Xu, 2020). The weak interference of stimuli of different sensory modalities suggests that their representation during maintenance is orthogonal. This orthogonality between sensory modalities can be explicitly represented by the dedicated sensory cortices, whereas the neuronal representations in association regions such as the prefrontal cortex, which respond to stimuli of all modalities (Raposo *et al.*, 2014), is more susceptible to overlaps (Fuster, 2015; Kandel *et al.*, 2000; Raposo *et al.*, 2014). On the other hand, the dependence of the memory bias on the target-distractor similarity along the same features (Figure 1.2b) is predicted by models that assume a continuous neural encoding of features such as orientation, location, or color (Compte *et al.*, 2000; Edin *et al.*, 2009).

A related result concerning the vulnerability of working memory representations is that distractors presented early during the delay have a stronger impact on behavior (Suzuki & Gottlieb, 2013). This result has been shown for monkeys performing an ocular-delayed response task (ODR) (Figure 1.2c). We cannot infer this result from the decay of working memory fidelity with time, which would predict a stronger impact for late distractor presentations. However, this dependence can be captured by a computational model, as we will show in Section 4.3.

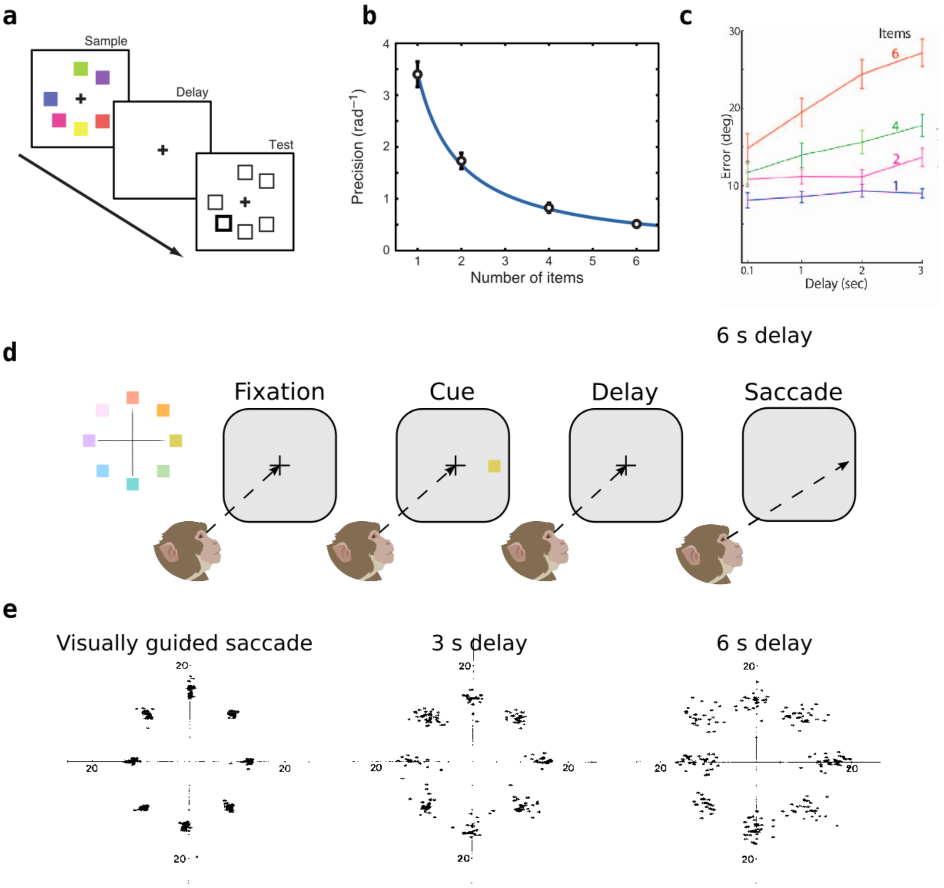

Figure 1.1: Working memory can hold a limited number of items for a limited time. a,b Working memory precision in humans performing a color report task. **a** a different number of colored squares is presented on each trial (6 in the example shown). After a delay period, the color of the square presented at the cued location must be reported. **b** Subjects' performance drops as a function of the number of items presented in the cue array. Precision is calculated as the reciprocal of the standard deviation of the response error (color similarity is mapped to a 0 to 2π interval). **c** The response error increases both with the number of items and the delay length in a color-location working memory task. **d,e** In an ocular-delayed-response task (ODR), monkeys have to fixate in the center of a screen (fixation spot indicated by the cross) where a cue is presented (Cue). The monkey keeps fixating during a delay period of variable duration (Delay). When the fixation spot is switched off, the monkey has to do an eye movement indicating the location of the presented cue (Saccade). **e** Eye position (as measured by an eye-tracker set-up) at the end of different task conditions: left, a visually guided task (no delay); middle, a 3 s delay; right, a 6 s delay.

a,b Reproduced from Bays *et al.* (2009) by permission of the publisher under a Creative Commons license. **c** Reproduced from Pertzov *et al.* (2017) with permission from American Psychological Association. **e** Reproduced from Funahashi *et al.* (1989) under the permission of the American Physiological Society.

1.2.3 Lesions and ablations: the importance of prefrontal cortex

Before neurophysiological advances started to enable increasingly detailed observations of the neuronal processes, subjects with brain damage or practiced ablations provided insights about the neural basis of working memory. These cases have informed us about the location of brain function already long before the advent of neurophysiological measuring techniques (Fuster,

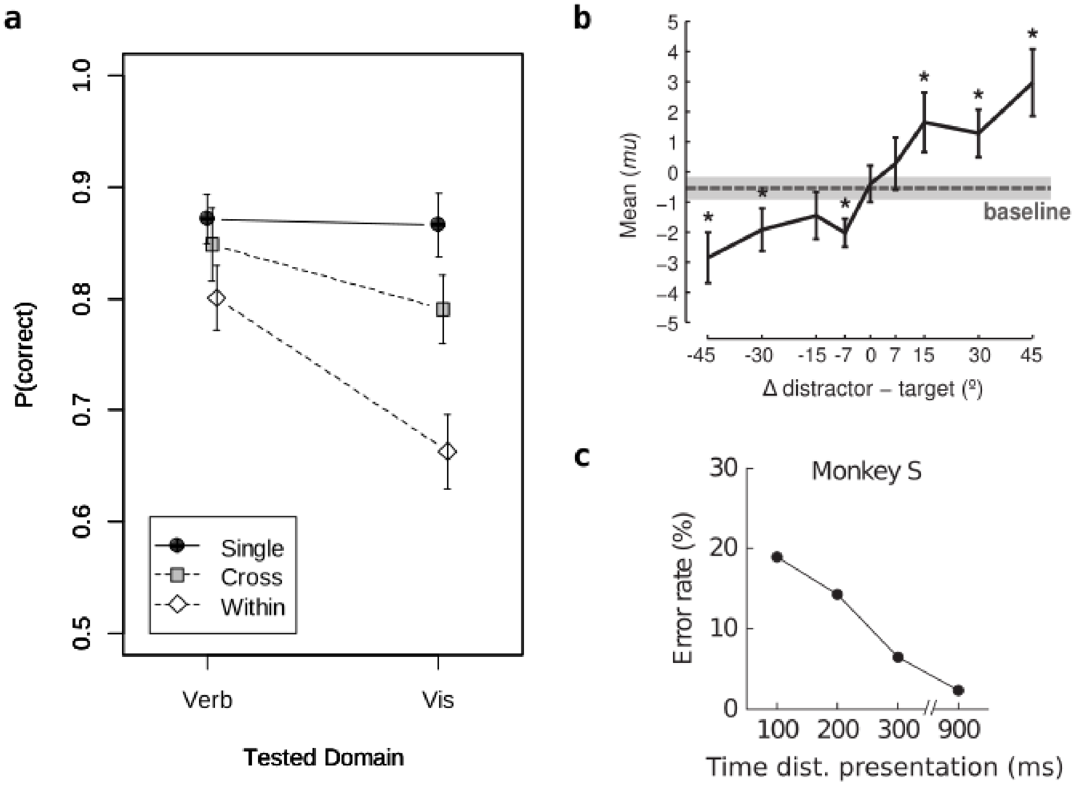

Figure 1.2: The impact of distracting stimuli on working memory depends on their similarity with the target stimulus and presentation time. **a** Distractors of the same category or type as the target stimulus (within) have a greater impact on performance than distractors of different types (cross). Single is the label for the condition with no distractor. The different categories are visual and verbal stimuli. **b** Memory bias as a function of the difference between the target and the distractor orientation in an orientation working memory task. **c** Distractors presented early in the delay have a greater impact on the performance of an ocular-delayed-response task by a monkey.
a Reproduced from Oberauer *et al.* (2018) with permission from American Psychological Association , **b** reproduced from Rademaker *et al.* (2015) with permission from with permission from American Psychological Association. **c** Reproduced from Suzuki & Gottlieb (2013) with permission from Springer Nature.

2015).The famous case of Phineas Gage (Harlow, 1848, 1869), who lost part of his left frontal cortex during an accident, is probably one of the earlier accounts of the relationship between anatomy and function. Along with a drastic change in Gage's personality (for which the case is often cited), the physicians reported a noticeable decrease in the capacity to plan future actions (Fuster, 2015; Harlow, 1848, 1869), which relies strongly on working memory. The correlation between frontal lesions and short-term memory deficit has been reported repeatedly in posterior studies with humans (Lewinsohn *et al.*, 1972; Milner, 1982)[2], with lesions in other cortical regions having a significantly smaller impact on behavior.

Cortical ablations have also been used to establish the importance of the prefrontal cortex in experiments with non-human species. In his experiments, Jacobsen (Jacobsen, 1935, 1936) uses

[2] It was a usual practice to remove (lobotomize) parts of the brain in subjects who suffered from diverse symptoms such as depression, schizophrenia or epilepsy (Fuster, 2015)

a delay task in which monkeys have to remember the location of a food reward to retrieve it after a memory delay (Figure 1.3). He found that lesions of the frontal cortex had a significantly higher impact on behavior than lesioning other regions (motor, premotor, parietal, and temporal lobes).

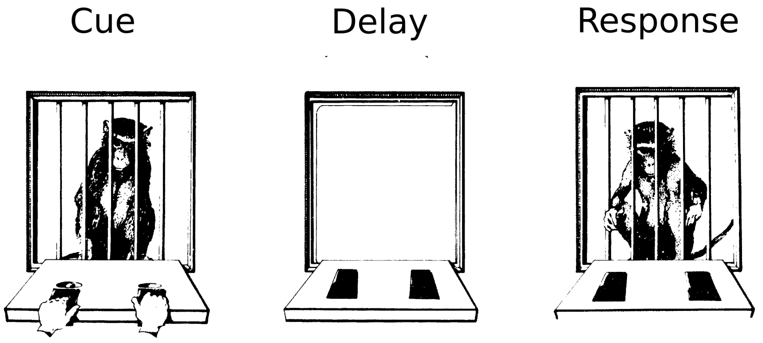

Figure 1.3: Classical working memory task with monkeys A monkey is briefly shown in which out of two nearby standing plates, a food reward is hidden (Cue). After a delay period, during which the monkey cannot see the plates (Delay), the animal can retrieve the food (Response). The location of the reward is randomly varied by the experimenter from trial to trial.
Reproduced from (Goldman-Rakic, 1992) with permission from the Wiley publishing group.

1.3 Neurophysiology and the discovery of delay activity

We will now consider further insight into working memory function brought up by electrophysiological measurements. Luckily, lobotomy is no longer a common way of treating symptoms related to brain function. Invasive experiments with animal species are nowadays the primary way of obtaining measures of highly resolved neural activity, such as calcium activity, local field potentials (LFPs), single neuron spike counts, or intracellular voltage recordings. One of the last developed techniques, optogenetics, allows experimenters to activate or inactivate neurons with a temporal precision of milliseconds, helping to establish causal relationships in neural circuits (Chen *et al.*, 2022). On the other hand, human electrophysiology relies to a great extent on the use of non-invasive techniques such as magnetic resonance imaging techniques (MRI and fMRI) and electric or magnetic encephalography (EEG and MEG; Glover, 2011; Guest & Love, 2017).

1.3.1 Persistent activity

In 1971, J. M. Fuster, concomitantly with experimenters of the Primate Center in Japan, discovered one of the most important correlates of working memory in the cortex: persistently active cells (Figure 1.4). The authors recorded from prefrontal cells of monkeys performing an ocular-delay response task (Fuster & Alexander, 1971, Figure 1.1d) and found that some of them exhibited sustained stimulus-selective activity during the memory delay. This activity profile has been measured repeatedly in delayed response paradigms and is regarded as the clearest neural correlate of working memory (Leavitt *et al.*, 2017; Sreenivasan & D'Esposito, 2019; Zylberberg & Strowbridge, 2017), as it offers a non-ambiguous and straightforward interpretation. Many studies have reported persistent activity in other regions of the brain (Sreenivasan & D'Esposito, 2019); however, the firing rate modulation in these areas during the delay is often weaker and does not span the entire delay (Leavitt *et al.*, 2017). The accumulating evidence from electrophysiological measurements over the last decades (Sreenivasan & D'Esposito, 2019) keeps underscoring the predominance of association regions, particularly the prefrontal cortex, for working memory maintenance.

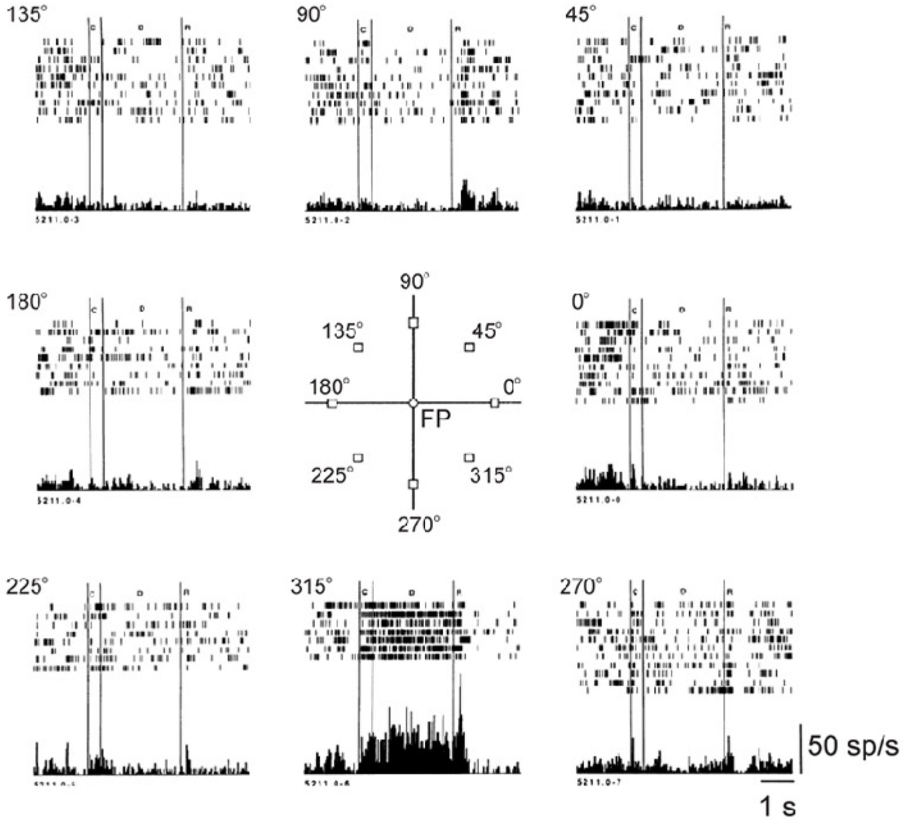

Figure 1.4: A prefrontal neuron with sustained stimulus-selective activity during the delay, neural correlate of working memory Raster plots (top) and peri-stimulus time histograms (PSTHs) (bottom) of a monkey's prefrontal neuron during the performance of an ODR task. Each panel corresponds to trials in which the cue was presented at one of the eight locations. When the cue is presented at 315 °, the neuron increments its activity, and it sustains the increased firing throughout the delay (bottom center panel).

reproduced from (Funahashi *et al.*, 1989) under the permission of the American Physiological Society.

1.3.2 Dynamics and stability in the working memory code

While the persistent cells appear to play a prominent role as memory carriers, they do not explain all the relevant experimental observations. Different types of activity profiles are observed among the task-modulated cells, which are not persistent. The importance these neurons have for the working memory code is not clearly understood, but insight can obtained by analysis at the population level. To analyze the population code during delay-response tasks, many authors have trained algorithms to decode the stimulus identity from the neural activity. When a decoder is trained and tested on data corresponding to different time windows during the task (cross-temporal decoding), its accuracy can be plotted as a squared pattern (Figure 1.5). While a stable (non time-evolving) code allows an algorithm trained at a given time to perform accurately when tested at other times (off diagonal regions of the pattern), a dynamic code (time evolving) prevents the algorithm from this generalization. As a result, stability causes square regions of above chance accuracy in the cross-temporal decoding while dynamic code produces diagonal patterns and poor off-diagonal generalization (Dehaene *et al.*, 2015).

Based on the cross-temporal analysis, many studies have found periods of strong dynamics in the working memory code that are incompatible with a completely persistent representation (Figure 1.5; Mendoza-Halliday & Martinez-Trujillo, 2017; Spaak *et al.*, 2017; Stokes *et al.*, 2013; Stroud *et al.*, 2023; Wolff *et al.*, 2015).

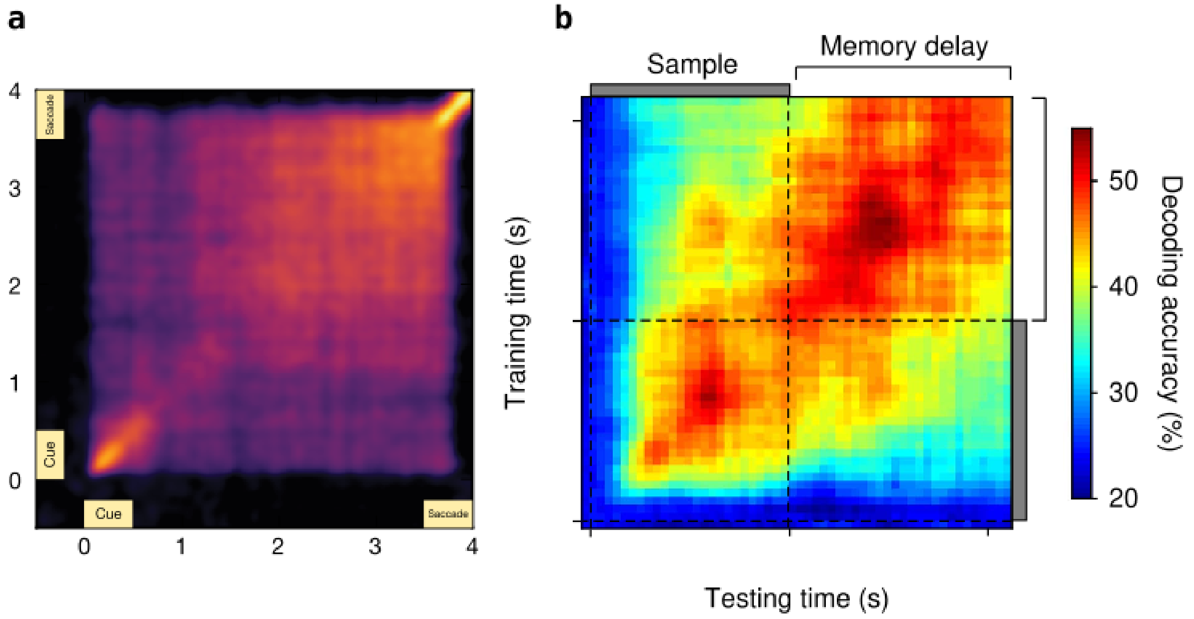

Figure 1.5: Cross-temporal decoding reveals that working memory code in the prefrontal cortex undergoes a transition from cue to delay epochs. Training and testing a decoder on the data at different combinations of time windows highlights regions of stability and dynamics, respectively. Off-diagonal generalization indicates that the code is stable, while off-diagonal significant drops in accuracy are associated with a time-evolving (dynamic) code. **a** Cross-temporal decoding pattern for data from monkeys performing an ODR task: the code is dynamic during the cue presentation and at the beginning of the delay, and it increasingly stabilizes towards the end of the delay. **b** Cross-temporal decoding pattern, data recorded from monkeys performing a 2-alternative task (choose left or right), the strong off-diagonal generalization during the memory delay period indicates the stability of the code during this epoch. In contrast, the accuracy drops substantially when the decoder is trained during cue and tested on delay and vice-versa.

a Reproduced from Spaak *et al.* (2017) with permission of the publisher under a Creative Commons Attribution 4.0 International License (CC-BY), **b** reproduced from Mendoza-Halliday & Martinez-Trujillo (2017) with permission of the publisher under a Creative Commons Attribution 4.0 International License.

The general hallmark across the different studies is that the working memory codes during cue and delay periods are mutually close to orthogonal. This orthogonality implies strong dynamics from one period to another (Figure 1.5). Eventually, the code becomes stable during the delay (Figure 1.5). Whereas the stability of the code can be related to persistent selective activity (Sreenivasan & D'Esposito, 2019), it is less clear what the across-epoch dynamics represent. Transient activity profiles of single neurons (Funahashi *et al.*, 1989; Markowitz *et al.*, 2015; Mendoza-Halliday & Martinez-Trujillo, 2017), which have even been observed in the shape of sequential activity (Batuev *et al.*, 1980), are examples of dynamic code. However, these observations do not explain a dynamic-to-stable transition in the code. This thesis addresses this question by analyzing an experimental data set and proposing a computational model that explains the main observed features.

Activity-silent working memory

It has also been proposed that working memory may be activity-silent, with memories stored in synaptic traces through short-term synaptic plasticity (Mongillo *et al.*, 2008). Human EEG activity provide some support for this idea (Wolff *et al.*, 2015) by showing that the stimulus decodability, which decreases to chance during a memory delay, can be recovered by flashing the subjects with a non-selective stimulus. This reactivation would be compatible with the stimulus being encoded in the shape of modified synaptic connection strengths because unspecific input could retrieve the information encoded in the synapses. However, these findings remain controversal (Barbosa *et al.*, 2021). Some authors have proposed that spiking and synaptic mechanisms could coexist and underlie the interference between recent memories (Barbosa *et al.*, 2020). In this thesis, however, only spiking-related mechanisms will be discussed.

1.3.3 Mixed selectivity and structure in prefrontal networks

A general characteristic of the prefrontal cortex is the heterogeneity of the single neuron activities (Asaad *et al.*, 2000; Dang *et al.*, 2021; Rigotti *et al.*, 2013). Neurons in the prefrontal cortex can respond to different task features (Asaad *et al.*, 2000; Finkelstein *et al.*, 2021; Markowitz *et al.*, 2015; Mendoza-Halliday & Martinez-Trujillo, 2017; Yang *et al.*, 2022) as well as to stimuli from different modalities (Raposo *et al.*, 2014). This type of response profile, usually called mixed-selectivity (Asaad *et al.*, 2000; Dang *et al.*, 2021) (Figure 1.6c,d), contrasts with the more stereotyped dynamics of neurons in lower regions and sensory cortices, which respond to specific features (frequency, orientation, location, color ...).

When a neuron's selectivity is linearly mixed, there is no interaction between the contribution of different features to the neuron's activity (Figure 1.6c). This type of mixed selective response implies that a neuron's tuning curve for a given feature (e.g. location) can be scaled under the modulation of another task feature (e.g. task condition) but does not change its shape (Figure 1.6c). On the other hand, non-linear mixed selectivity implies that the contribution of different features to a neuron's response can be mutually dependent. This dependence can produce context-dependent tuning curves (Figure 1.6d). The example neurons in Figure 1.6 illustrate the respective cases; in the case of non-linear mixed selective, the neuron's preferred cue location depends on the task condition.

Recent studies (Fusi *et al.*, 2016; Rigotti *et al.*, 2010, 2013) have argued that the prevalence of mixed-selectivity, in particular, non-linear mixed selectivity in the prefrontal cortex, offers computational advantages in terms of the input output operations a network can perform (see Rigotti *et al.* (2013) for a geometric explanation and discussion). The account of the prefrontal cortex's complex response profiles and the potential functional advantages of its mixed selectivity, (Asaad *et al.*, 2000; Rigotti *et al.*, 2010, 2013), combined with the widespread use of dimensionality reduction techniques (Cueva *et al.*, 2018, 2020; Mante *et al.*, 2013), has led to a reduced emphasis on analyses based on classical selectivity and functional and anatomical structure (Barak *et al.*, 2013; Murray *et al.*, 2017a; Stroud *et al.*, 2023).

However, other studies have contributed to a complementary view. Recording from mice during a delay task, Wu *et al.* (2020) reported a substantially larger number of cells with selectivity for a single task feature as compared to mixed selective cells (Figure 1.6e). In the same study, the authors observed that the neurons with sustained delay activity were mainly located in superficial cortical layers. The layer specificity of the mnemonic responses has been observed in studies involving different species.

In monkeys, Markowitz *et al.* (2015) found that different types of prefrontal response profiles were more likely to be found at different topographical locations and in different cortical layers (Figure 1.7a). Moreover, as we will show in Section 4.1, most neurons in the Markowitz *et al.* (2015) data set have stable selectivity across task epochs and conditions. Combining macaque data from different working memory tasks Bastos *et al.* (2018) also found a higher density of neurons with delay activity in superficial layers (Figure 1.7c). Additionally, Bastos *et al.* (2018) observed differences between the power spectrum in different layers, with gamma oscillations dominating in the superficial and alpha-beta in deeper layers (Figure 1.7d). Whereas the authors attribute activity in superficial layers to maintenance, deeper layers are thought to be in charge of loading information from upstream and transmitting information to downstream regions. Similar results were obtained by applying high-resolution fMRI on humans. Finn *et al.* (2019); Lawrence *et al.* (2018) observed delay activity in superficial frontal layers and response-related activity in deeper layers (Figure 1.7b).

Evidence for functional specialization and non-linear mixed selectivity offer contrasting views or scenarios of prefrontal function (Dubreuil *et al.*, 2022; Hirokawa *et al.*, 2019; Raposo *et al.*, 2014). These scenarios are not mutually exclusive, but the extent to which one or the other prevails can be informative about the mechanisms underlying memory encoding and maintenance. For example, layer specificity of prefrontal activity could enhance the robustness to distracting stimuli. We know that distracting stimuli produce a smaller impact in prefrontal than other upstream areas (Lorenc *et al.*, 2018; Suzuki & Gottlieb, 2013; Yoon *et al.*, 2006), which indicates that distractors are filtered earlier in the processing line (Finkelstein *et al.*, 2021; Lorenc *et al.*, 2018, 2021; Suzuki & Gottlieb, 2013; Yoon *et al.*, 2006). However, even if filtered to some extent, distracting stimuli reach the prefrontal cortex (Finkelstein *et al.*, 2021; Suzuki & Gottlieb, 2013). Restricting delay activity to superficial layers can be a way of further protecting memory content from the distracting input, which will first arrive to deeper layers (Kandel *et al.*, 2000). Likewise, neurons active during different task epochs (Finkelstein *et al.*, 2021; Inagaki *et al.*, 2019; Markowitz *et al.*, 2015; Mendoza-Halliday & Martinez-Trujillo, 2017; Yang *et al.*, 2022) can be linked to dynamic transition observed in the code at the population level (Figure 1.5). We will explore these possibilities in Sections 4.1 and 4.2, based on the analysis of a data set and on the development of a computational model.

Before presenting our own results, we will review some of the mechanisms that computational models have proposed to explain working memory in the prefrontal cortex.

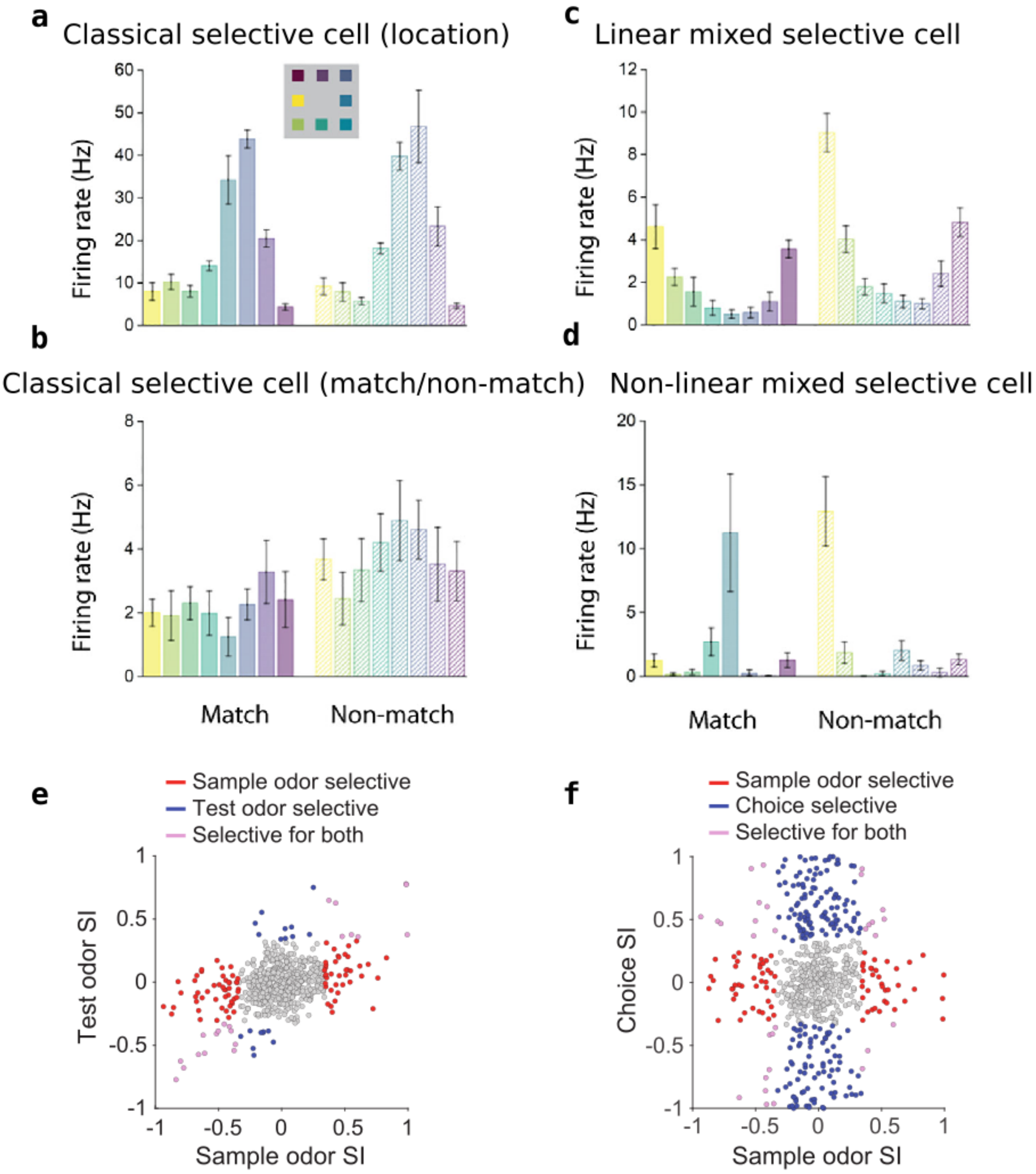

Figure 1.6: Prefrontal neurons can be selective to different combinations of task features. **a-d** Four single-neuron examples illustrating different types of selectivity in a match/non-match working memory task. **a** Neuron selective to stimulus location. **b** neuron with selectivity to cue-target similarity (match/non-match). **c** a neuron with selectivity for stimulus location *and* cue-target similarity (match/non-match). **d** selectivity for stimulus location depends on the match/non-match condition, indicative of non-linear mixed selectivity. **e,f** In mice anterior-lateral motor cortex (ALM), an analogous region to our prefrontal cortex, a greater proportion of neurons are selective for a given task feature and not the other (red and blue dots) than for both (purple).

a-d reproduced from Dang *et al.* (2021) under the terms of the Creative Commons Attribution 4.0 International License (CC-BY). **e,f** reproduced from Wu *et al.* (2020) with permission of Elsevier

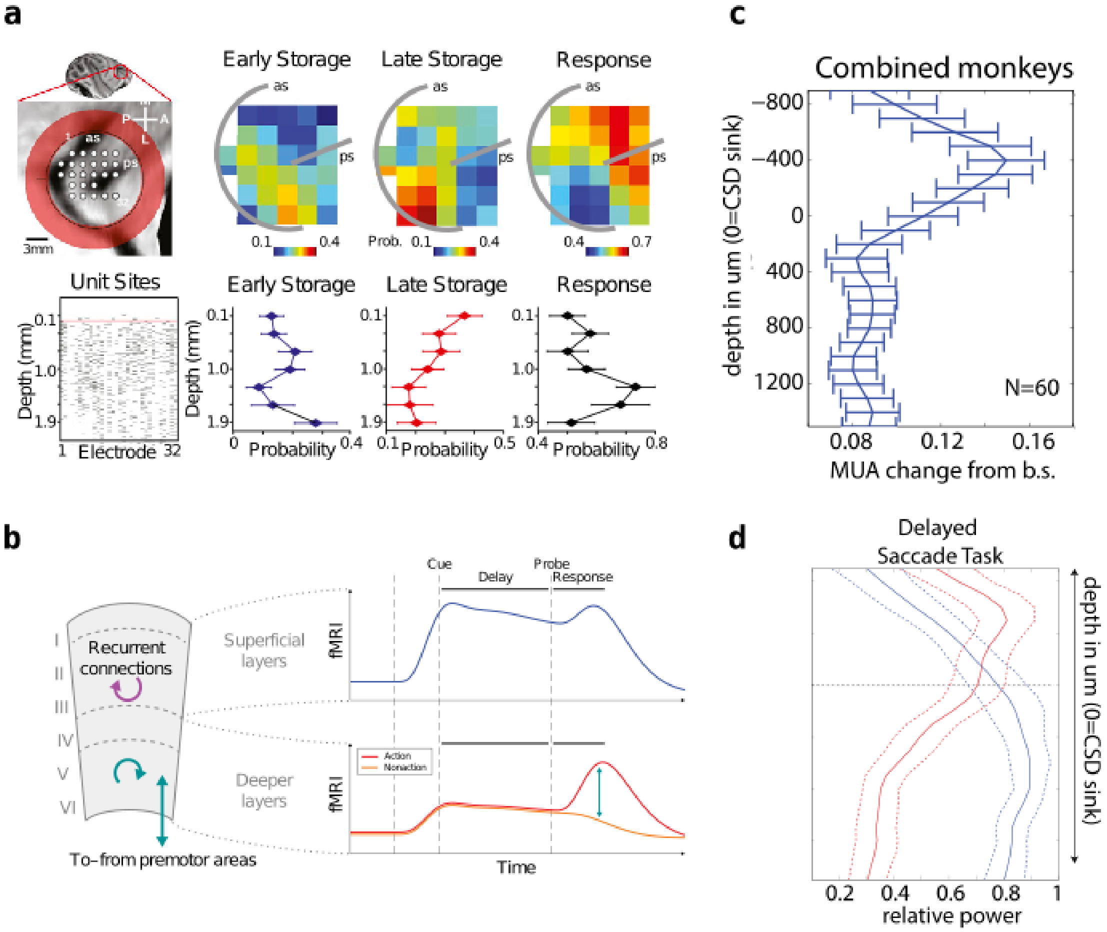

Figure 1.7: Evidence for structure in prefrontal cortex a Anatomical segregation of neurons with functional different response profiles in monkey PFC. Top, location of the recording site in the brain, and density of neurons of three distinct functional types (labeled as Early storage, Late storage, and Response neurons) as a function of the position, indicated by the electrode terminals. Bottom, maps of the recording depth at which spiking activity was recorded on each electrode and density of neurons of the three different types as a function of the depth. **b** High-resolution functional magnetic resonance (fMRI) shows layer-dependent activity in human prefrontal cortex during a working memory task. Top, increased delay activity of neurons in superficial layers. Activity in deeper layers is strongly modulated during the response but not during the delay. **c-d** layer-dependent activity in the PFC of monkeys performing WM tasks. **c** Change in activity from baseline during a memory delay for different recording depths. (average over 60 monkeys). **d** power in different frequency bands for the same monkeys as a function of recording depth.

a Reproduced from Markowitz *et al.* (2015) under the permission of the National Academy of Sciences **b** reproduced from Finn *et al.* (2019) with permission of Nature Research **c,d** Reproduced from Bastos *et al.* (2018) with permission of the publisher under a Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1.4 Computational models of working memory

Computational modeling is a playground to develop hypotheses regarding the mechanisms underlying certain brain functions, and to analyze data with a solid theoretical foundation.

The disadvantages are that modeling provides great freedom to hypothesize mechanisms and to explore the parameter space (unlike real experiments), which makes model validation non-straightforward. We will briefly review classical working memory models and some recent works that provide the necessary context and a base for our own modeling approach.

1.4.1 Mechanisms and models of persistent activity

As mentioned above, the discovery of persistent activity is probably the most significant finding in the working memory literature (Leavitt *et al.*, 2017). It has received a proportional attention in the computational modeling community.

Persistent activity has been related to two fundamental mechanisms: bistability at the single-cell level and bistability at the network level (attractor networks). Autonomously bistable cells have been observed in brain regions such as the hippocampus (Traub & Jefferys, 1994). Computational studies (Miles & Wong, 1987; Traub & Wong, 1982) have shown that persistent activity in a population of cells can be induced by the simultaneous presence of a proportion of cells with autonomous bistability and overall sufficiently strong recurrent connection. Although cell-autonomous bistability has not frequently been reported in prefrontal cells (Zylberberg & Strowbridge, 2017), its contribution to prefrontal persistent activity cannot be ruled out.

Models implementing bistability at the network level rely on strong recurrent connections between neurons with similar selectivity (Sreenivasan & D'Esposito, 2019; Zylberberg & Strowbridge, 2017) and long-range inhibition (Figure 1.8a). Additionally, the presence of synaptic currents with slow dynamics, such as NMDA-receptor-mediated currents, enhances the network's stability (Wang, 1999). For binary stimuli, attractor models usually comprise two excitatory populations that compete through reciprocal inhibition (Wang XJ, 2002; Wong & Wang, 2006). When the stimuli are evenly sampled from a continuous and periodic feature (e.g. locations on a circle, as in the ocular-delayed response task), the two-population system is naturally extended to a network with ring topology and Mexican hat-like connectivity (Figure 1.8a). The ring attractor model, also called bump attractor (Compte *et al.*, 2000, see Figure 1.8), has had great success in reproducing some of the key experimental findings related to working memory (Barbosa *et al.*, 2020; Wimmer *et al.*, 2014). In particular, the model predicts the decrease in accuracy of the response in delayed-response tasks (Figure 1.1d,e) as the diffusion of the activity bump during the memory delay (Wimmer *et al.*, 2014). Pair-wise correlations of single-neuron activity during the delay that mimic those in the experiments (Wimmer *et al.*, 2014) further support the bump attractor. The model also captures the dependence of distractor impact with target-distractor similarity (Compte *et al.*, 2000; Pertzov *et al.*, 2017; Rademaker *et al.*, 2015). Moreover, extensions of the model that include short-term synaptic plasticity have been used to explain the effect of past-trials memories (serial biases, Barbosa *et al.*, 2020).

A common critique of ring-attractor models is their vulnerability to heterogeneities in the connectivity. The rotational invariance of the connectivity in ring attractor networks (connectivity strength between two neurons only depends on the difference between their preferred cues Figure 1.8a) is a natural consequence of the periodicity of the stimulus. Any disruption asymmetric in the connectivity causes the memory bump to drift towards so-called hotspots (Hansel & Mato, 2013). Some models have proposed mechanisms, such as short-term plasticity, that can reduce the drift in the bump when the connections are heterogeneous (Hansel & Mato, 2013; Renart *et al.*, 2003). In a more general framework, Darshan & Rivkind

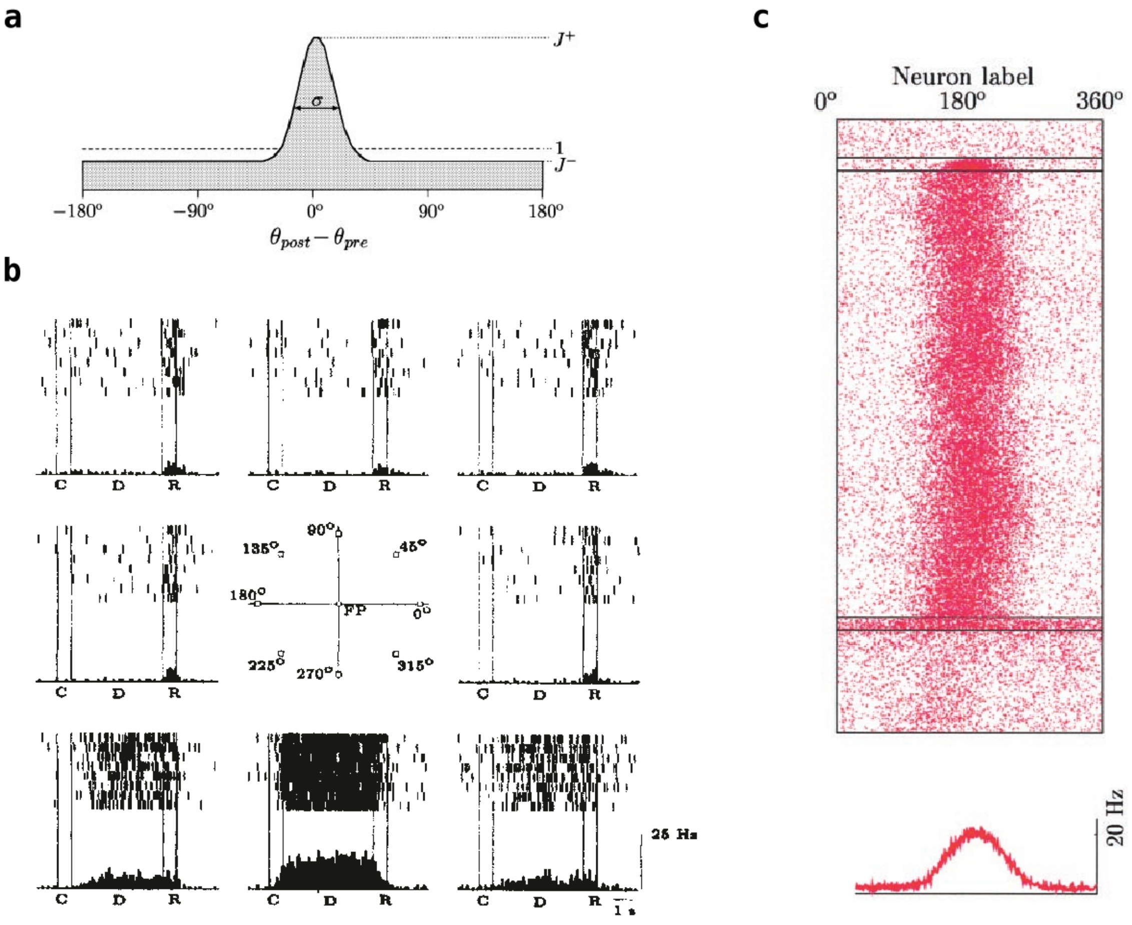

Figure 1.8: Persistent activity modeled by a network with ring topology with strong recurrent connections The ring-attractor model captures the selective and persistent firing observed in prefrontal neurons (Figure 1.4) **a** footprint of the synaptic connectivity as a function of the difference in preferred cue between neurons in the network. Excitation dominates over inhibition only in the short range. **b** A model neuron mimicking the selective persistent response to stimulus observed in Funahashi *et al.* (1989), (Figure 1.4). The model neuron is most selective to a stimulus presented at 315°, and with sustained firing during the delay. **c** Network activity in the ring model. Raster plot of 2048 excitatory neurons, arranged according to their preferred cue location (from 0 to 360 °). y−axis, time. When the cue is presented (indicated by the black lines), a bump of activity forms, which is sustained during the delay until the network is flashed with a non-selective stimulus that destroys the bump (resets network to baseline).

Reproduced from Compte *et al.* (2000) under the permission of Oxford University Press.

(2022) have recently shown that asymmetries in the connectivity are compatible with ring attractor dynamics even without further stabilization mechanisms.

1.4.2 Models including dynamics

With all its virtues, the ring-attractor is the canonical model of stability in the prefrontal working memory. Alternative models or extensions of the ring attractor network have been proposed to explain the dynamics observed in the code (Figure 1.5).

A model example that exhibits dynamic delay activity is given by the networks used in Barak *et al.* (2013). The authors trained the linear readout weights of an RNN to solve a classical working memory task (Romo *et al.*, 1999). They showed that the network's performance was similar to that of an attractor network designed for the same paradigm. Notably, the network output during the entire delay is chaotic and only becomes informative by the time of the motor response (Figure 1.9a). This model can account for dynamic activity and non-linear mixed selectivity but fails at capturing epochs of stability, which are present in the prefrontal code (Figure 1.5).

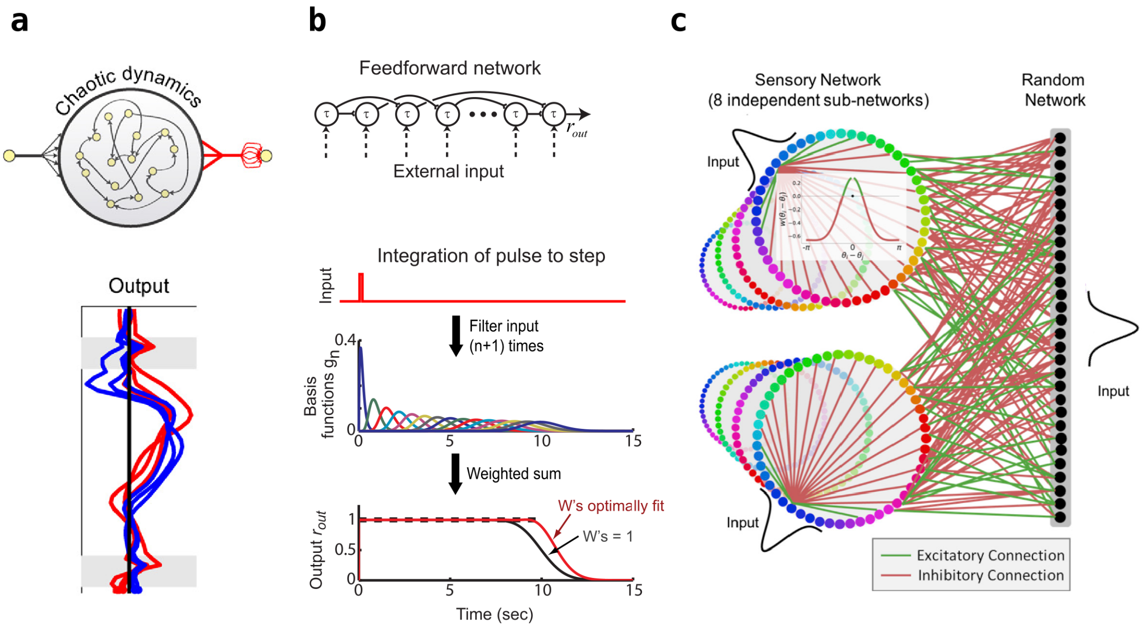

Figure 1.9: Dynamic code in working memory models. **a** A recurrent network working in a chaotic regime solves a working memory task where the frequency of an initial tone must be compared to a second one, provided after a delay period (task paradigm as in Romo *et al.*, 1999). Top, network scheme. Only the linear readout weights are modified during the network training. Bottom, The output of a trained network is shown for different trials of the two different stimulus combinations ($f_1 > f_2$ red lines, and $f_2 > f_1$ blue lines). The gray shadings indicate stimulus presentation times. The network output is time-varying until the presentation of the second stimulus makes the two conditions (blue and red colors) distinguishable. **b** Network with sequential activity. From top to bottom: network schematic; external input pulse; Sequential activation of the network's units, each neuron is represented by a different color; network output, for properly adjusted readout weights, the output is flat, ensuring a stable readout. **c** Network composed of an arbitrary group of sensory modules (colored rings on the cartoon) that project reciprocally and randomly to another network (labeled as random network). Recurrent connections in the sensory modules and the random network are too weak to sustain delay activity. Maintenance is achieved by properly tuning the reciprocal connections between sensory modules and the random network. The stimulus is presented to the pertinent sensory modules *and* to the random network to achieve a dynamic transition in the code from cue to delay. For details see Bouchacourt & Buschman (2019).
a Reproduced from Barak *et al.* (2013) with permission of Elsevier, **b** reproduced from Goldman (2009) with permission of Elsevier, **c** reproduced from Bouchacourt & Buschman (2019) with permission of Elsevier.

Some works have captured mechanisms of dynamic activity that are compatible with sustaining a stable representation (Druckmann & Chklovskii, 2012; Ganguli & Latham, 2009; Goldman, 2009; Murphy & Miller, 2009; Murray *et al.*, 2017a). Goldman (2009) proposed a network of feedforwardly connected neurons that can sustain delay activity (Figure 1.9b). A stable readout of this sequential activity is obtained by correctly choosing the linear readout weights (Figure 1.9b). Analogously, Druckmann & Chklovskii (2012) have shown that general non-sequential patterns of dynamic activity can produce stable readout, provided the connectivity matrix meets specific requirements (largest eigenvalue equal to 1). Intuitively, time-varying single-neuron activity can produce stable readout when positive modulations in some neurons are compensated by negative modulations in others. Hansel & Mato (2013) proposed a model where the effect of heterogeneous connections is stabilized by short-term synaptic plasticity. Their memory bumps are biased toward heterogeneity-induced hotspots, but the drift is slow compared to behavioral time scales. This model succeeds at capturing various prototypical response profiles observed in prefrontal neurons (Funahashi *et al.*, 1989).

These results widen the view of persistent delay activity from a phenomenon observed at the level of single cells to a feature exhibited by the population activity. However, none of them gives an account of the *transition* from dynamic to stable code that is observed in the experiments (Figure 1.5, Mendoza-Halliday & Martinez-Trujillo (2017); Parthasarathy *et al.* (2019); Spaak *et al.* (2017).

Finally, some authors have proposed non-normal amplification (Hennequin *et al.*, 2012; Kao & Hennequin, 2019) as a mechanism to generate initial dynamics and later stability in the code. Murray *et al.* (2017a) illustrated this idea with a simple linear model. To produce the desired result (initial dynamics and later stability), their stimulus input had to be partially aligned with the stable subspace. The partial alignment of the input allows the non-normal dynamics to produce transient activity orthogonal to the stable subspace before the code stabilizes. Stroud *et al.* (2023) extended this circuit proposal to a more general framework that uses non-linear recurrent networks. These authors argue that non-normality can be a general property of prefrontal circuits, providing fast and energetically optimal information loading.

These models (Murray *et al.*, 2017a; Stroud *et al.*, 2023) do account for a change from dynamic to stable in the population code. In both of them, the transition from dynamic to stable has to do with the non-normality of the connectivity. On a more intuitive level, non-normality is a property exhibited by almost any connectivity matrix that is non-symmetric (Horn & Johnson, 1985), such as networks with predominant feedforward or feedback structure. We argue that the qualitative hallmark of the experimental data, the cue-to-delay transition in the code, is inevitably associated with a feedforward relay of information in the frontal networks. Considering this, any model that explains the results should include a certain degree of feedforward structure. It is still an open question how this feedforward structure is implemented. We will address this point in the Section 4.1 and Section 4.2 of this thesis.

1.4.3 Mixed selectivity and structure in prefrontal networks

The model proposed by Bouchacourt & Buschman (2019) can also capture the dynamic and stable epochs of the working memory code. Their network proposal explicitly includes different modules or subnetworks (Figure 1.9c). An arbitrary number of sensory networks, which receive the stimulus input, are reciprocally connected to a second network (labeled as random network). Crucially, to exhibit transient dynamics at the beginning of the delay, additional projections must be included from the stimulus to the random network (Figure 1.9c, right). This model contrasts with all previously presented in that it explicitly includes structure in the shape of separate networks. The dynamic transition can be explained in terms of the activation of the different subnetworks, aligning with the above-commented findings on layer dependency of prefrontal and with the analysis we present in Sections 4.1 and 4.2. A shortcoming of the model is that the exact proposed architecture cannot easily be related to observed functional or anatomical specialization.

In this thesis, we will investigate the role of functional specialization in the prefrontal cortex. We will first analyze recordings from behaving monkey (Section 4.1) and then present a computational model that is based on the analysis of the experimental data (Section 4.2). Our results suggest that the observed dynamics during working memory as well as the prefrontal resistance against distracting stimuli (Section 4.3) can be strongly related to groups of neurons firing at different task epochs.

Chapter 2

Goals

In this thesis, we investigate the neural mechanisms that underlie the dynamics of the working memory code. Our overall goal is to understand to which extent the presence of specialized subcircuits within prefrontal networks underlies the observed changes in the code between the cue and delay epochs.

In section Section 4.1, we analyze a set of recordings from the prefrontal cortex of behaving macaque monkeys. Our aims are:

1 To quantify the degree of structure in the distribution of the feature selectivity.

2 to relate the observed structure to the neurons' selectivity for time, stimulus location, and task condition.

3 to contrast our analysis with the growing view of non-linear mixed selectivity as the primary mechanism behind prefrontal dynamics.

In section Section 4.2, we present a computational model that implements a concrete three-subpopulation structure that captures the activity of three of the experimentally observed functional profiles. We want

4 relate the activity of the different subnetworks to the dynamics of the code at the population level (cue to delay transition). With this purpose, we want to compare the patterns of cross-temporal decoding of the different model subnetworks with the respective experimental counterparts.

5 In contrast with more complex network proposals, we want to illustrate that the dynamics observed in the WM code are compatible with different neuronal subnetworks that undergo attractor dynamics.

6 Along with central aims, we want to propose a mechanism capable of generating the observed ramping activity in some neurons during the delay. We will frame the discussion of the ramping generating mechanisms in the context of previous works that have addressed this question (Finkelstein *et al.*, 2021; Inagaki *et al.*, 2019)

In section Section 4.3, we test how the model behaves in the presence of distracting stimuli. This analysis serves as a means to:

7 test whether the subpopulation structure has some functional advantages. For this, we will compare our model's behavior with that of a single bump attractor ring.

As an overall aim, we emphasize that multiple subnetworks with simple and interpretable dynamics (as our attractor networks) can be an alternative to single networks with complex dynamics that often, are difficult to interpret.

Chapter 3

Methods

3.1 Visual working memory task with macaques

3.1.1 The ocular-delay response task

Full experimental details for the data set can be found in Markowitz *et al.* (2015). Two adult rhesus macaque monkeys (*Macaca mulatta*) were used for the study. The animals were trained for several weeks to perform an oculomotor-delayed-response or ocular-delay response task (ODR, also called memory-guided-saccade task, see Figure 3.1). The variation of the task which requires WM, which we refer to as memory task (Figure 3.1a), is a standard paradigm which has been used in many experimental works (Constantinidis *et al.*, 2001a; Funahashi *et al.*, 1989; Fuster & Alexander, 1971; Goldman-Rakic, 1988). In the memory, the monkeys must first maintain their gaze on a fixating target during 0.5 s. Then, while the monkeys fixate, a cue stimulus is presented during 0.3 s on one out of eight locations on the screen. Removal of the stimulus gives way to the delay period, which for this experiment was varied between 1 and 1.5 s. The extinction of the fixation target indicated the animal to perform a saccade (eye movement) towards the remembered cue position.

In the visual variation of the ODR task (Figure 3.1), which we refer to as visual task, the cue-stimulus is not removed from the screen until the fixation target signals the animals to respond. Including the visual variation of the ODR task is an important feature of this dataset, which makes it possible to compare the neural activity between a condition requiring working memory and a condition with similar behavioral demands that does not require working memory.

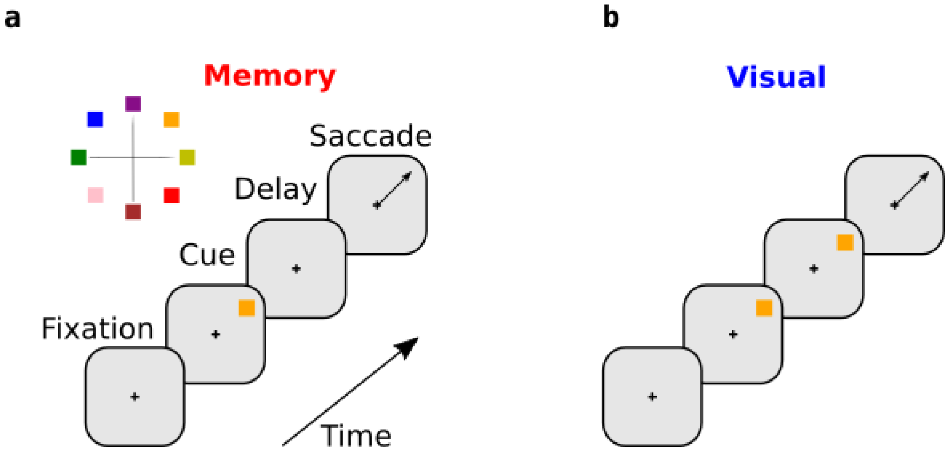

Figure 3.1: Experimental task: Memory guided saccade task and its visual variation During both tasks the monkey has to maintain fixation until the they are cued to saccade. The cartoon illustrates an example trial.

3.1.2 Recording of neural activity and spike sorting

After a surgical procedure which involved removing a part of the animals' skull (craniotomy), a 32-unit electrode (low-profile recording chamber Gray Matter Research, MT) was implanted in each monkey's right arcuate cortex. The authors recorded from isolated neurons while each monkey performed up to 500 trials of randomly interleaved memory- and visually-guided delayed saccades to one of eight targets for a liquid reward. The animals' eye position was constantly monitored with an infrared optical eye tracking system (sampling at 120 Hz).

Each recorded neuron was labeled as fast-spiking (FS) or regular-spiking (RS) according to its spike waveform. FS and RS presumably correspond to inhibitory and excitatory cells (Constantinidis & Goldman-Rakic, 2002; Gonzalez-Burgos *et al.*, 2005). The analyses shown are all conducted with the RS cells.

3.2 Data analysis

3.2.1 Single neuron analysis

The quantification of stimulus selectivity and task selectivity was done as in the original work (Markowitz *et al.*, 2015). Both procedures are described bellow.

3.2.1.1 Quantifying spatial (stimulus) selectivity

We quantified the spatial tuning of each neuron using a z-score (as in Crammond & Kalaska (1996)). Initially, we calculate each unit's firing rate across all trials for each Cue location (f_i). Next, we assigned an angular displacement, φ_i , to each Cue location and then calculated the first trigonometric moment, R_i , of each unit's response across all eight angles, as follows: Assigning an angle to each cue location (φ_i) we calculated each neuron's first trigonometric moment, R_i:

$$\begin{aligned} C &= \sum_i^8 f_i \cos(\varphi_i) \\ S &= \sum_i^8 f_i \sin(\varphi_i) \\ R_m &= \sqrt{\left(C / \sum f_i\right)^2 + \left(S / \sum f_i\right)^2} \end{aligned} \tag{3.1}$$

We shuffled cue location labels across trials to obtain a null distribution and estimated the corresponding trigonometric moment (this procedure was repeated 10^4times). We obtained the p-value as the fraction of shuffled moments that exceeded the original moment of the data, and passed it to a normal inverse cumulative distribution to obtain the tuning z-score. Neurons were considered selective if their z-score was above 1.65.

As in Markowitz *et al.* (2015), we excluded neurons with inverted tuning, whose preferred stimulus response during the delay was lower than their baseline firing rate (Zhou *et al.*, 2012).

3.2.1.2 Quantifying selectivity for task condition (three-group classification)

This procedure is analogous to the one explained above. We quantified the task selectivity of each unit's preferred target response using a permutation test.

For each neuron, we first estimated the trial-averaged firing rate in response to its preferred target during the last 300 ms of the delay of both task conditions (memory,visual). Then, we compared the actual difference in firing rates across tasks to a null difference estimate obtained by shuffling the task labels 10^4 times. The p-value (fraction of the resampled rate differences that exceeded the actual rate difference) was then passed to a normal inverse cumulative distribution function to obtain the task selectivity z-score. Units with z-score < -1.65 (1-tailed t-test) were classified as *perceptual* neurons (*early storage* neurons in Markowitz *et al.* (2015)), and units with z-score $> +1.65$ (1-tailed t-test) were classified as *mnemonic* neurons (*late storage* neurons in Markowitz *et al.* (2015)). All units with z-scores between $(-1.65, +1.65)$ were labeled as *persistent* neurons (putative *response* neurons in Markowitz *et al.* (2015)).

3.2.1.3 Single neuron exponential fits

We fitted an exponential (Equation (3.2)) function to the trial-averaged firing rates for the neurons in the *perceptual* and *mnemonic* groups during the memory.

$$A \exp\left(\frac{t}{\tau} + b\right) \tag{3.2}$$

Parameter A is associated with the neuron's maximum firing rate; τ characterizes the timing of the change in firing rate and should be compatible with (on the order of) the delay length; b is related to an offset or baseline firing rate, scaled by amplitude A. The range of the parameter A was limited between 0 and the neuron's maximum firing rate during the corresponding condition (memory task, preferred cue location). By constraining the value of A, we prevented abnormal parameter values, such as a large amplitude A and a large τ, which would make the exponential a close-to-linear and deprive the parameters of their meaning.

The class *Parameters* and function *minimize* from package *lmfit* (python) were used to perform the fits.

3.2.1.4 Analysis of intertrial variability

We were interested in the origins of the gradual (close-to-linear) increase in firing rate observed in the PSTHs of the *mnemonic* cells. Latimer *et al.* (2015) showed that such an increase (usually referred to as ramping activity) in trial-averaged responses does not necessarily arise from gradual increase in activity but that it could instead be due to instantaneous jumps in firing rate at different times on different trials. The Fano factor is defined as the ratio between the variance and the mean of a counting process. We used this measure to quantify the inter-trial variability in the data and then compared it to surrogate data to test whether or not the observed PSTHs are compatible with instantaneous jump-like activations (as the one observed in Latimer *et al.* (2015)). For each time window during the memory delay (50 ms window width), we calculated the Fano factor of each *mnemonic* neuron as the ratio between the variance and the mean of its spike count across the trials corresponding to the neuron's preferred location. We then averaged over all the *mnemonic* neurons to compare with the surrogate data.

We have a surrogate neuron for every *mnemonic* neuron ($N = 81$), and for each neuron, we simulated as many trials as the corresponding *mnemonic* neuron has for its preferred cue condition. Each surrogate trial was simulated by a step function (heaviside in numpy-python) whose minimum and maximum were selected to match the baseline and final amplitude of the corresponding *mnemonic* neuron's PSTH. The function's discontinuity was placed at a random point (uniformly distributed) during the delay time, different for each simulated trial.

3.2.2 Dimensionality reduction

3.2.2.1 Principal Component Analysis

Dimensionality reduction techniques are often applied when the size of a data set makes its interpretation challenging. Principal Component Analysis (PCA) finds new variables (dimensions in neuronal space) ordered according to the amount of variance they explain in the data. In this way, the method increases interpretability while minimizing information loss.

A formal explanation of PCA follows: The neural data can be arranged in a matrix $\mathbf{X}$ of N rows, which correspond to the number of observations and p columns, which are the (random) variables or features. In our case, N is the number of neurons and p spans all the combination of features (where every time point can be considered as a different feature, e.g. $p = \#\text{neurons} \times \#\text{tasks} \times \#\text{stimulus} \times \#\text{time samples}$). The Principal Components (PCs) are the vectors found when looking for a linear combination of the neurons' activity ($\sum_{i=1}^{N} a_i \mathbf{x}^\top = \mathbf{X}^\top \mathbf{a}$) with maximum variance, where $\mathbf{x}$ are the row vectors of $\mathbf{X}$ (containing each neuron's activity for each combination of features). Such a linear combination has a variance given by ($\text{var}\left(\mathbf{X}^\top \mathbf{a}\right) = \mathbf{a}^\top \mathbf{S}\mathbf{a}$), where $\mathbf{S}$ is the covariance matrix of the data. If we consider a centralized matrix $\mathbf{X}$, such that $\left\langle \mathbf{X}_i^\top \right\rangle = 0 \quad$ for $i = 1 \dots N$ (which we can always do if we subtract for each neuron its activity average over the p features) then covariance matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$. Identifying the linear combination with maximum variance is then equivalent to obtaining a $n-$dimensional vector $\mathbf{a}$ which maximizes the quadratic form $\mathbf{a}^\top \mathbf{S}\mathbf{a}$. For the problem to have a well-defined solution, the norm of $\mathbf{a}$ has to be restricted to an arbitrary number. The problem is then equivalent to maximizing $\mathbf{a}'\mathbf{S}\mathbf{s} - (\mathbf{a}'\mathbf{a} - 1)\lambda$ where λ is a Lagrange multiplier. Differentiating with respect to the vector $\mathbf{a}$ we get the equation

$$\mathbf{S}\mathbf{a} = \lambda \mathbf{a} \tag{3.3}$$

From Equation (3.3) it is clear that $\mathbf{a}$ is an eigenvector and λ an eigenvalue of the covariance matrix $\mathbf{S}$. The vector $\mathbf{a}$ with maximum variance will be the one associated with the largest eigenvalue of $\mathbf{S}$ [1]. It can be shown (Bartholomew, 2010) that the vectors which successively maximize the variance while being mutually orthogonal is the full set of eigenvectors of $\mathbf{S}$. The Principal Components of the data matrix $\mathbf{X}$ are thus the eigenvectors of its covariance matrix.

It can be shown (Jolliffe *et al.*, 2016) that the eigenvectors of $\mathbf{S}$ (the PCs) can be obtained via the single value decomposition of $\mathbf{X}$. This method is used by the python function we applied to our data (see below).

The PCs can alternatively be obtained as a result of a minimization problem. We introduce this derivation because it can be naturally extended to the targeted PCA (demixed-PCA, dPCA) method which we explain in the following section. For the same centered data matrix $\mathbf{X}$ a set of $q \leq N$ PCs is the best linear approximation to the data in a $q-$dimensional space. (see et. all. Hastie (2009); Kobak *et al.* (2016)). In other words, the PCs are obtained when minimizing the following cost function

$$L_{PCA} = \left\| \mathbf{X} - \mathbf{D}^\top \mathbf{D}\mathbf{X} \right\|^2 \tag{3.4}$$

where the PCs are the rows of the rank-q matrix $\mathbf{D}$. Matrix $\mathbf{D}$ can be viewed as a decoding matrix which compresses the data $\mathbf{X}$ into a space of dimension q. The transpose $\mathbf{D}$ can

[1] $\mathbf{S}$ is semi-defined positive its spectrum is equal or greater than zero.

be understood as an encoding matrix which, in turn decompresses the activity from the q-dimensional space back to the n-dimensional space, approximately reconstructing the original data.

We use the class *decomposition* from python package *sklearn* to apply PCA on our data. Initially, the matrix of the data has dimensions $(N, K, C, T-$#neurons,#tasks,#stimulus,#time samples$)$ To obtain the PCA decomposition we first flatten the matrix to a 2-dimensional matrix (The columns of the flattened matrix span all possible (p) combinations of task features, as mentioned above). Then we apply the method *PCA.fit* which outputs the PCs. To interpret the results, for each combination of stimulus location and task type $((C = 8) \times (K = 2) = 16$ combinations), we project the time course of the neural activity $(N \times T)$ on the space of PCs. For visualization, PCs can be plotted respectively against time or against one another. In the combined PC plots, the trajectories are parametrized by the time.

3.2.2.2 demixed-Principal Component Analysis

For a detailed explanation of demixed-Principal Component Analysis (dPCA) with illustrating examples see Kobak *et al.* (2016). Similarly to PCA, dPCA finds linear combinations of the neurons' activity that contain relevant information about the experimental task. The difference between both methods is that dPCA finds directions in neuronal space with maximal variance due to specific task features (or to combination of task features). In this way, dPCA disentangles the contribution of different features to the total variance.

dPCA introduces two changes in Equation (3.4). First, instead of requiring the compression and decompression matrices to reconstruct the neural activity directly, it separately favors the reconstruction of neural activity related to a certain feature or combination of features. Thus, to obtain dPCs related to task feature ϕ, the compression and decompression would be required to minimize the error with respect to matrix $\mathbf{X}_\phi$, from which the average over every other task feature has been subtracted $\mathbf{X}_\phi = \mathbf{X} - \sum_{\phi' \neq \phi} \langle X \rangle_{\phi'}$, where ϕ are the task features (in our case: task condition, stimulus and time). The second modification with respect to PCA is that in order to gain a more flexible mapping, the decoding and encoding matrices are allowed to be different. The result is the following cost function,

$$L_{dPCA} = \sum_{\phi} L_{dPCA,\phi} = \sum_{\phi} \left\| \mathbf{X}_\phi - \mathbf{F}_\phi \mathbf{D}_\phi \mathbf{X} \right\|^2 \tag{3.5}$$

Where $\mathbf{F}_\phi$ and $\mathbf{D}_\phi$ are the encoding and decoding matrices associated with feature ϕ. Each term in the sum can be minimized separately. Minimizing each $\mathbf{L}_\phi$ constitutes a *reduced rank regression* problem, which can be solved analytically (see Material and Methods in Kobak *et al.* (2016)). We used the package written in Matlab, available at `http://github.com/machenslab/dPCA`.

3.2.3 Decoding stimulus information from PFC population activity

Classification algorithms that can decode the content of neural representations are frequently used to interpret data in cognitive and systems neuroscience (Henderson *et al.*, 2021; Kim *et al.*, 2016; King & Dehaene, 2014; Stokes, 2015; Stokes *et al.*, 2013; Wolff *et al.*, 2015). They are particularly useful to identify the neural representations at the level of neural population

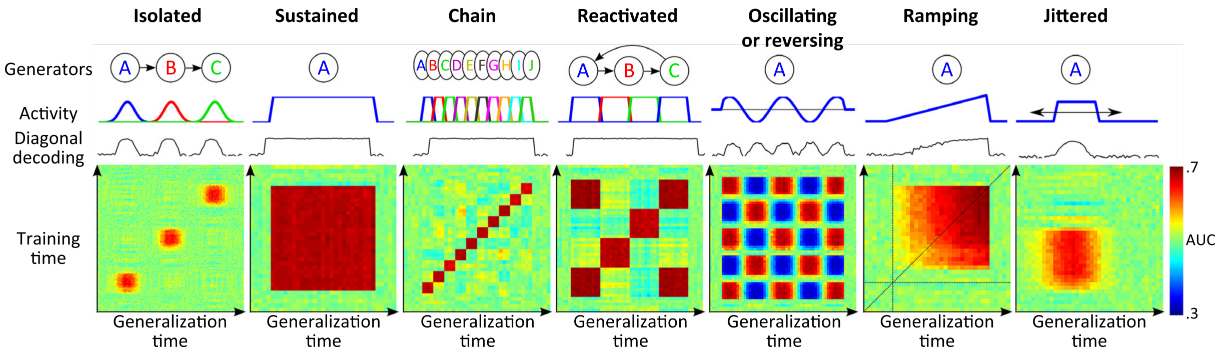

Figure 3.2: Cross-temporal decoding patterns reflect qualitative properties of the neural code: Cartoon examples, from left to right: separate groups of neurons firing at different sparse times, sustained (persistent) firing, chain of activation tiling the length of the trial, sequence with reactivations, oscillating signal, gradual activity increase (ramping), the effect of jitter (different latencies for different trials)

activity. In general, a decoding analysis involves training an algorithm with a subset of the neural data and then testing its performance on the remaining portion of the data.

Here, we used such a population decoding approach was used to quantify stimulus information at the population level. We trained and tested support vector machine (SVM) classifiers on surrogate populations of neurons (pseudopopulations) constructed from the individual neuron recordings (Sarma *et al.*, 2016). In the pseudopopulations we included all neurons with more than 10 correct trials per condition (per task condition and stimulus combination). The spike count window was T = 50 ms. For each neuron, the spike counts were normalized to zero mean and unit variance across all trials and time bins.

As classifiers, we used linear multi-class SVMs, C-SVC from LIBSVM (Chang & Lin, 2011). Nonlinear SVMs with Radial Basis Function (RBF) kernels yielded the same results, so we used linear SVMs throughout the analysis. To estimate the decoder performance, we applied leave-one-out cross-validation. During each of 10000 repeats, we selected a different set of 10 trials for each cue location for each neuron. Next, we randomly assigned 9 of the 10 trials to the training set and the remaining trial to the test set. This gave us a pseudopopulation response with 72 training trials and 8 test trials. Overall performance was computed by averaging test performance across all repetitions.

3.2.4 Cross-temporal population decoding

The stability of cue decoding was quantified by comparing decoding accuracy throughout the trial. In particular, we used *cross-temporal decoding*, that is, we trained the decoder at one time point and then tested it on data from every time point in the trial. This allowed us to measure how a trained decoder generalizes across time and from a given experimental condition to another (King & Dehaene, 2014). As explained in detail in King & Dehaene (2014), testing how the decoder generalizes across times can reveal or reflect qualitative properties of the neural code (see Figure 3.2). The fundamental hallmark across the different types of code is that a trained decoder can generalize across times (off-diagonal points in the performance matrix) when the neural code remains stable. When the neural code changes significantly (dynamic code), so does the decoder's performance.

In our data set (Markowitz *et al.*, 2015), and other data sets related to similar WM paradigms (Constantinidis *et al.*, 2001a; Mendoza-Halliday & Martinez-Trujillo, 2017; Spaak *et al.*, 2017; Stokes *et al.*, 2013), the cross-temporal decoding pattern reveals a lack of generalization between cue and delay epochs (see Figure 4.2). It also suggests the presence of ramping activity (see Figure 3.2) during the delay.

3.2.5 Measuring clustering in the space of features with the ePAIRS algorithm.

Representing neurons in feature space can be useful for understanding computations, structure, and the distribution of selectivity. For example, by visualizing neurons in feature space, we can identify clusters or groups of neurons that respond similarly to specific features or stimuli. These clusters could correspond to different subpopulations of neurons that perform specialized functions within a larger neural network.

To identify these clusters, we can use clustering algorithms to group neurons that share similar feature selectivity. Since neurons with similar feature selectivity (forming clusters), will lie close to each other when represented on the space of features, a way to assess the presence of these clusters is to measure the distribution of nearest neighbor angles in the feature space.

Inspired in recent works (Dubreuil *et al.*, 2022; Hirokawa *et al.*, 2019; Raposo *et al.*, 2014; Yang *et al.*, 2022), we used the "elliptical projection angle index of response similarity " (ePAIRS) to measure the presence of clusters in our data set. This test is a variation of the "projection angle index of response similarity "'(PAIRS) method which was first used in Raposo *et al.* (2014). A formal explanation follows.

Before applying the ePAIRS test, we reduced the dimensionality of our data (using either PCA or dPCA). This step is important to filter out the noise and irrelevant variance in the data and to obtain a set of feature dimensions we can interpret. . Coherent with our previous notation (see subsection on dPCA), we consider the matrix $\mathbf{D}$ whose N rows correspond to the number of considered neurons and q columns, the number of features (PCs or dPCs). Considering each row in $\mathbf{D}$ as a point in the q-dimensional feature space, the ePAIRS test compares the original distribution of directions $\mathbf{d}_i/\|\mathbf{d}_i\|$ to a null distribution which is obtained by bootstrapping from a multivariate Gaussian with the same covariance ($\Sigma = \frac{1}{N}\mathbf{D}^\top\mathbf{D}$) as the original data. The algorithm is described by the following steps:

1 for each point $\mathbf{d}_i$ its k nearest neighbors as found (the points which maximize $\cos\widehat{\mathbf{d}_i\mathbf{d}_j}$). We have used $k = 13$ for our computations, smaller values of k do not alter the results qualitatively.

2 The mean α_i angle is computed for each neuron, defining the empirical distribution $\hat{p}_{data}(\alpha)$.

3 Steps 1 and 2 are applied $N_{bootstrap}$ times (we used $N_{bootstrap} = 10000$) to random distributions of N q-dimensional points, sampled from a multi-variate Gaussian $N(0,\Sigma)$, to obtain the null distribution $\hat{p}_{null}(\alpha)$

4 Finally, the original and null distributions are compared using a two-sided Wilcoxon rank-sum test, which yields a p value. The effect size is computed as

$$c = \frac{\mu_{null} - \mu_{data}}{\mu_{null}} \tag{3.6}$$

- $c > 0$ means smaller distance between neurons than expected by chance, i.e. structure or clusters.
- $c < 0$ means greater distance than expected by chance, which would correspond to neurons evenly space in feature space.

Figure 3.3 illustrates how the ePAIRS method distinguishes between a case where the selectivity between two features is randomly mixed (Figure 3.3a) and a case where the selectivity is clearly non-randomly mixed (Figure 3.3b). The original distribution of nearest neighbor distances is overlapping with the null distribution when the distribution of selectivity is random (Figure 3.3c) and significantly different from the null when the selectivity is non-randomly distributed (Figure 3.3d). The difference between the ePAIRS and the PAIRS methods is that in ePAIRS, the null distribution are generated by bootstrapping from multivariate Gaussian distribution *with the same covariance* present in the data, while the PAIRS method samples from a spherical Gaussian distribution (the covariance matrix is diagonal with all elements in the diagonal having the same value). This difference can cause false positive results of the PAIRS method, when apply to data with different variance along different features (which is usually the case, see Hirokawa *et al.* (2019) for examples).

3.2.6 Combined dPCs for 2-feature comparisons

To get a finer-detailed view of how the neurons' selectivity is distributed among different task features, we measured each neuron's contribution to different pairs of activity modes. We used the same approach applied in Yang *et al.* (2022) to assess the distribution of feature selectivity in mouse ALM region. Each neuron's contribution to a pair of features is represented as a 2-dimensional vector in the space spanned by the corresponding features (Figure 3.4a). The angle this vector forms with one of the axes is taken as a measure of how the neuron's selectivity is distributed between the features. For every pair of features, the original distribution of angles is compared to a null distribution generated by bootstrapping over multivariate Gaussian distributions with data-matched variance.

The feature-specific dimensions used in Yang *et al.* (2022) were obtained directly as linear combinations in the space spanned by the neurons. They could define linear feature-specific dimensions because their task had 2 possible stimulus and 2 possible choices, so the corresponding stimulus and choice dimensions were 1-dimensional. For the analysis of the contribution to pairs of features, the authors then used the absolute value of components or weights of each neurons along the corresponding feature vectors. For our analysis, we combined some of the feature-specific dPCs to obtain meaningful feature-specific dimensions. In particular, we combined the first two stimulus-dPCs because they provide complementary information related to variance along orthogonal directions on the plane where the stimuli were presented (see Figure 4.11). In a similar way, we combined the first two time-dPCs, because they alone explained a substantial amount of the variance (14.5% and 7.3%). In both cases, the components of the combined dimensions are obtained as euclidean norms:

$$\begin{aligned} w_{\text{stimulus},i} &= \sqrt{w^2_{\text{stimulus},1,i} + w^2_{\text{stimulus},2,i}} \qquad i = 1\dots N \\ w_{\text{time},i} &= \sqrt{w^2_{\text{time},1,i} + w^2_{\text{time},2,i} + w^2_{\text{time},3,i}} \qquad i = 1\dots N \end{aligned} \tag{3.7}$$

As a consequence of combining (non-linearly) dPCs to obtain the feature-specific vectors,

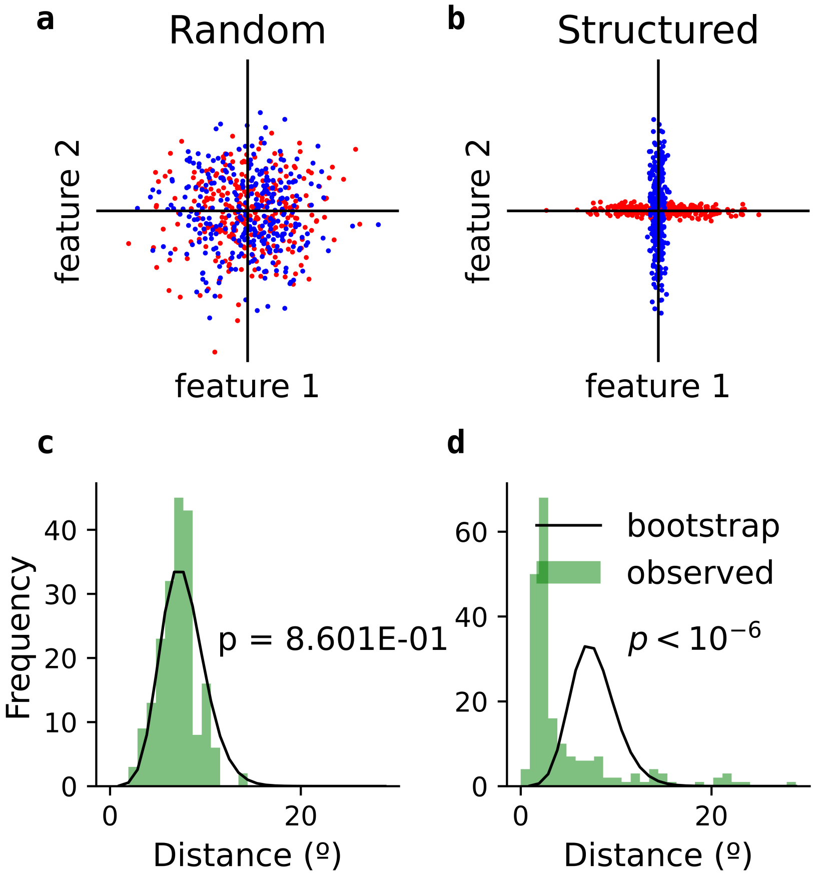

Figure 3.3: ePAIRS identifies non-random structure based on the distribution of nearest neighbor angles Extreme cartoon examples to illustrate the method: Random mixed (**a**) and non-random mixed (**b**) selectivity for features 1 and 2. **c,d** original distributions of nearest neighbor distances (green) and null distribution generated with the ePAIRS algorithm (black).

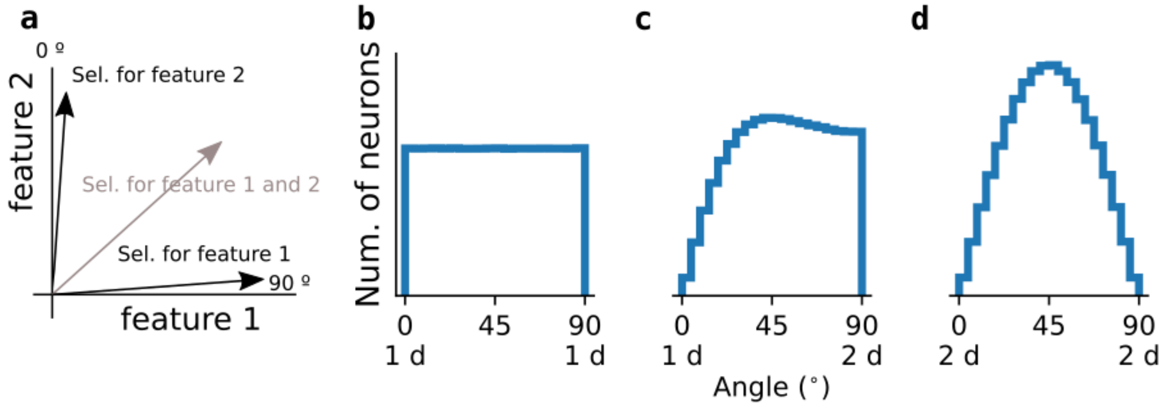

Figure 3.4: Random distribution of selectivity between two dimensions: a example selectivity of three cartoon neurons. **b-d** Null distribution of angles for a surrogate population of 1000 neurons for different combinations of feature dimensions. **b**: both dimensions are the absolute value of weights obtained as random samples from the same distribution. **c**: one dimension is obtained as in **b**, the other dimension, as the square root of the sum of two squared weights, each obtained from as samples of the same random distribution. **d**: both dimensions are obtained as the second dimension in **c**.

the null distributions we obtain as controls are not flat. We show examples of null distributions of surrogate data in Figure 3.4.

3.2.7 Classification according to dPC weights

The components or weights of the dPCs provide a measure to classify the neurons. The neural trajectories along the 1st task-dPC (the dPC which explains the highest amount of variance due to the difference in task condition) bifurcate according to the task condition (memory, visual) after the time when the stimulus is removed in the memory task. Therefore, we used the weights on this dPC as a measure of how much the neurons are modulated by the task condition. To obtain a three-group classification, since the distribution of weights in this dPC is unimodal (Figure 4.14), we split the neurons according to different percentile values: below the 25th percentile, above the 75th percentile and between the 25th and 75th percentile. The three populations we obtain are similar to those in Markowitz *et al.* (2015).

3.3 Models

3.3.1 Spiking neuron model: leaky-integrate and fire

The leaky-integrate and fire (LIF) neuronal model has been used extensively to model neural networks. We have chosen this model because it offers a good compromise between biological plausibility and computational efficiency.

The LIF model was inspired by the work of the French physiologist Louis Lapicque (Brunel & Van Rossum, 2007; Lapicque, 2007), but it was first presented as we know it today by B. Knight (Knight, 1972). The basic feature of the model is that it describes the subthreshold voltage integration but not the rapid voltage dynamics of the spike generation. This is justified by the two processes happening at clearly different time scales and because the voltage exchange during a spike is highly stereotypical (Burkitt, 2006). When the membrane potential v reaches the established threshold for generating action potential, a spike is modeled by setting v to a given reset value v_r.

The subthreshold voltage integration is model as a function of the currents flowing to the neuron's synaptic channels ($I_{\alpha,i}$ with $\alpha = \text{AMPA}, \text{GABA}, \text{NMDA}\ldots$) and a passive leakage current I_L which effectively tends to bring v to its resting value:

$$C\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = \sum_{\alpha} I_{\alpha,i}(t) + I_{L,i}(t) \tag{3.8}$$

In the *conductance-based* LIF models the synaptic currents are described by Equation (3.9)[2].

$$I_{\alpha,i} = \left(v_i - E_\alpha\right)\sum_j g_{\alpha,ij}s_{\alpha,j}(t) \tag{3.9}$$

The voltages E_i are constant and are referred to as reversal potentials. They are related to the equilibrium potential of the different ion channels. Whenever v crosses the value of a given reversal potential, the synaptic current flow associated to this channels switches. The value of

[2] For *current-based* models see e.g. Burkitt (2006); Cavallari *et al.* (2014).

the conductances $g_{i,j}$ will depend in general on the type of synapse (e.g. AMPA, GABA or NMDA) and on the presynaptic and postsynaptic pair of neurons. The variables $s_{\alpha,i}$ are called gating variables and they are related to the amount of neurotransmitter which is available at the synaptic cleft (or to the fraction of synaptic channels that are open). A common way of describing the dynamic of a gating variable (Compte *et al.*, 2000; Wang, 1999) is

$$\frac{\mathrm{d}s_{\alpha,i}}{\mathrm{d}t} = -\frac{1}{\tau_\alpha} s_{\alpha-i} + w_\alpha \sum_k \delta(t - t_k) \tag{3.10}$$

where τ_α is a decaying time constant specific to each type of synapse and t_k are the spiking times of the presynaptic neurons. The second term in the right-hand side of Equation (3.10) is Dirac's delta distribution, which represents that any presynaptic spike at time t_k will increase s by an amount w_k.

We have used Equation (3.10) for AMPA and GABA synapses. For NMDA we included an additional gating variable x

$$\frac{\mathrm{d}s_{\mathsf{NMDA},i}}{\mathrm{d}t} = -\frac{1}{\tau_{\mathsf{NMDA}}} s_{\mathsf{NMDA}} + \alpha_s(1-s) \quad \frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{1}{\tau_x} x + \sum_k \delta(t - t_k) \tag{3.11}$$

additionally, the NMDA conductance is modeled as voltage dependent (Compte *et al.*, 2000; Jahr *et al.*, 1990).

$$g_{\mathsf{NMDA}}(v_i) = g_{\mathsf{NMDA},0} \frac{1}{(1 + \exp(-0.062 v_i)/3.57)} \tag{3.12}$$

where $g_{\mathsf{NMDA},0}$ is parameter which we vary for different simulations. The temporal dynamics of NMDA, slower than the other synaptic variables, are crucial to generate persistent firing rates within a physiological range $(10-50\ Hz)$ (Wang, 1999). Without the stabilization generated by the NMDA dynamics, strong recurrent excitation (see section below and Wang (1999)) would produce persistent firing rates above the physiological observed values and for a narrower range of background inputs.

All the synaptic currents we use are implemented as in Compte *et al.* (2000).

In the general formulation of the LIF model, an additional term I_{inj} is sometimes included on the right-hand side of Equation (3.8) to account for currents injected directly into the neurons' soma by artificial means. We have modeled the effect of the stimulus (light cues on a screen) on the frontal circuits as an injected pulse with a Gaussian profile. (Gaussian in the space of neurons)

$$\begin{aligned} I_{\mathsf{inj},i} &= I_s \exp\left(\frac{(\theta_i - \theta_{\mathsf{stim}})^2}{2\sigma^2}\right) \quad \text{when stimulus on} \\ I_{\mathsf{inj},i} &= 0 \quad \text{when stimulus off} \end{aligned} \tag{3.13}$$

where I_s is a parameter we varied across different simulations (in a range between 10 and 100 pA), σ is the width of the pulse (we used $42\,\mathrm{deg}$), θ_{stim} is the angle at which the stimulus is presented and θ_i is the preferred orientation of neuron i. In the actual experiments (Markowitz *et al.*, 2015), the stimulus cues are illuminated dots on a screen. Since the cues appear at a fixed distance from the fixation point, we take the angle of the stimulus as the relevant variable. From a computational perspective, what matters is that we are modeling a continuous period variable.

By modeling the stimulus as an injected current we are not considering the time varying, voltage dependent processes related to stimulus related synaptic stimulation. Stimulation has been modeled in a similar way in other works (Compte *et al.*, 2000) (cite more)

3.3.2 Firing rate model

For some simulations, we used a mean-field model (Wilson & Cowan, 1972; Wong & Wang, 2006) whose behavior is qualitatively equivalent to that of a spiking neural network. Instead of describing a neuron's spike timing and membrane voltage, this model describes the neuron's firing rate, which is the number of action potentials generated during a finite time window. The advantage of this model is its simplicity, which allows for much faster simulations.

To reduce the spiking network model to the firing rate model we use requires several assumptions. For a more detailed explanation, see Wong & Wang (2006). We briefly outline the most important steps. The first simplification is related to the mean-field approximation, introduced originally and explained in detail in Wilson & Cowan (1972). In this approximation, the non-linear dynamics associated with the action potentials are accounted for by a transfer function $\phi(i)$, which describes the relation between the currents a neuron receives and its electric activity (input/output function). We use the function used in Wong & Wang (2006):

$$\phi(I) = \frac{aI - b}{1 - \exp\big(-d(aI - b)\big)} \tag{3.14}$$

where a, b, d are constant parameters (see Wong & Wang (2006)).

As shown in Wilson & Cowan (1972), the dynamics of the firing rate variable r can then be written in terms of ϕ as:

$$\tau_r \frac{\mathrm{d}r_i}{\mathrm{d}t} = -r_i + \phi(I_{syn,i}) \tag{3.15}$$

where τ_r is related to the time a neuron needs to integrate its inputs.

Conceptually, the rate variable r represents the firing rates of a *spatially localized neural population* (Wilson & Cowan, 1972). As is often done in the literature (Engel *et al.*, 2015; Murray *et al.*, 2017b; Wimmer *et al.*, 2015) we assume for our simulations that each rate variable r_i represents a unit i with preferred location θ_i, whether this unit is composed by one or several cells is not relevant for the model.

The model in Wong & Wang (2006) initially considers the total synaptic input I_{syn} as the sum of the same three types of synaptic currents we used for the LIF model (spiking neural model); AMPA, GABA and NMDA. Since the membrane voltage is not described by this model, the currents depend only on the value of the conductance g and of the gating variables s:

$$I_{syn,i} = g_{\alpha,ij} s_\alpha \quad \text{with } \alpha = \text{AMPA, GABA, NMDA} \tag{3.16}$$

Where the dynamics of the gating variables is given by:

$$\begin{aligned}\frac{\mathrm{d}s_{\mathrm{AMPA},i}}{\mathrm{d}t} &= -\frac{s_{\mathrm{AMPA}}}{\tau_{\mathrm{AMPA}}} + r_i \\ \frac{\mathrm{d}s_{\mathrm{GABA},i}}{\mathrm{d}t} &= -\frac{s_{\mathrm{GABA}}}{\tau_{\mathrm{GABA}}} + r_I\end{aligned} \tag{3.17}$$

where r_I is the rate of the inhibitory population, which the model considers initially. The dynamics for the NMDA variable are given by:

$$\frac{\mathrm{d}s_{\mathrm{NMDA}}}{\mathrm{d}t} = -\frac{s_{\mathrm{NMDA}}}{\tau_{\mathrm{NMDA}}} + (1 - s_{\mathrm{NMDA}})\gamma r_i \tag{3.18}$$

Where the effect of the two time constants contributing to the NMDA dynamics in the LIF model is now effectively included in Equation (3.18), see Wong & Wang (2006) for a derivation of Equation (3.18).

At this stage, a further simplification is made by considering that the dynamics of the NMDA gating variable are much slower than the dynamics of the other AMPA and GABA gating variables and the neuron integration dynamics ($\tau_{\mathrm{NMDA}} \gg \tau_{\mathrm{AMPA}}, \tau_{\mathrm{GABA}}, \tau_r$). This implies that the only the steady-state values of the variables $r, s_{\mathrm{AMPA}}, s_{\mathrm{GABA}}$ are used, and only the differential equation for s_{NMDA} (Equation (3.18)). In the original formulation in Wong & Wang (2006), the authors considered initially a network with two excitatory populations, which receive different inputs, and a non-selective inhibitory population. After the simplifications, the effect of the inhibition is eventually included as negative contributions across the excitatory populations.

The original formulation in Wong & Wang (2006) based on the analogous LIF model (Wang XJ, 2002)) describes a 2-unit system in the context of 2-alternative decision making. For both the LIF and the rate model, the original formulation is naturally extended to a system with selectivity for a continuous periodic feature, such as orientation, by introducing the appropriate shape of connectivity weights g_{ij}. We will discuss below how a positive recurrence among neurons with similar selectivity can generate selective and persistent activity in a network.

3.3.3 Connectivity of the ring-attractor model

Neuronal selective behavior to external stimulation is accomplished by synaptic connectivities of the center-surround type. In this type of connectivity, also referred to as lateral inhibition, excitatory connections dominate over inhibitory connections only between cells with similar selectivity. For cells with sufficiently distinct selectivity, inhibition is stronger. [3] This synaptic configuration in some parts of the nervous system allows us to resolve fine differences in the sensory space (e.g. nearby locations, nearby colors).

The center surround connectivity is usually described as mexican hat-type because of its shape Figure 3.5 For periodic stimulus such as orientations, or as in our case, locations associated with orientations (dots on a circle), the connectivity is defined as a function of the difference between the preferred direction of cells ($\Delta\theta = \theta_i - \theta_j$), and is in this way rotationally

[3] In sensory networks, the anatomic arrangement of the cells matches the topology of the system, such that cells with nearby preferred stimuli are also adjacent. This, however, is not necessarily the case for more upstream networks where the anatomy of the circuits does not parallel or reflect the structure of the connections.

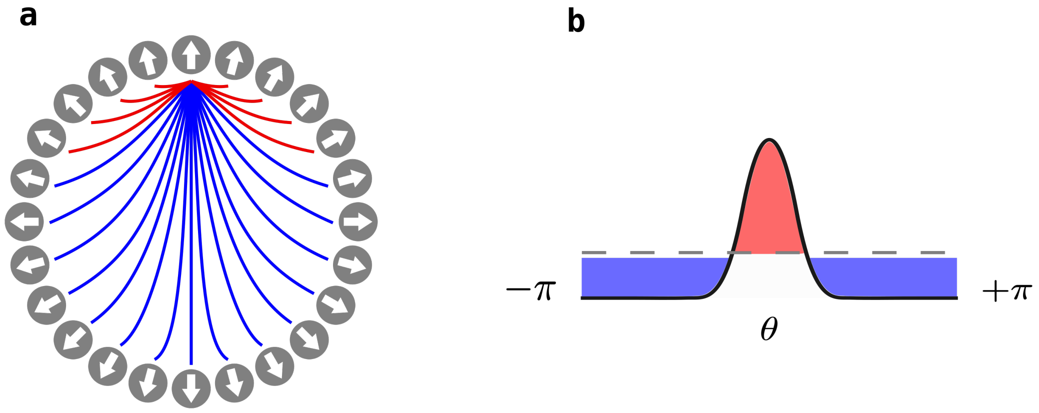

Figure 3.5: Topology and connectivity profile of the ring-attractor network: **a** Ring-network topology. Neurons excite other neurons with similar selectivity (nearest neighbors in the ring, excitation highlighted in red) and inhibit the rest of the network (inhibition highlighted in blue). **b** Connectivity profile. Amplitude of the effective connections as a function of the difference in selectivity (distance on the ring) Excitation, red. Inhibition, blue

invariant.

$$g_{ij} = J_m + J_p \exp\left(\frac{(\theta_i - \theta_j)^2}{2\sigma^2}\right) \tag{3.19}$$

In the firing rate model, the connectivity describes the connection between excitatory neurons (inhibitory neurons are not included in the model). Inhibition is implemented through the negative values of $g_{i,j}$. In the LIF model, excitation and inhibition are implemented by different populations of neurons. We include a structured profile as the one in Figure 3.5 for the connections between excitatory neurons (with all values of $g_{E,E} > 0$). The rest of the connections $(E \to I, I \to E, I \to I)$ we model as flat. The inhibition and excitation combined, work as the effective profile in Figure 3.5.

For the spiking neuron model, we used the structured connectivity profile (Equation (3.19)) for the connection between excitatory neurons, we used flat profiles for the rest of the connections.

A network with a connectivity profile as shown in Figure 3.5 can exhibit different qualitative behaviors which depend largely on the strength of the excitatory connections. They are summarized in Figure 3.6. When the amplitude of the excitatory connections is high enough, the network has attractor dynamics, and it responds during stimulation forming a bump of activity centered on the neuron which has been maximally stimulated. Importantly, the shape of the bump is determined in the first place by the intrinsic network dynamics Once the stimulation ceases, the activity bump will fade, and the network will return to its baseline activity. In this dynamic regime, the bump state is reached through a super-critic Hopf bifuraction and the system is accordingly monostable. This regime can describe transient responses to stimulus, such as the ones observed in sensory neurons or some of our PFC neurons. We used this regime for some of our subnetworks (see details below). When the amplitude of the excitatory connections is further increased, the network can reach a bistable regime, where the activity bump can be sustained without external stimulation. The bifurcation through which the bump is formed in this regime is sub-critical. The firing rates or spiking activity of the neurons

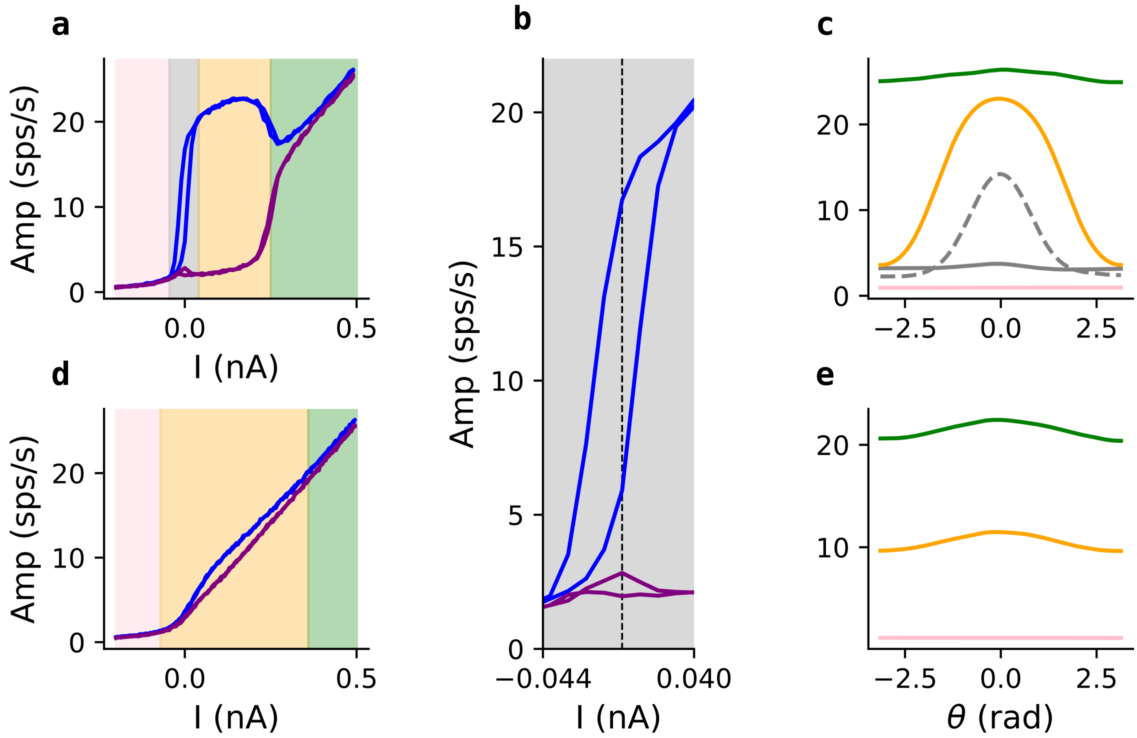

Figure 3.6: Different activity regimes of a ring network with lateral inhibition Regime with bistability (**a**-**c**); **a**: bifurcation plot. maximum (blue) and minimum (purple) firing rate amplitude for different values of the bifurcation parameter (non-selective injected current). Regions with different stable solutions highlighted in **a**, homogeneous low activity state (no bump) (pink), bistable region (gray), non-homogeneous bump state (yellow) and homogeneous high activity state (green). **b**: (inset) bistable region in **a**. **c**: example network activity profiles (smoothed with a Gaussian filter, $\sigma = 20$) for the different regions highlighted in **a**. Regime without bistability **d**,**e**. **d** bifurcation plot homologous to **a**, without bistable region. **e** same as **c**.

during stimulus presentation is very similar in both the transient and the persistent regimes, However, a network with bistable behavior can sustain the intrinsic bump of activity even after the stimulus is removed. This behavior has been used before to model WM in frontal circuits (Compte *et al.*, 2000; Wimmer *et al.*, 2014)

For the firing rate model, we do not model separately excitatory and inhibitory neurons. The effect of inhibition is included effectively by the negative part of the connectivity ($J_m < 0$ in Equation (3.19)).

3.3.4 Connectivity between ring-networks

Our model can reproduce some crucial aspects of the experimental data due to the usage of different stability regimes (monostable and bistable) for different sub-networks and to the way these sub-networks are connected. We designed the inter-ring connectivity structure in such a way that we would obtain the differential behavior of the populations in Markowitz *et al.* (2015) (Figure 4.5). Qualitatively speaking, we need the following stimulus-selective behaviors:

- A network which responds in the presence of external stimulation (labeled as *perceptual* neurons here)
- A network which fires during the memory-demanding period (the delay in the memory) but not otherwise
- A network which fires from stimulus presentation until motor response, regardless of whether the stimulus remains on the screen or not.

As one can anticipate, different inter-ring connectivities can satisfy these conditions. We explored different types of connectivities, which will be described in the following. Common to all the sub-network to sub-network connectivities is a structured synaptic footprint as Equation (3.19) where the subindexes i, j belong to different sub-networks. This connectivity ensures that neurons in one ring are maximally connected to neurons in another ring that have the same stimulus selectivity.

3.3.5 Ramping input

The ramping signal is implemented as a linear change in the synaptic conductances mediated by NMDA (g_{NMDA}). The steepness of the ramping signal is controlled by the parameter r_γ:

$$\begin{aligned} g_{\mathrm{NMDA}}(t) &= \gamma(t) g_{\mathrm{NMDA}} \\ \gamma(t) &= r_\gamma t \end{aligned} \tag{3.20}$$

Note that this modulation affects both excitatory and inhibitory neurons in the *storage* network.

Chapter 4

Results

4.1 Structure and functional specialization in the prefrontal cortex

Section summary

In this section, we combine different analysis techniques to investigate how selectivity for different task features during working memory is distributed among prefrontal neurons. First, we illustrate the diversity of the single-neuron activity profiles and the changes between cue and delay epochs in the population code. Classifying the stimulus-selective neurons based on their selectivity for the task condition (following the approach in (Markowitz *et al.*, 2015)) reveals different stereotypical activity profiles. However, assessing the statistical differences between the stereotypical profiles is difficult because the population averages mask single-neuron variability. To assess the presence of functional groups unbiasedly, we analyze the activity of all excitatory neurons using dimensionality reduction and clustering algorithms. We observed a significant level of clustering in the space spanned by the demixed-Principal Components, which indicates a non-random distribution of feature selectivity. Finally, we relate the overall non-random feature selectivity distribution to the specific task features: time, stimulus location, and task condition. We obtain results compatible with the single neuron analysis and show that the different stereotypical or functional profiles contribute to the working memory code.

4.1.1 Diverse activity profiles among prefrontal neurons

The prefrontal neurons engaged during WM tasks exhibit different types of activation profiles. Some cells (as the one in Figure 4.1a) exhibit a type of activity known as persistent (see Chapter 1). Persistent neurons respond selectively to the presentation of a cue stimulus and sustain their increased firing rate throughout the memory delay. As explained in the Chapter 1, persistent activity is regarded as one of the main correlates of short-term memory maintenance in the cortex (Funahashi *et al.*, 1989; Funahashi, 2017; Fuster & Alexander, 1971), and it has been modeled successfully (Camperi & Wang, 1998; Compte *et al.*, 2000; Murray *et al.*, 2017a; Wimmer *et al.*, 2014). However, there exist a variety of different neuronal response profiles in prefrontal cortex (PFC), whose relevance for WM is less understood: In a way that resembles the activity of neurons in sensory regions, some neurons respond transiently to a stimulus (as the example in Figure 4.1b), returning to baseline firing once the stimulus is removed. Other neurons (as the example unit in Figure 4.1c) undergo gradual increases in firing rate during the delay period. This type of response profile has been observed in several studies (Constantinidis

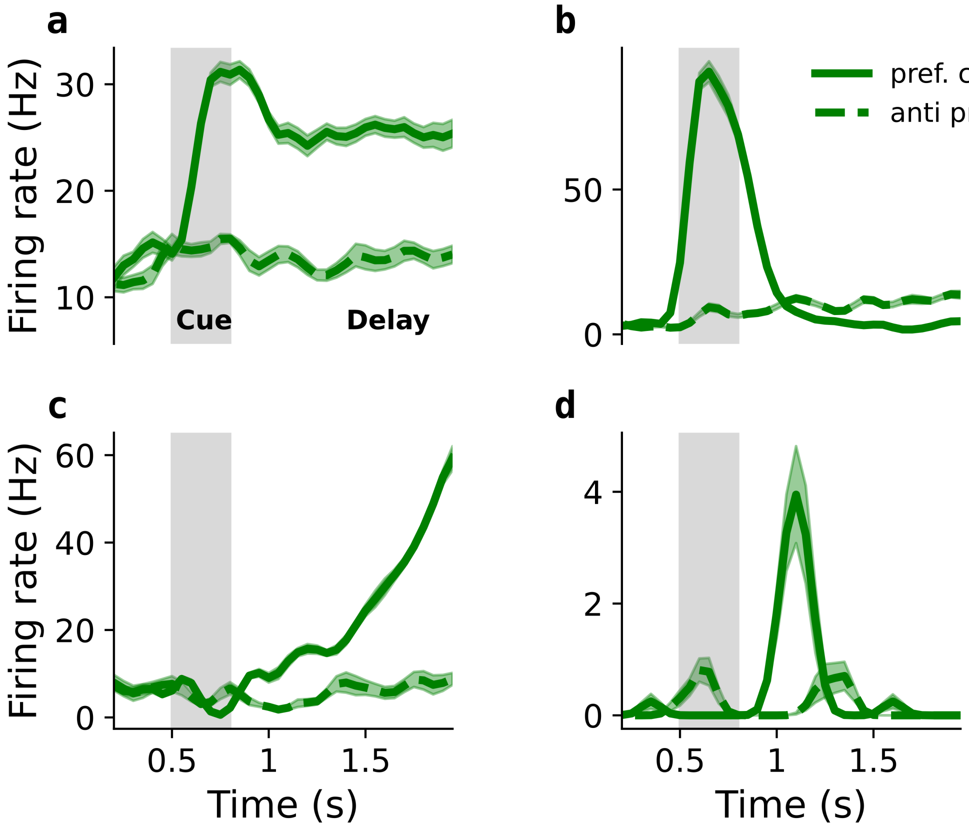

Figure 4.1: The activity profiles of the single PFC neurons involved in a WM task are diverse. PSTHs of four example neurons for trials at which the stimulus is at the neuron's preferred location (solid) and anti-preferred location (180° from preferred location, dashed). All neurons shown are selective to stimulus. **a** Persistent type, firing during cue and delay. **b** transient activation during stimulus presentation. **c** increasing ramping activity during the delay. **d** transient activation during early delay.

et al., 2001b; Finkelstein *et al.*, 2021; Funahashi *et al.*, 1989; Inagaki *et al.*, 2019) which involve memory delays of stereotyped length, and it is thought to be a correlate of elapsed time or response anticipation (Durstewitz, 2003; Emmons *et al.*, 2017; Finkelstein *et al.*, 2021; Paton & Buonomano, 2018).

Apart from the already commented types of response, other profiles can be observed, such as transient activations during a fraction of the delay (as in the example in Figure 4.1d) or cells whose firing rate decreases during stimulus presentation or delay (not shown). These single neuron examples constitute a "fine detailed" picture of the PFC dynamics, showing that there is a variety of different stimulus-selective responses.

4.1.2 The working memory code undergoes a transition between the cue and delay epochs

Analyzing the recordings at the population level gives a more synthesized way of interpreting the nature of the neural code. In the last years, it has become increasingly common to complement the analysis based on the single-cell responses with a measure based on the activity of all cells. A cross-temporal decoding pattern (as the one in Figure 4.1e) can be useful to illustrate some properties of the neural code.

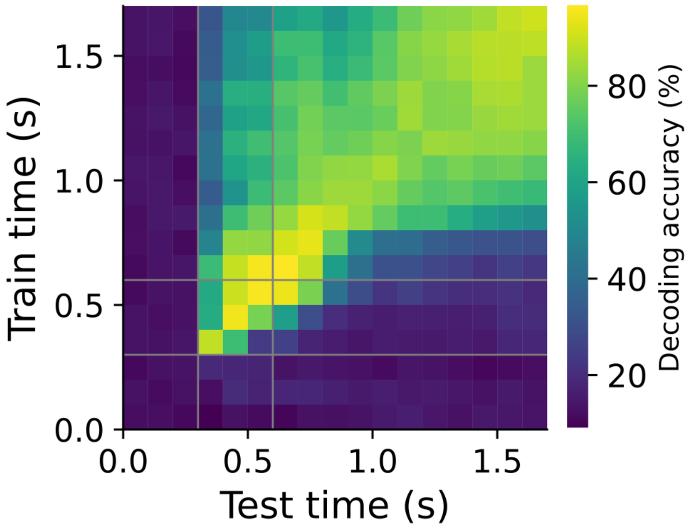

Figure 4.2: Cross-temporal decoding reveals that the WM code is dynamic during the transition from cue to delay, and stable during late delay. Pattern of decoder accuracy. Each data point (small squares) indicates the accuracy with which a decoder trained on the data at the time specified by y−axis is able to decode what the presented stimulus location was at the time indicated by the x−axis. Gray lines separate stimulus from delay epoch.

The off-diagonal regions of above chance accuracy in Figure 4.1e represent epochs during which the code is stable, allowing the decoder to generalize from a given time point to another. On the other hand, significant drops in decoding accuracy when moving away from the diagonal indicate a substantial change in the neuronal code. Qualitatively, the pattern observed in our data set (Figure 4.1e) is equivalent to the one obtained in several other studies of the

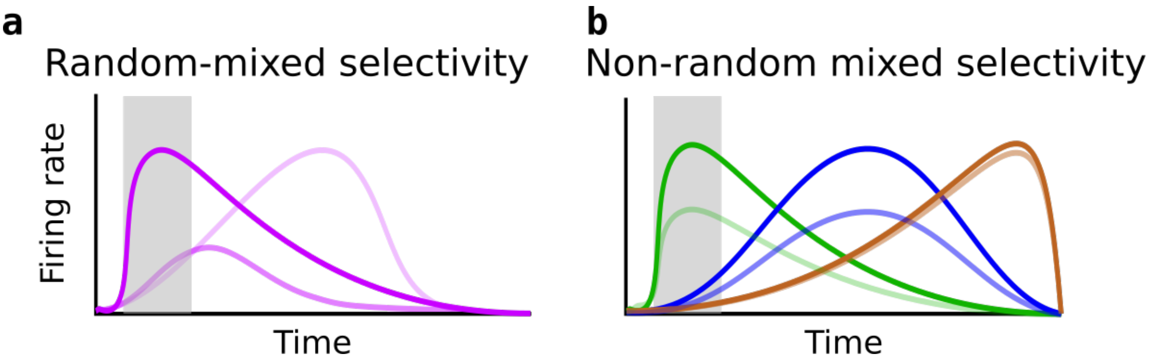

Figure 4.3: Different patterns of neuronal activation can explain the experimental observations Two extreme scenarios are illustrated that could explain the observed dynamics in the cue-to-delay region (Figure 4.2). **a** Random mixed feature selectivity; neurons have non-structured combinations of feature selectivity. Consequently, some neurons (one example neuron shown) will be active for different combinations of task epoch and stimulus identity. Responses to different cues are highlighted with different color intensities. **b** Non-random (structured) selectivity; neurons respond to specific combinations of task epoch and stimulus identity. Three neurons (green, blue, and brown) are shown. They all have a fixed preferred cue and are active respectively during distinct task epochs. Response of the three neurons to a cue presented away from their respective preferred locations.

same WM paradigm (Constantinidis & Wang, 2004; Mendoza-Halliday & Martinez-Trujillo, 2017; Spaak *et al.*, 2017; Stokes *et al.*, 2013): the WM code undergoes a significant change in the transition from cue to delay epoch, and during the delay, the representation becomes stable. While many models have been proposed to explain how stability can come about in the WM code (Camperi & Wang, 1998; Compte *et al.*, 2000; Renart *et al.*, 2003; Wimmer *et al.*, 2014), the origins of this dynamic cue-to-delay transition still need to be better understood. Investigating its underlying mechanisms constitutes one of our primary motivations for this work. In the sections below, we will discuss analyses carried out on the neural recordings that will establish a connection between the single-neuron picture (Figure 4.1) and the decoding pattern (Figure 4.2) and set a basis for the formulation of a computational model which will be compatible with both.

4.1.3 Possible scenarios that could explain the dynamics in the data

Currently, there is no agreement about the mechanisms that underlie the observed dynamics in the data (Figure 4.2). Before we explain our analysis, we will consider which scenarios would be compatible with the observed dynamic transition (Figure 4.1e). A possibility would imply a random distribution of the stimulus-selectivity during cue and delay epochs among the neurons (Figure 4.3a). In this scenario, some neurons would respond maximally to different stimulus locations during cue and delay epoch (dynamic selectivity). In a contrasting scenario (depicted in the cartoon in Figure 4.3b), neurons with stable stimulus-selectivity could be active at different task epochs. Altogether, this second scenario would correspond to a higher degree of structure in the neural responses than the first one. We want to investigate which scenario is more compatible with the experimental observations. For this purpose, we carry on the analyses explained below.

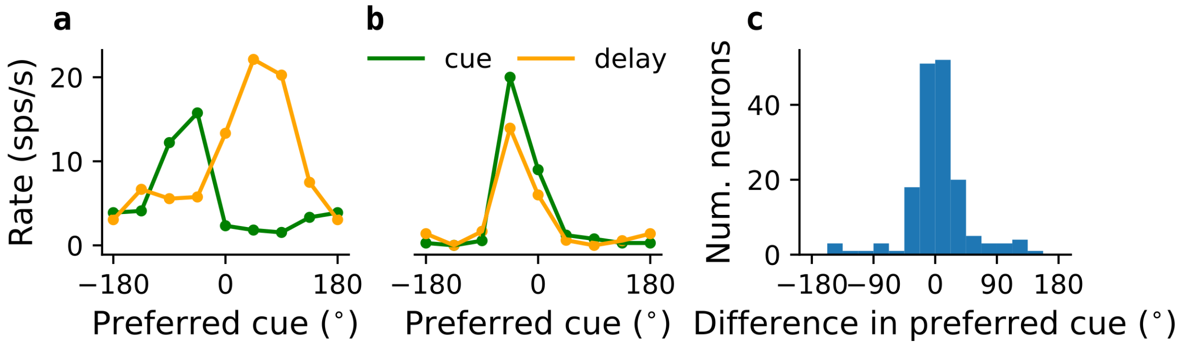

Figure 4.4: Most neurons do not exhibit changes in preferred selectivity during WM task. Preferred cues of neurons with significant selectivity ($N = 166$) during cue *and* delay periods were compared. **a** Example neuron whose preferred stimulus location during the cue period differs from the one during the delay. **b** Neuron whose selectivity is stable. **c** Histogram of the magnitude of the selectivity changes for all neurons with significant selectivity in both cue and delay periods.

4.1.4 Experimental evidence for stable stimulus selectivity and functional clustering

Analysis of the tuning properties of PFC neurons reveals stable stimulus selectivity across task epochs.The mixed selectivity hypothesis from above (Figure 4.3a) implies that a significant part of the neurons should change their stimulus-selectivity from the cue to the delay period. To directly test this, we quantify changes in stimulus selectivity between cue and delay epochs. We do not see many neurons exhibiting changes in the preferred location (Figure 4.4). Moreover, a two-way ANOVA for circular data revealed no statistically significant change in preferred cue for N = 166 neurons that were active during cue and delay (P = 0.83). This result suggests that the data are more compatible with the scenario depicted in Figure 4.3b.

In our further analysis, we will capitalize on the unique experimental design of Markowitz *et al.* (2015). To distinguish visual from memory effects, they included a variation of the oculomotor delayed-response (ODR) task where the stimulus was maintained on the screen until the animals were cued to respond. This task design allowed them to classify each neuron into one of three possible categories according to its firing rate during the last 300 ms before the animal was cued to respond. More specifically, the classification criterion was whether a neuron's firing rate during the classic mnemonic task (memory) was significantly different from the firing rate during the visual task (see Section 3.2.1.2, and Markowitz *et al.* (2015)). The units whose rate was significantly different were split into two classes depending on the task (memory, visual) to which they were most selective. The average firing rates of the three resulting classes are shown in fig. 4.5c-e. It is easy to see that the time course of the population averages in Figure 4.5 resembles some of the ones of the single neurons in Figure 4.1. The population average in Figure 4.5c corresponds to neurons that always fire in the presence of the stimulus, experiencing a significant decrease in activity upon stimulus removal. As we already mentioned above, this stimulus-dependent activity profile resembles the one of a sensory neuron. For this reason, we will refer to this type of neuron as *perceptual*. The population average in Figure 4.5d shows a ramping response in the mnemonic condition, which is almost

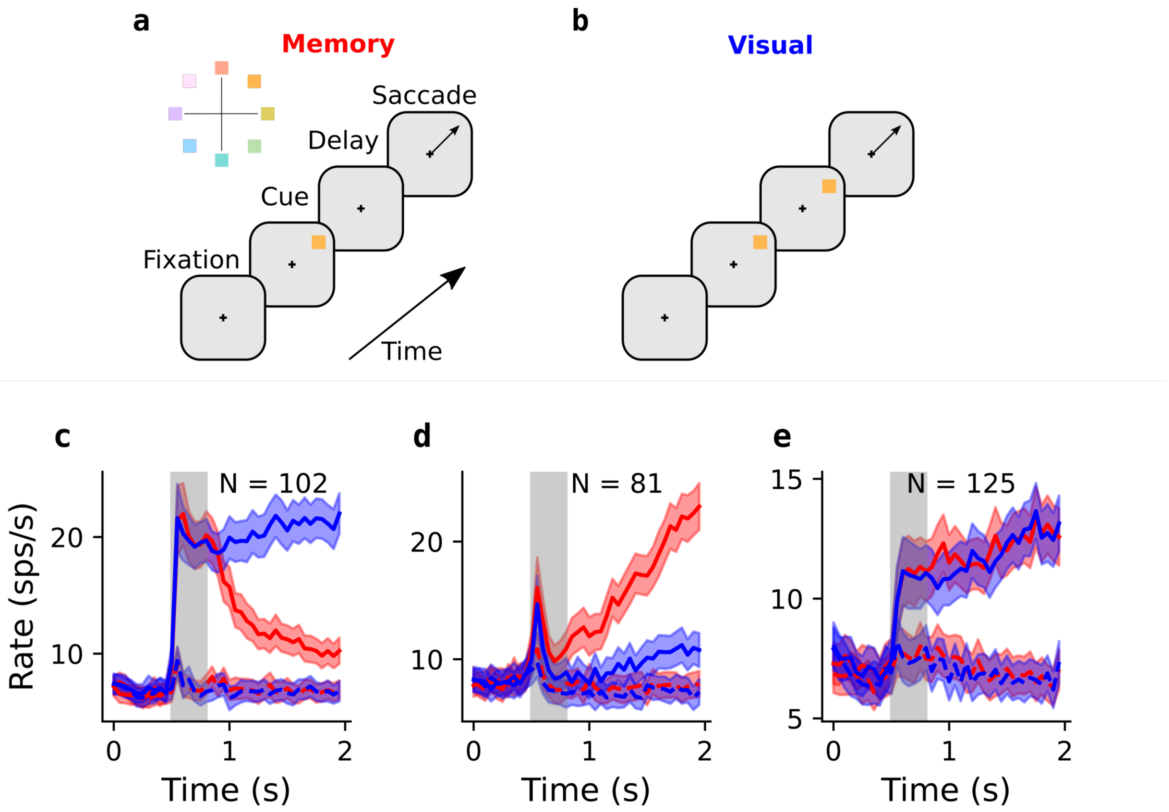

Figure 4.5: A visual variation of a memory-guided saccade task allows to classify neurons into three categories. a: schematic of the classic memory task (left) and the visual task (right) during which the visual stimulus is not removed from the screen until the animal is cued to respond. **c-e** Population average firing rates for the three groups obtained in a classification according to task condition selectivity during the delay. **c** population responding in the presence of stimulus and beginning of memory delay, **d** population responding during memory delay, gradually increasing firing rate throughout the trial. **e** Population with sustained selective activity during both task conditions.

absent or strongly attenuated for the visual condition. These neurons carry stimulus-related information only when the task condition requires WM. Thus, we will refer to these neurons as *mnemonic*. Finally, the average firing rate of the neurons in Figure 4.5e, resembles the one of the "persistent type" neuron in Figure 4.1a. It reflects the maintenance of stimulus-selective information during both visual and mnemonic conditions. We will label them as *persistent* neurons. Classifying the stimulus selective neurons into the three groups in Figure 4.5c-e favors interpreting these groups as distinct functional sets or subpopulations (see the description in Markowitz *et al.* (2015)). However, this reasoning may seem circular because the selection is based on significant differences in firing rates at the end of the trial. Moreover, the PSTHs in Figure 4.5 are population averages, and as such, they do not account for the heterogeneity of the single neuron activities. To understand to which extent the neurons in Figure 4.5 correspond to different functional classes, we next analyzed the neuronal activity in more detail.

To get an insight into how stereotyped the activities of the respective populations are, we first analyzed the time course of individual neurons during the memory task. We focused our attention on the *perceptual* and *mnemonic* groups. The average firing rate of these neurons changes in time during the memory condition, and we expected this change to be directly related to the dynamic cue-to-delay transition. We fit the firing rates during the delay of all *perceptual* and *mnemonic* neurons with an exponential curve with an offset (see Section 3.2.1.3). In both groups, the activity profiles of neurons with a R^2 value above a chosen threshold (see Section 3.2.1.3 for detail) are consistent with the activity profile of their population averages (see Figures 4.6 and 4.7). The τ of the *perceptual* neurons in Figure 4.6 are associated with fast decay in activity after stimulus removal. Likewise, the τ of the *mnemonic* cells shown in Figure 4.7 (note we are showing the absolute value of τ) are associated with gradual increases in activity during the delay. This result confirms that the average PSTHs of the classes found in Markowitz *et al.* (2015), shown in Figure 4.5c,d, represent stereotyped activation profiles. The distinct activity profiles among neurons could be related to their computational function. We will address this point in the following chapters, along with the introduction of our computational model.

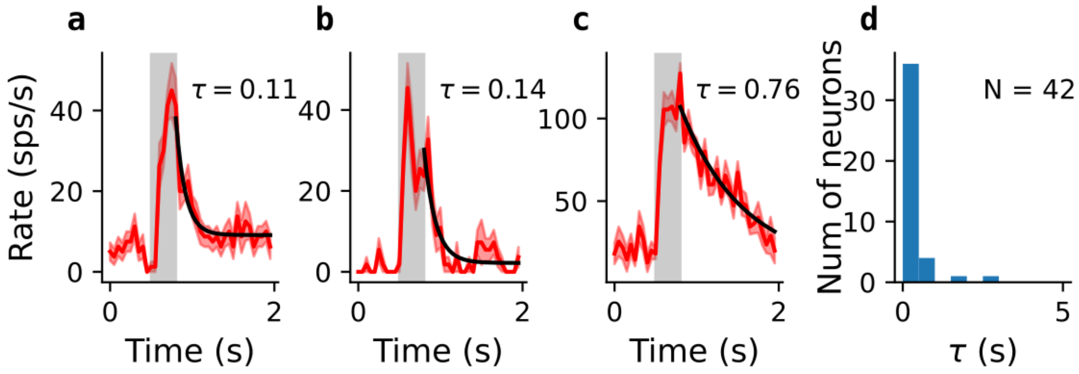

Figure 4.6: Single neuron time course characterization of *perceptual* PFC neurons. a-c Exponential curve fitting a decrease in firing rate of example neurons from the *perceptual* class after stimulus removal. **d** Histogram of time constants obtained when fitting the rest of the *perceptual* neurons in Figure 4.5c, neurons with $R^2 > 0.4$ included.

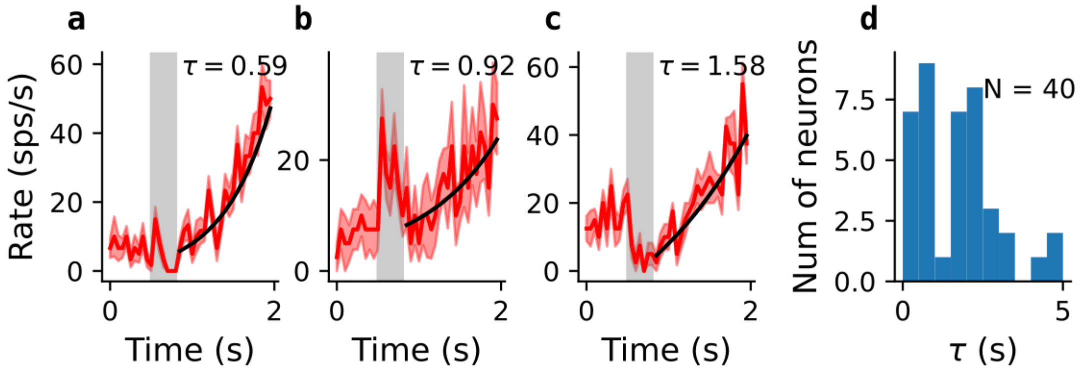

Figure 4.7: Single neuron time course characterization of *mnemonic* neurons. a-c Exponential curve fitting an increase in firing rate of example neurons from the *mnemonic* class after stimulus removal. **d** Histogram of time constants obtained when fitting the rest of the *mnemonic* neurons (Figure 4.5d), neurons with $R^2 > 0.4$ included.

4.1.4.1 Ramping PSTHs and single neuron activity

As argued by Latimer *et al.* (2015), ramping PSTHs as in Figure 4.7a-c are not always indicative of ramping activity at the individual neuron level (see cartoons in Figure 4.8i,j for illustration). Indeed, a ramp can result from averaging the activity of neurons with a step-like activation at different time points (Figure 4.8i-j). To distinguish between these scenarios, we examined the inter-trial variability of the *mnemonic* neurons and compared it to the variability of surrogate data (Figure 4.8). Ramping activity can be observed during single trials for some of the *mnemonic* neurons (Figure 4.8a-f). To analyze it more systematically, we measured the inter-trial variability as the Fano factor. We averaged across all neurons in the *mnemonic* group (Figure 4.8g,h) and compared it to the same variability measure on a network of surrogate neurons with step-like activations (Figure 4.8i-k). The differences between the shape of the average variability obtained from the data neurons (Figure 4.8h) and the one obtained from the surrogate step-activating neurons (Figure 4.8k) indicate that the ramping observed in the recorded cells does not arise from step-like activations at random times (Latimer *et al.*, 2015). In sum, we conclude the ramping of the *mnemonic* cells (Figure 4.5d, Figure 4.8g) reflects an essential aspect of the data rather than being an artifact.

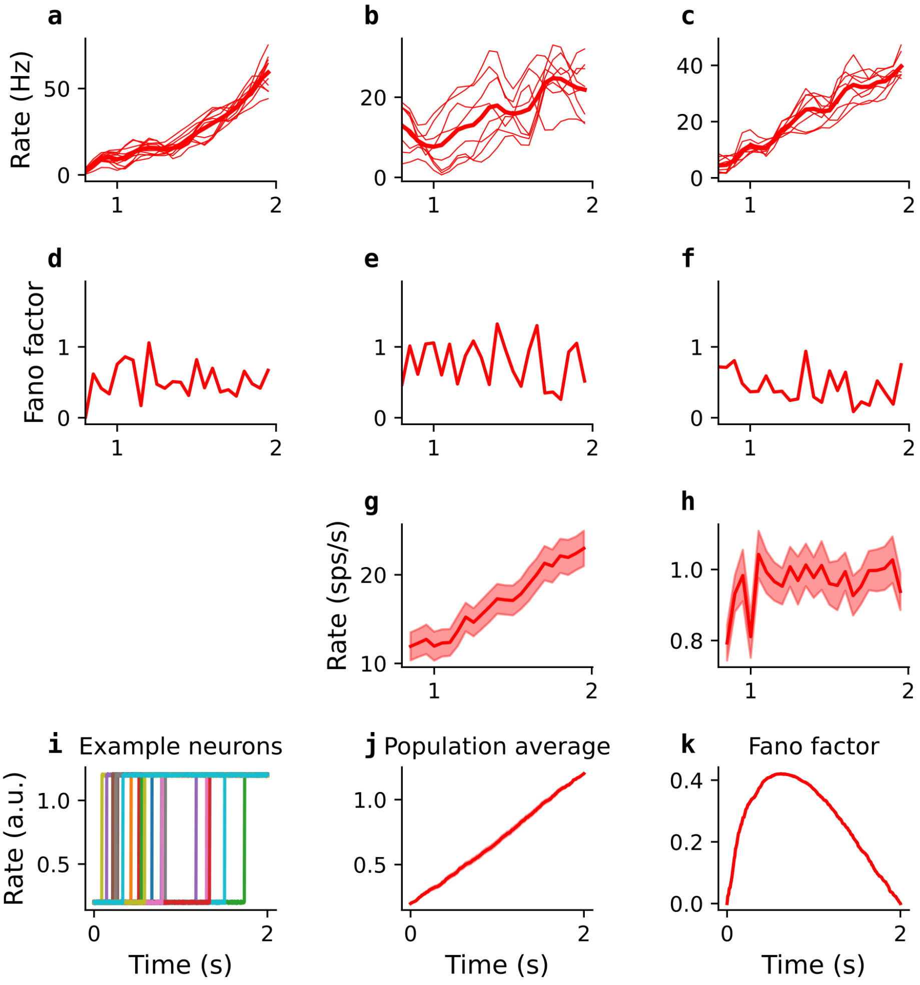

Figure 4.8: The inter-trial variability of the *mnemonic* neurons is not compatible with single-trial step-like activity a-c firing rates during single trials (red thin lines) and average firing rate over trials (red lines) for the example ramping neurons shown in Figure 4.7a-c (red lines). spike count window $T = 50$ ms, smoothed with Gaussian filter ($\sigma = 1$). **d-f** Fano factor for each of the neurons (**a-c**) **g** average PSTHs of the *mnemonic* neurons during the delay of the memory task. **h** average Fano factor of the *mnemonic* neurons during the delay. **i-k** Example illustration with surrogate data. **i** example neurons with step-like activity, steps times sampled from a random uniform distribution. **j** average activity of 1000 surrogate neurons with step activity. **k** average Fano factor for the 1000 neurons.

Conclusions from the single neuron analysis: there is stable selectivity and stereotypical responses

In summary, the analysis of the single neuron firing rates shows that in spite of a rich variety of activity profiles, certain stereotypical activity profiles are over-represented. Analyzing the neural activity using cross-temporal decoding reveals that the neural code undergoes a dynamic transition from cue to delay epoch and then becomes stable during the delay. We investigated how the dynamics at the population level are connected to the single neuron heterogeneity, and we found that the dynamics could be associated with groups of neurons that are active at different task epochs. Indirect support to this hypothesis is given by the absence of changes in the stimulus-selectivity among the neurons (Figure 4.4). More specifically, the average PSTHs of the *perceptual* and *mnemonic* groups (Figure 4.5, Markowitz *et al.* (2015)) seem to correspond to neurons most active during cue and delay periods, respectively. However, we were aware that the heterogeneity of the single neuron response can be masked by averaging over many neurons. Thus, we further analyzed the single neuron time courses to verify how many of them were faithfully captured by the population averages (Figures 4.6 and 4.7). We find that a portion of the cells ($\sim 30\%$) have time courses that are different from the average time course of the whole group. Still, the firing rates of more than 50% of the cells are consistent with the average PSTHs found in Markowitz *et al.* (2015) (Figure 4.5), which means that there is a significant number of *perceptual* and *mnemonic*-type of cells among the recorded neurons. Taken together, these results seem to suggest that the presence of cells with contrasting activation profiles may be linked to the observed dynamics in the population code. However, given the circularity of the classification criterion (see above), we wondered whether we would obtain a similar result using a less biased analysis.

In the following section, we will discuss an analysis of the same recordings in an unbiased way, using dimensionality reduction techniques. Reducing the dimensionality also offers a way of visualizing the relevant neural dynamics by representing the activity along a reduced number of dimensions (we will use demixed-Principal Components, see Section 3.2.2.2 and following section)). Finally, this approach allows us to quantify the degree of similarity between neuronal responses, revealing how much neurons form *clusters* (as the interpretation in Markowitz *et al.* (2015) suggests).

4.1.5 Analysis based on dimensionality reduction techniques supports the presence of structure in the prefrontal networks

The analyses presented in this section reveal the presence of non-random mixed feature selectivity in the neural data. We measure the amount of clustering in the data by quantifying the structure in the Principal Components (PCs) space. The PCs are directions in neuronal activity space that capture most of the variance in the data. They can be strongly correlated with specific task features or combinations of features (Jolliffe *et al.*, 2016; Kobak *et al.*, 2016). Using a set of PCs naturally enables us to get rid of noise from the data, making it more interpretable. Moreover, when the PCs are directly associated with specific task features, the observed structure can be related to how the neurons encode these features. For the results shown in this section, we used a variation of Principal Component Analysis called demixed Principal Component Analysis (dPCA, see Section 3.2.2.2) that constrains the obtained PCs to capture variance due to specific features or combination of features (Kobak *et al.*, 2016). We measured the presence of structure in the dPC space, using a clustering algorithm (ePAIRS) and finally related the results to the single-neuron-based findings in the last section.

4.1.5.1 Functional clusters in the feature space

As a first step to investigate the presence of structure in our data, we used a non-parametric statistical test called elliptical projection angle index of response similarity (ePAIRS) (Dubreuil *et al.*, 2022; Hirokawa *et al.*, 2019; Raposo *et al.*, 2014). This test is based on computing the nearest neighbors' distances between neurons in feature space and comparing the distribution of distances to a null distribution generated by bootstrapping from the original data (see Chapter 3). The cartoon example in Figure 4.9 illustrates how the ePAIRS can distinguish between two extreme scenarios with contrasting population structures. When feature selectivity is distributed randomly (Figure 4.9a), neurons do not form clusters in the feature space. As a consequence, the distribution of angular distances between neurons for the original data is equivalent to the one of a generated null distribution (Figure 4.9c, see Chapter 3). On the other hand, the non-random distribution of feature selectivity is reflected as neuronal clusters in the feature space (Figure 4.9b). In this case, the original distribution of distances between neurons in feature space has a significantly lower mean than that of a generated null distribution (Figure 4.9d, see Section 3.2.5). From a computational point of view, the behavior of a group of neurons with highly structured (highly non-random mixed) selectivity (as the example Figure 4.9b,d) can be faithfully represented by as few as two units. However, when the feature selectivity is randomly distributed (Figure 4.9a,c), there is no such equivalent lower-dimensional representation.

We obtain a significant measure of clustering when applying the ePAIRS to our data (Figure 4.10a). As a first agnostic measure, we applied the algorithm (ePAIRS) on a 20-PC space, which capture a 62% of the total variance (Figure 4.10). This measure is agnostic or unbiased because we do not select each PC according to any specific criterion. We only ensure we take enough PCs to capture a significant amount of the total variance.

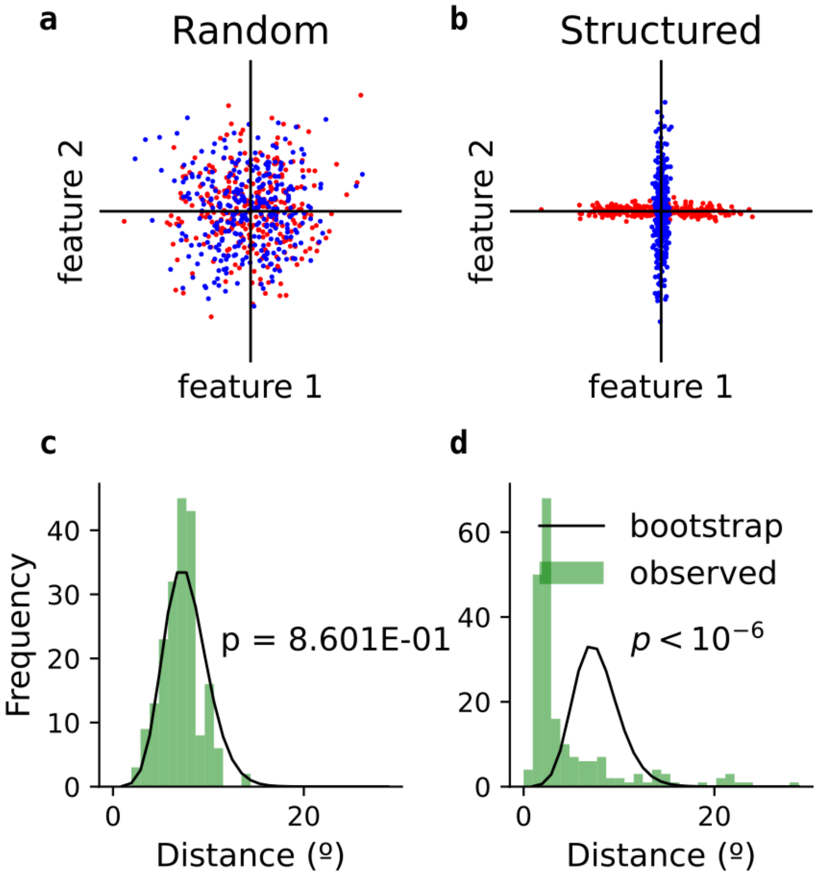

Figure 4.9: Population structure example: extreme scenarios Cartoon illustrating two extremes types of population structure. **a** Complete random structure. Each neuron's weights (x, y dimensions in the plot) are drawn randomly from a normal distribution, $N(0,\sigma)$. **b** The weight on one of the features is drawn from the same normal distribution used in **a**, $N(0,\sigma)$, while the weight on the other feature is drawn from a distribution with lower variance ($N(0,\sigma/10)$). Effectively, this mimics selectivity to one of the features and not to the other. **c,d** Applying the ePAIRS method to the weights in **a,b**. The green histogram shows the distribution of nearest neighbor distances in the surrogate data, and the black lines are the corresponding histograms of nearest neighbors obtained by bootstrapping from a distribution with the same covariance as the original data (but disrupting population structure, see Section 3.2.5).

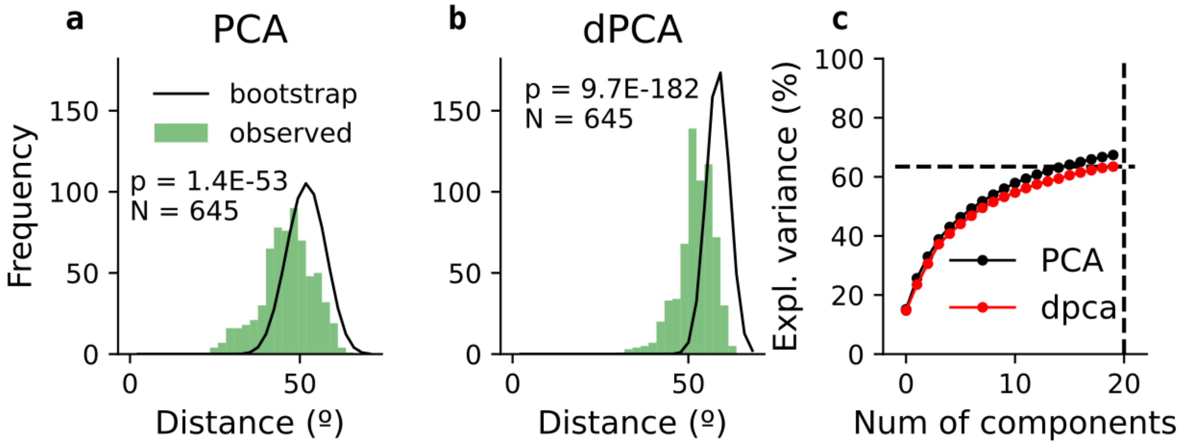

Figure 4.10: ePAIRS method reveals non-random population structure in prefrontal recordings. **a,b** Distribution of angles between nearest neighbors (average distance for 13 nearest neighbors) of the original data (green) and null distribution (black). **a** ePAIRS applied to the first 20 PCs, **b** ePAIRS applied to the first 20 dPCs. **c** Variance explained by including different numbers of PCs and dPCA.

4.1.5.2 Neural trajectories along demixed Principal Components reflect task-relevant information

While the ePAIRS test indicates the presence of non-random structure in the PFC population activity, it does not inform us about the geometry of this structure. To reveal the nature of the population structure, we first inspected some of the most-variance-capturing dPCs separately. As we expected from previous analysis (Chung & Abbott (2021); Kobak *et al.* (2016)), the two first dPCs which captured stimulus-related variance reflected the geometry of the task. When plotting the neural activity of the excitatory neurons along the stimulus dPCs respectively (Figure 4.11a,b), the trajectories during both task conditions were split according to the location of the presented stimulus. The eight separate locations can be recovered when combining the information in both dPCS by plotting one against the other (Figure 4.11c).

This result is the very reflection of the presence of stimulus selectivity at the level of population dynamics. This stimulus selectivity is consistent with the selectivity described at the level of single neurons above. How much the neural trajectories should diverge according to the task condition is a priori less clear. From the analysis of the neuronal firing rates (Figure 4.5), however, we expect significant differences during the two task conditions. Indeed, when plotted along the first task-variance-explaining dPC, the neural trajectories corresponding to different task conditions diverge shortly after the time at which the stimulus is withdrawn in the memory condition (when the two tasks become distinguishable). The information in the dPCs commented above can be combined into a 3-dimensional representation (Figure 4.11e) that synthesizes the previous results and reveals the geometry of neural activity in both the memory and the visual task. This three-dimensional representation (Figure 4.11e) illustrates the relation between the neural codes corresponding to the visual and memory conditions. Both task conditions are indistinguishable if looked at in the space spanned by the stimulus dPCs (Figure 4.11c). However, the task-condition related dPC (Figure 4.11d) separates both representations, showing that they are symmetric but not overlapping (Figure 4.11e). Computationally, this representation allows a stimulus decoding to be independent of the task

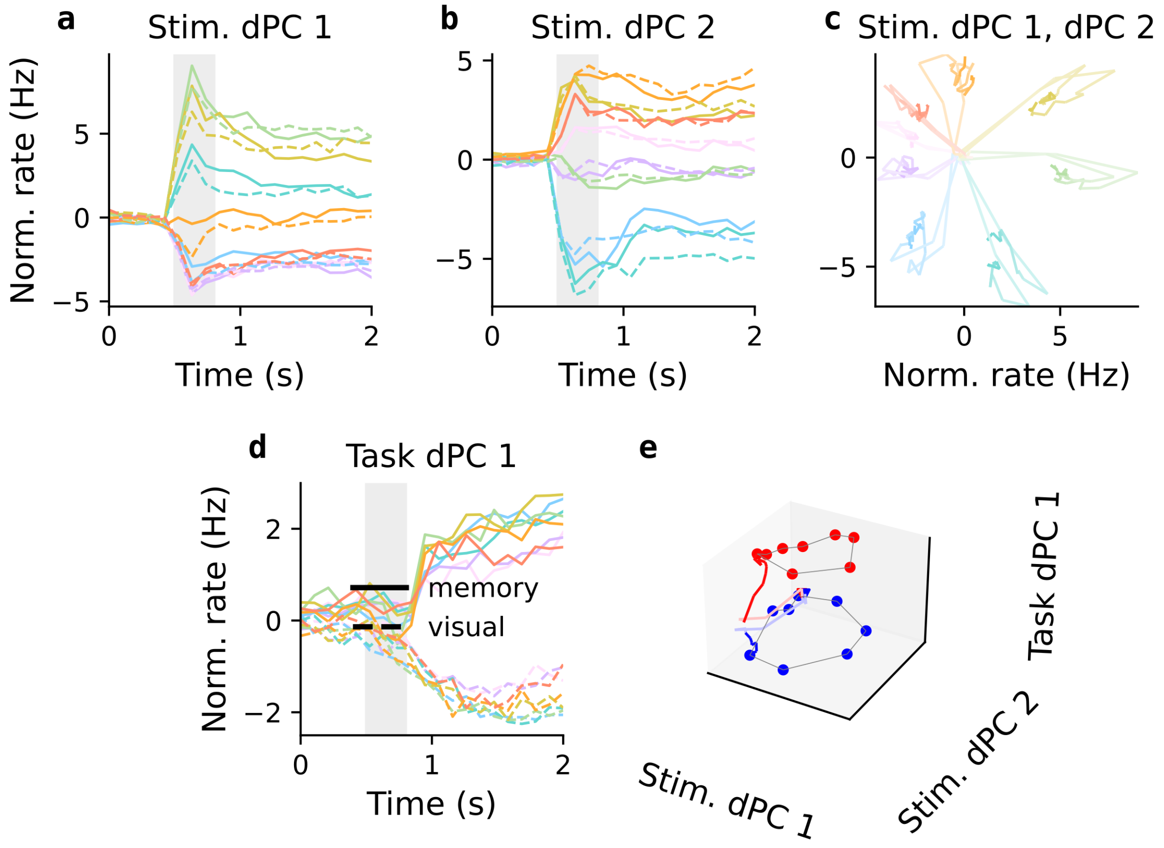

Figure 4.11: Demixed-Principal component analysis of the prefrontal neuronal recordings. (a-d) Time course of example dPCs. **(e)** Firing rates projected on the first two stimulus-dPCs. **(f)** Neuronal firing rate of the excitatory neurons projected onto the first and second stimulus-dPCs (x and y-axis) and first task-dPC (z-axis). The dots indicate the network activity at the end of the delay (red: memory, blue:visual). Two example trajectories for a given cue (90°) is shown: they start a common spontaneous state (filled square) and evolve in the same direction during the 300 ms cue period (faint red and blue lines) before they diverge during the delay period (solid lines).

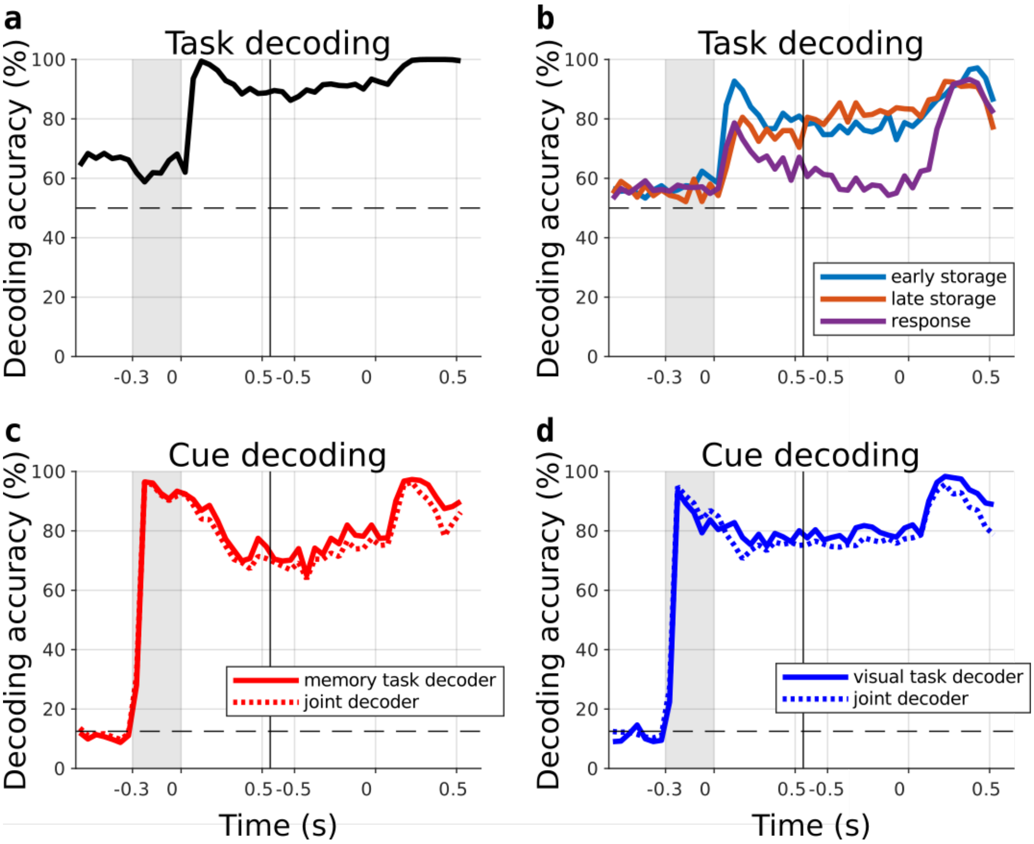

Figure 4.12: Decoding task condition and stimulus location provides information about the geometry of the neural representation a task decoding performance when using all the excitatory neurons. **b** task decoding performance for different functional subpopulations. **c** decoding of the presented cue when training on the visual condition and testing on the memory condition. **d** decoding of the presented cue when training on the memory condition and testing on the visual condition.

condition, even though the overall neural code is task-dependent.

An alternative way to understand the geometry of the neural representation is to decode stimulus location and task condition separately from the neural data (Figure 4.12). The task condition decoding becomes accurate after the cue is removed in the memory task (Figure 4.12a,b), where both conditions become distinguishable. Applying this procedure to the subpopulations separately highlights that the *perceptual* and *mnemonic* are more informative about the task condition (task selective). The above chance accuracy when decoding across task conditions (Figure 4.12c,d) is supported mainly by the activity of the *persistent*.

4.1.5.3 Distinct contributions to the encoding of time, stimulus, and task condition

We obtained a more detailed view of the non-random structure present in the data by considering each neuron's contribution to a pair of features. This analysis was inspired by the one in Yang *et al.* (2022) (see Fig 3d in Yang *et al.* (2022), Chapter 3). We can represent each neuron's contribution to a pair of features (time-stimulus, time-task or task-stimulus) as a two-dimensional vector (Figure 4.13a). The analysis is based on computing the distribution of the angles these vectors form with the x-axis in the respective 2-dimensional feature spaces. The distributions of the original data (Figure 4.13b-d) are compared to null distributions (*orange* in Figure 4.13) obtained by bootstrapping. It is worth noting a difference between the feature-specific dimensions used for our analysis and those used in Yang *et al.* (2022). We combine the weights of different feature-specific dPCs to obtain the time and stimulus-specific dimensions. The reason to combine the first two stimulus-dPCs is that the information they provide is complementary due to the geometry of the task (see Figure 4.11a-c). In the case of the time-dPCs, we combined the first two because each explained a significant fraction of variance (14.9% and 7.3%) and because of their respective time courses. In both cases, the components of the combined dimensions are obtained as Euclidean norms:

$$w_{\text{combined feature},i} = \sqrt{w^2_{\text{feature},1,i} + w^2_{\text{feature},2,i}} \qquad i = 1\ldots N \tag{4.1}$$

Due to these combined-dPC dimensions, the null distributions are not flat, unlike the ones in Yang *et al.* (2022) (see Figure 4.13b-d and see Chapter 3 for a detailed explanation).

The original angle distribution is significantly different from the null distribution for all the two-feature comparisons (Figure 4.13b-d, Kolgomorov-Smirnov test $p < 0.001$, see Section 3.2.6). Each 2-feature comparison offers a different view of the non-random structure of the mixed feature selectivity present in the data.

Fewer neurons are coding a mixture of stimulus and time than expected by chance (Figure 4.13b). This difference can be related to the existence of stimulus-selective neurons with flat activity profiles (no information about time), such as the neurons in the *persistent* group (Figure 4.5e). The under-representation of neurons coding time but not task condition (Figure 4.13c, left) indicates a correlation between the coding of these features: Neurons that encode time are usually also selective to the task condition. This correlation can be understood in light of our previous analysis. The activities of the *perceptual* and *mnemonic* (see Figure 4.5b,c) contain information about the task condition, and the ramping of the*mnemonic* neurons can be easily mapped to the elapsed time. Finally, the neurons' contribution to task-condition and stimulus (Figure 4.13d) reveals two facts. One is that there are more neurons

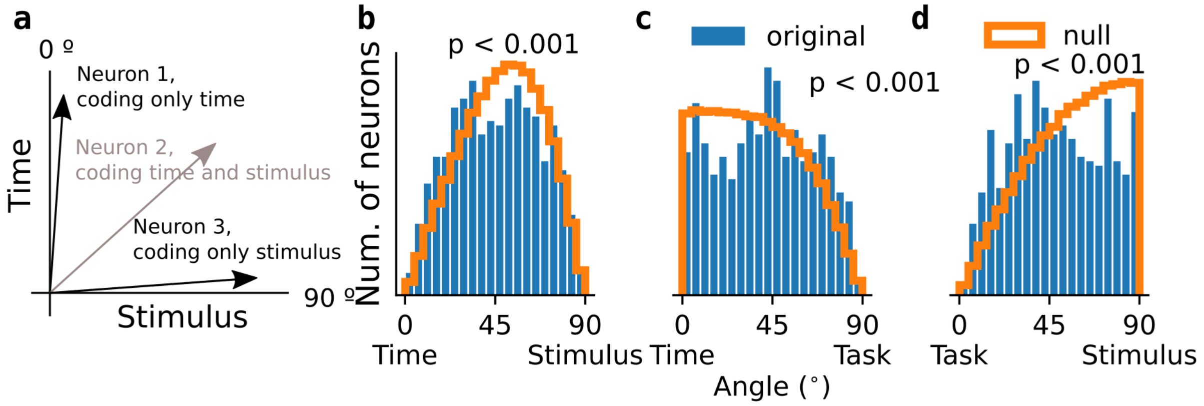

Figure 4.13: Degree of clustering, two-feature comparisons. **a** Method explanation: each neurons selective for a pair a feature determines an angle in the 2-d feature space. Since the weights are always positive (either obtained as "normalized combination" of feature weights or a the absolute value) the angles lie in between 0 and 90 °. **b-d** Original distribution of angles (blue) and null distribution (orange) for the three different feature pair combinations. time-stimulus (**b**), time-task (**c**), task-stimulus (**d**).

coding for task condition than expected by chance. The other is that stimulus-selective neurons are usually selective to the task condition. Again, both facts align with the over-representation of the stereotypical firing profiles of the *perceptual* and *mnemonic* neurons, which is different for the two task conditions and stimulus selective (Figure 4.5c-e).

By classifying the neurons according to their contribution to the first task-dPC, we get three groups (Figure 4.14) that have a high overlap with the ones obtained in Markowitz *et al.* (2015) (Figure 4.5b-d). The neurons' weights on the first task-dPC are split according to a chosen threshold value (see Section 3.2.7). We compared the original distribution of weights with one obtained from a data set where the task condition labels had been randomly shuffled (Figure 4.14d,e). The shuffled distribution disrupts the information related to task condition. We observed heavier tails in the distribution of the original data (Figure 4.14e,f). This result again points to the non-random distribution of selectivity and specifically highlights the importance of *perceptual* and *mnemonic* types of neurons.

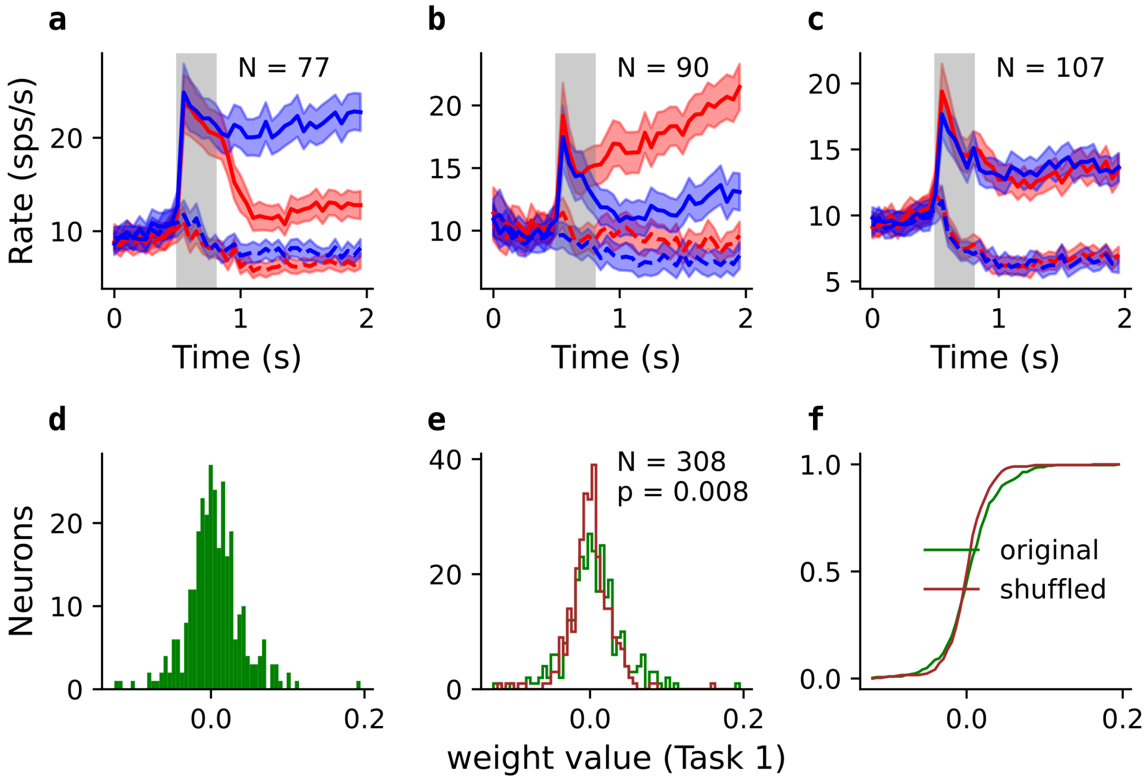

Figure 4.14: Classifying neurons according to dPC weight value . (a-c) Average PSTHs of neurons grouped according to weight value (components of dPC in neuronal space) on task-dPC. **(d)** Histogram of weight of the first task-dPC.**e** Comparison between the histogram of the weights on the original 1st task-dPC (green) and a histogram of the weights on the 1st task-dPC obtained when random shuffling the task weights. **f** cumulative distribution function (CDF) for the histogram in **e**.

4.1.6 Cross-temporal decoding analysis underscores the relevance of different activity profiles

We have established the presence of non-random structure in the PFC population activity and its relation to functionally specialized groups of neurons. We now return to the question of the origin of the cue-delay transition observed in population decoding (Figures 4.2 and 4.15a,b).

We first compare the cross-temporal decoding patterns corresponding to the memory and visual conditions (Figure 4.15). The dynamic transition is mainly present in the memory condition (Figure 4.15b). The decoding accuracy drops when training during the first 100 ms of cue presentation and testing during the rest of the visual task (Figure 4.15c). This is likely due to the transient activations observed in all the stimulus-selective populations (Figure 4.5c-e) during cue presentation. Otherwise, the code is stable, and decoding is possible across cue and delay. This difference in decoding between the two task conditions can already be associated with the *perceptual* and *mnemonic* neurons, whose activity profiles in the two task conditions are different.

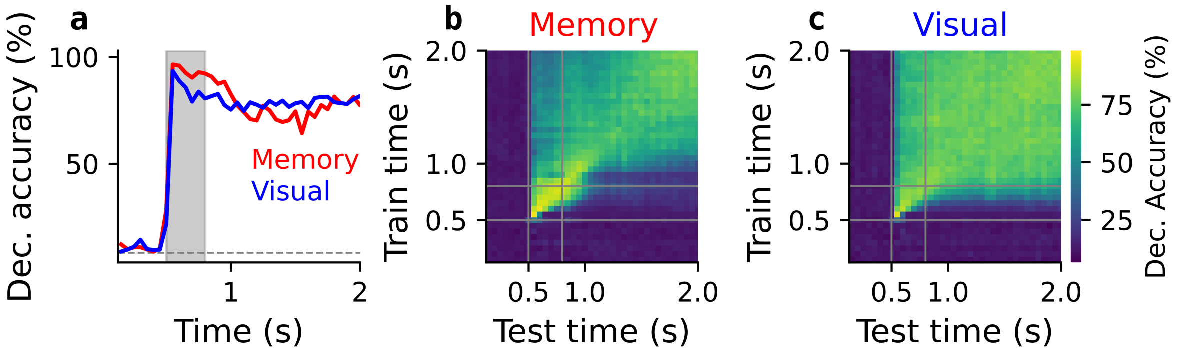

Figure 4.15: Population decoding in the memory and visual task. **a** train and test at same time point (in time decoding) decoding performance vs time for 2 task conditions. Cross-temporal decoding pattern for the memory task. Cross-temporal decoding visual task. Pseudopopulations used for decoding contained all excitatory neurons with enough trials per cue condition ($N = 650$).

The importance of these neuronal types (*perceptual*, *mnemonic*) is further emphasized by the results obtained by applying cross-temporal decoding on neurons from different classes. When decoding from *persistent* neurons, the performance slightly decreases after the presentation of the stimulus (during both task conditions) remaining above chance until the motor response (Figure 4.16g-i). Analyzing this pattern in isolation could lead to the conclusion that *i)* these neurons alone explain the dynamics in the cue-to-delay transition and that *ii)* the same neuronal strategy is used in both task conditions. However, the respective decoding patterns of the other two populations challenge this view, underscoring the importance these other types of neurons might have for the behavior. We find that the existence of neurons of the *perceptual* type contributes significantly to the change in the cross-temporal decoding's performance at the cue-to-delay transition. When decoding only from neurons pooled from the *perceptual*

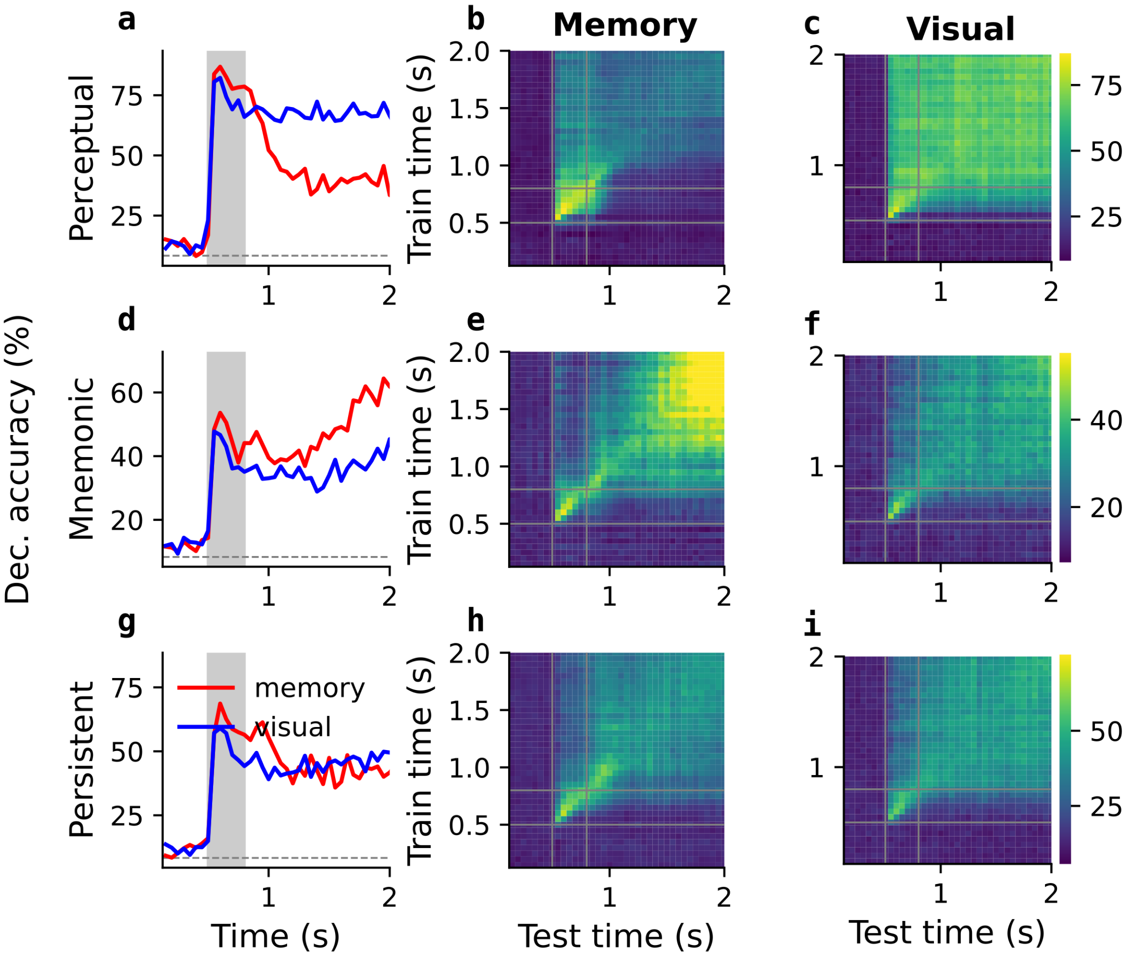

Figure 4.16: Decoding from subpopulations separately a-c Decoding analysis for the *perceptual* neurons. **a** train and test at the same time point (in time decoding) decoding performance vs time for the memory task (red) and visual task (blue) condition. **b** Cross-temporal decoding pattern for the memory task. **c** Cross-temporal decoding visual task. **d-f** Same as **a-c** for the *mnemonic* neurons. **g-i** Same as **a-c** for the *persistent* neurons.

type population, the performance drops about 50 % (from 80 % to 40 % correct) when the stimulus is withdrawn in the memory condition (Figure 4.16a-c). In the same way, the increase in decoder performance during the memory delay is explained by the *mnemonic* type of neurons, many of which experiment an increase in firing rate toward the end of the delay (Figure 4.16d-f).

Crucially, the highest performance measure during cue and late delay is obtained when decoding from *perceptual* and *mnemonic* neurons, respectively.

In summary, the separate decoding patterns reveal that the decoder relies strongly on the *perceptual* neurons during the cue epoch and on the *mnemonic* neurons during the delay.

This analysis suggests that the cue-to-delay transition (Figure 4.1e) corresponds to *perceptual* neurons being most stimulus-informative during the cue period, while *mnemonic* neurons are most informative during the memory delay.

Conclusions

Motivated by the three-group classification in Markowitz *et al.* (2015), we extended their analysis to understand how their average PSTHs relate to the activity of single prefrontal neurons. A closer inspection of the single neuron time courses (analyzing the distribution of their time constants and the intertrial variability) reveals a higher degree of response heterogeneity, which the average PSTHs mask. However, a significant fraction of the neurons have indeed stereotypical activation profiles that match the average PSTHs of Markowitz *et al.* (2015) (Figure 4.5).

We complemented the single neuron analysis with several measures applied to reduced spaces of Principal Components. This population-based analysis widens the picture and offers a compact and helpful way of visualizing the neural dynamics. Moreover, by measuring the presence of clusters in the space of demixed-PCs, we could quantitatively establish the presence of non-random mixed selectivity in the data. To relate the observed structure to the dynamic cue-to-delay transition (Figure 4.2) we trained a decoder on separate sets of neurons. We find that the contrasting stereotypical activity of the *perceptual* and *mnemonic* neurons is directly related to the pronounced changes in the decoding.

4.2 A three-population model explains salient features of the prefrontal recordings

Section summary

In this section, we present a computational model comprising three functional subnetworks. The subnetworks mimic the stereotypical activity profiles observed in the data (Section 4.1). Each is modeled by a ring-attractor network working on a different dynamic regime, which is determined by the amplitude of its recurrent connections. The model explains the dynamic transition from cue to delay observed in the experimental recordings at the population level and makes specific predictions about its mechanistic origins. In particular, it highlights the relevance of the contrasting types of activity profiles (*encoding* neurons active during cue presentation and *storage* neurons active during memory delay), which underlie the orthogonality of the code across task epochs. The concrete network implementation makes specific predictions about how information is transmitted from the stimulus through the network, suggesting that feedforward structure is essential to the prefrontal working memory circuits.

We propose a standard three-ring network that implements maintenance due to the bistability of the *storage* ring (*bistable network*) and an alternative network, with a monostable *storage* ring, where memory is sustained with the help of external modulation (*ramping network*). The two network variations make equivalent predictions regarding the functional subpopulation structure and the dynamics of the code, but they propose conceptually different interpretations. While the *bistable network* is compatible with the view of PFC as an autonomous WM hub, the *ramping network* illustrates how memory maintenance could rely on distributed inter-area interactions (Christophel *et al.*, 2017).

Finally, the cross-temporal decoding analysis illustrates the relation between the subnetwork dynamics and cue-to-delay transition in the code.

4.2.1 Selectivity for a continuous stimulus: the ring-attractor model

In this section, we introduce a network model that explains salient qualitative features of prefrontal cortical activity analyzed in Section 4.1. The possible connections between the structure in the data and the dynamics at the population level (Section 4.1) motivated us to model the three functionally distinct populations of the Markowitz *et al.* (2015) data (Figure 4.5) as three distinct model sub-networks: A sub-network with strong activation during stimulus presentation, another sub-network that activates during the memory delay, and a third sub-network active both during the stimulus and delay epochs.

The task design in Markowitz *et al.* (2015), which involved a stimulus presented at evenly spaced positions on a screen, motivated us to use a network model with continuous stimulus representation. It has been shown that continuous representations can emerge from exposure to a discrete set of stimuli (Darshan & Rivkind, 2022). Besides, models of continuous stimulus encoding have successfully captured many experimental observations during delayed response paradigms (Compte *et al.*, 2000; Wimmer *et al.*, 2014). For these reasons, we chose to represent each model sub-network by a ring-attractor network (Compte *et al.*, 2000; Wimmer *et al.*, 2014).

In the ring attractor network, stimulus-selective activity is present due to the center-surround type of connectivity (see Figure 3.5, and Section 3.3.3), which is characterized by dominant excitation between neighboring neurons and long-range inhibition. In our model, the persistent activation in the absence of external stimulation relies on the amplitude of the positive recurrent connections and on the slow synaptic time constant of the NMDA neurotransmitter (see Compte *et al.* (2000); Wang (1999); Wong & Wang (2006)). Hence, our network model exhibits persistent activity due to network dynamics and not intrinsic bistability at the cellular level.

To model the single neurons within the networks, we used a stochastic model (leaky-integrate and fire model, LIF) that simulates spiking activity (see Section 3.3.1) and a firing rate model (mean-field) of the Wilson-Cowan type (see Section 3.3.2 and Wilson & Cowan (1972); Wong & Wang (2006)). The two models' behavior is qualitatively equivalent. The LIF model is more biologically plausible, while the firing rate model is computationally more efficient and analytically tractable. All results shown in this section are obtained with the LIF model.

Our model was developed to match the differential activity of the *perceptual*, *mnemonic*, and *persistent* neurons (Figure 4.5). These three activation profiles can be directly connected to the changes in decodability through the trial (Figure 4.2) that inform of a change in the neural representation. Moreover, the different profiles can inform the groups' respective functions in the context of the studied WM paradigm: *perceptual* neurons related to encoding processes, *mnemonic* neurons to maintenance and motor preparation, while the *persistent* neurons offer a steady, reliable readout. (see the discussion in Chapter 5). We, therefore, decided to model explicitly the activity of these three different groups.

We can summarize the qualitative features we wanted to capture as the following:

- A circuit that responds in the presence of external stimulation (labeled as *perceptual* neurons here)
- A circuit that fires during the memory-demanding period (the delay in the memory task) but not in the visual condition, when memory is not required.

- A circuit that fires from stimulus presentation until motor response, regardless of whether the stimulus remains on the screen or not.

Here, we present two canonical realizations that share a common three-ring architecture and capture the same essential features but offer different mechanistic interpretations. Both circuit proposals involve three sub-populations, which we label as *encoding*, *storage*, and *readout* populations (Figure 4.17). These populations mimic the activity of the *perceptual*, *mnemonic*, and *persistent* neurons in the Markowitz *et al.* (2015) data set. The first circuit (*bistable network*) we introduce receives excitation and suppression (excitatory to inhibitory neuron projections) from the stimulus and maintains the stimulus information during the delay through the bistability of the *storage* network. The second circuit (*ramping network*) does not include any network with bistability, and memory maintenance is implemented with the help of external non-selective projections to the *storage* network. We will first illustrate how the first circuit architecture behaves and the results it captures and then introduce the second one.

4.2.2 A three-population ring model with bistability

We illustrate how this model (we refer to as *bistable network*) works for the same two paradigms (memory, visual tasks) used in Markowitz *et al.* (2015). The model is composed of three sub-networks, modeled as ring networks with attractor dynamics, and the stimulus is implemented as a pulse with a Gaussian profile in the space of the neurons.

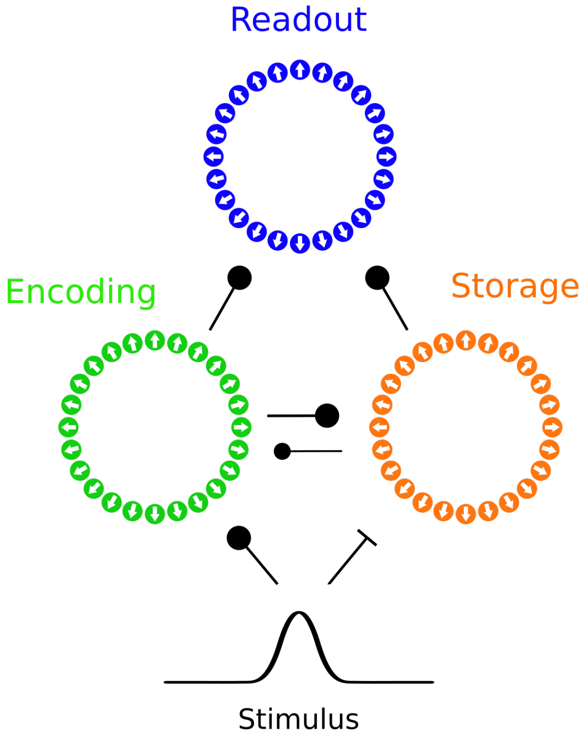

Figure 4.17: 3-ring model structure The stimulus, represented by the bell-shaped pulse, provides excitation to the *encoding* ring and suppresses (excitatory projection to inhibitory neurons) the *storage* ring. The excitatory neurons in the *encoding* ring project to the excitatory neurons in the *storage* ring, and some weaker projections are also present in the opposite direction (*storage* to *encoding*). The *encoding* and *storage* neurons provide excitation to the excitatory neurons in the *readout* ring.

The dynamics of each ring is described below:

- The excitatory neurons in the *encoding* ring receive excitation from the stimulus and provide excitatory connections to the excitatory neurons in the *storage* ring. The *encoding* recurrent connections are set to be strong enough to form an intrinsic bump of activity in response to external stimulation but too weak to sustain the bump in the absence of stimulation (monostable regime, see Chapter 3).
- The *storage* ring remains silent when the stimulus is present (cue period in the memory condition, whole trial in the visual). This stimulus-related suppression is guaranteed by excitatory projections from the stimulus to the inhibitory population in the *storage* ring. When the stimulus-mediated inhibition ceases, the *storage* gets activated by the excitation provided by the slowly decaying bump in the *encoding* ring (Figure 4.18a,g). In this circuit architecture (*bistable network*), the *storage* ring has stronger recurrence and operates in the bistable regime. The excitation provided by the *encoding* neurons is enough to allow the *storage* ring to form a bump at the stimulus location, and, thanks to its recurrent connections, sustain the bump during the mnemonic delay (Figure 4.18b,h).
- Finally, the *readout* network represents the activity of the neurons that respond selectively to the stimulus presentation and also sustain this selective activity after stimulus removal. This type of activity profile is the one exhibited by the *persistent* neurons in Figure 4.5e. In the circuit architectures that we propose, the persistent activation of the *readout* network is achieved through excitatory projections from both the *encoding* and the *storage* neurons to the *readout* neurons (Figure 4.17). In the memory condition, a bump forms in the *readout* ring due to the excitation provided by the *encoding* ring, and during the mnemonic delay, the *readout* bump is maintained thanks to the excitation provided by the *storage* bump (see Figure 4.18c,f,i). Since it receives excitation from the *encoding* neurons (active during stimulus presentation) and the *storage* neurons (active in the absence of stimulation), the *readout* ring does not need bistability and operates in the monostable regime, as does the *encoding* ring. During the visual condition, the *readout* ring is active due to the input coming from the *encoding* ring. The behavior of the *readout* ring can be described as stimulus but not task-selective, since unlike the other two rings, the *readout* ring is active during both task conditions.

The model's most essential aspect is the different activation profiles of the *encoding* and *storage* neurons. The *encoding* network behaves as a first stage processing of sensory information, hence its label (*encoding*), while the *storage* network is engaged for memory maintenance (at least / or also motor response). As shown below, this alternate activation of these sub-populations underlies the changes observed in the population code (Figure 4.15). The connections must be adjusted appropriately for the information relay between these populations to take place correctly. Too weak excitatory drive from *encoding* to *storage* would not allow the *storage* network to form a bump of activity or would prevent the *storage* bump from forming reliably at the correct location (see below, Figure 4.19). If the excitation is too high, on the other hand, the *storage* bump will form very quickly (no increasing activity during the delay) or lose stability (cite and show figure).

In addition to the basic mechanisms explained, we included excitatory-to-excitatory projections from the *storage* to the *encoding* network that produce sustained selective activity in the *encoding* neurons during the delay (Figure 4.18a,g). This feature can be observed in the *perceptual* neurons in the data (Figure 4.5). Further than replicating the experimental observation, the projections from *storage* to *encoding* have functional implications on memory stability, on which we will comment below and in the section Section 4.3. Since the information flows from the stimulus to the *encoding* network and is relayed to the *storage* units, projections from *storage* to *encoding* neurons are naturally regarded as *feedback* projections.

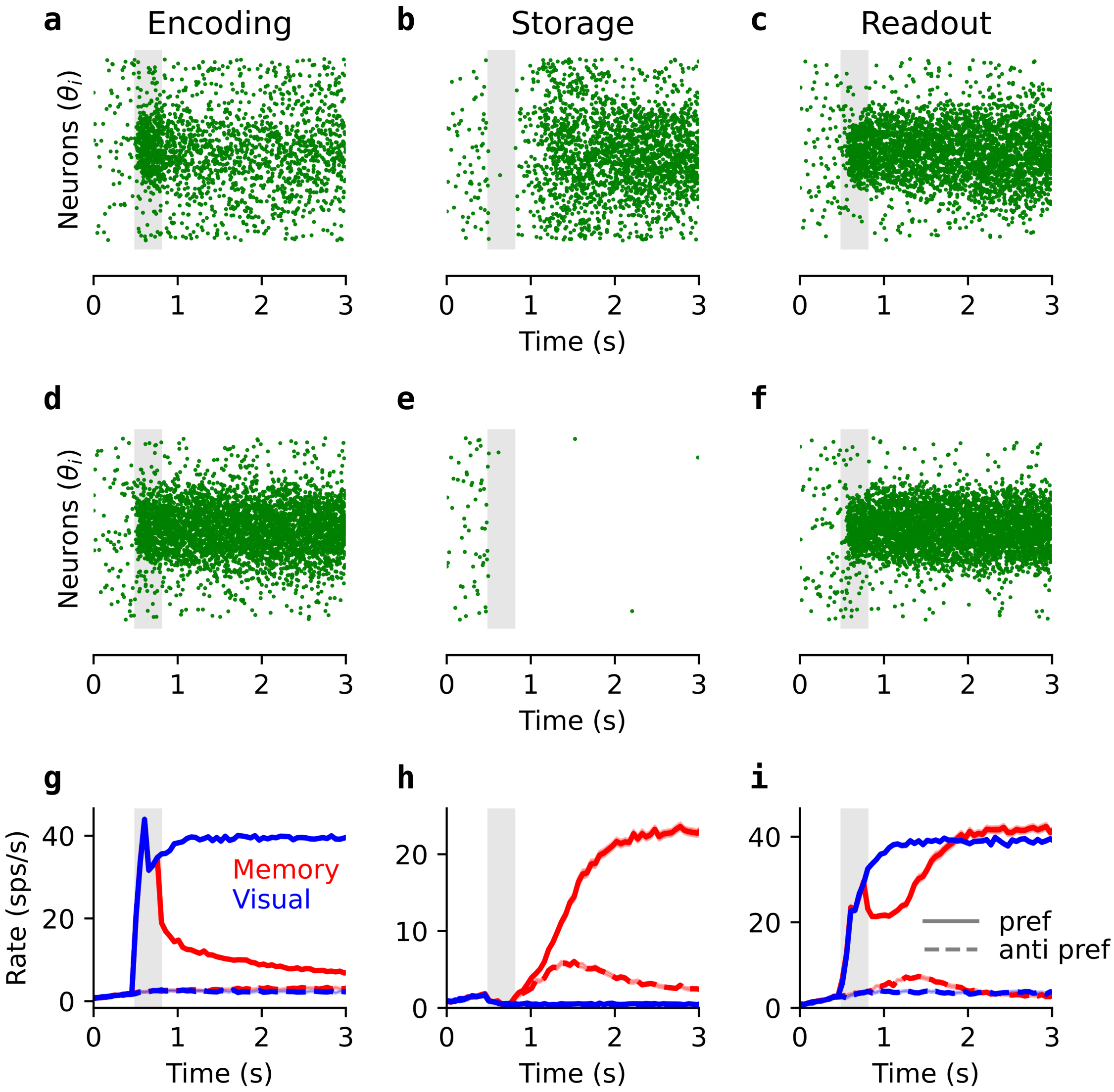

Figure 4.18: Three-ring model reproduces the functionally different activity profiles observed in PFC neural recordings a-f: raster plots, each dot represents a spike, the y-axis corresponds to the neurons arranged according to their preferred location ($[-\pi, \pi)$). **a-c** 3-ring activity during a memory trial, **d-f** 3-ring activity during a visual trial.Gray-shaded regions indicate the time during which the stimulus is presented. **g-i** Average firing rates of selected neurons from the different model networks. The PSTHs are averages over the PSTHs of n neurons for trials where their preferred direction was presented (solid lines) and trials where their *anti-preferred* (at $180°$ from the preferred direction) was presented (dashed lines). Activity for the memory and visual is shown in red and blue, respectively.

4.2.2.1 Top-down feedback projections stabilize memory

The feedback projections from *storage* to *encoding* ring can improve memory stability. Initially, these connections were included to reproduce the delay activity observed in the *perceptual* neurons in the data (Figure 4.5c). However, we observed that the memory error (measured here as the standard deviation of the bump position across trials) decreases when the feedback is included (Figure 4.19).

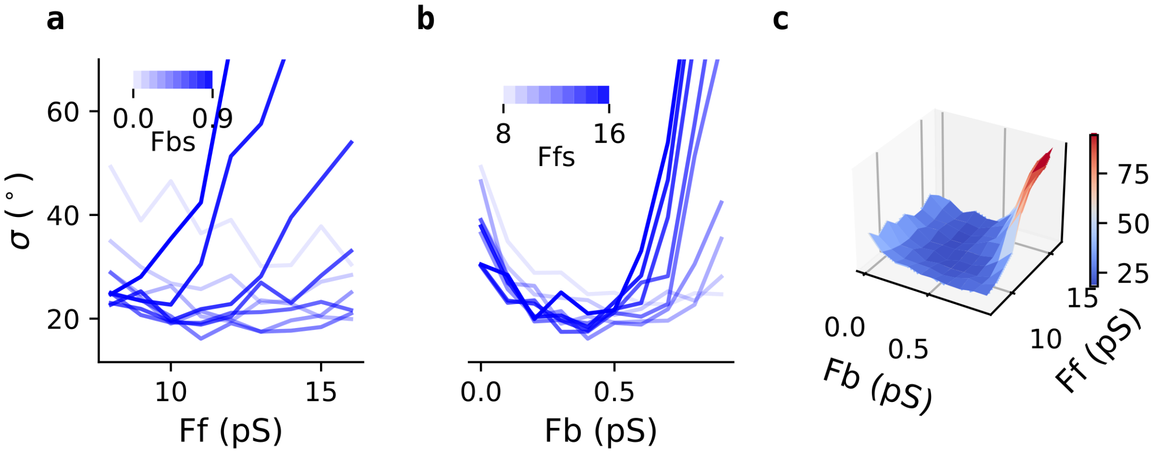

Figure 4.19: Feedback and feedforward projection must be in a certain range to allow for proper memory encoding. **a** Memory error as a function of the amplitude of the feedforward connections from *encoding* to *storage* ring. **b** Memory error for different values of the *storage* to *encoding* feedback amplitude. **c** Surface plot combining **a** and **b**, with the memory error on the $z-$axis.

This effect is due to the increase in the amplitude of the *storage* bump caused by the recurrent loop of *encoding* and *storage* network, which also leads to delay activity of the *encoding* network (Figure 4.20).

Enhanced memory stability is consistent with mathematical analysis of firing rate models showing that the bump diffusion depends inversely on the squared bump amplitude (Esnaola-Acebes *et al.*, 2022; Kilpatrick, 2013). As happens with the strength of the feedforward projections (see above), from *encoding* to *storage* ring, too strong feedback eventually makes the network unstable. As the feedback increases, so does the amplitude of the activity bump in the *encoding* ring (Figure 4.20). Consequently, the *storage* ring gets more excitatory drive from the *encoding* ring and can eventually lose its stability.

Finally, from a network perspective, including feedback projections make the circuit more realistic. The proposed circuit has an explicit feedforward structure (stimulus - *encoding*-*storage*) that contrasts with other WM models that do not consider any structure a priori (Barak *et al.*, 2013; Murray *et al.*, 2017a; Stroud *et al.*, 2023). In our work, we include structure explicitly, motivated by the experimental findings in Section 4.1. Without the feedback projections, our model has a purely feedforward structure for which there is no anatomical evidence. Including the feedback brings the model's structure closer to a realistic scenario.

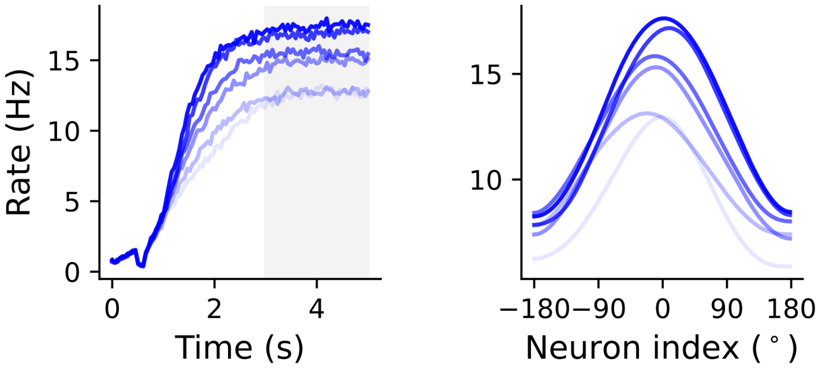

Figure 4.20: The feedback from *storage* to *encoding* circuit increases the amplitudes of the *storage* bump. The *storage* bump's amplitude increases with the top-down feedback as the *encoding* bump increases (not shown). **a** Bump amplitude against time for different values of the feedback (different scales of blue). **b** Population activity profiles calculated as an average over the time span indicated by the gray shading in **a**.

4.2.2.2 Cross-temporal decoding illustrates the subpopulations' contribution to the neural code

A central question we wanted to answer is what underlies the transition in the neural code from cue to delay epoch (Figure 4.2)(Murray *et al.*, 2017a; Spaak *et al.*, 2017; Stokes *et al.*, 2013). As we will illustrate, our model suggests that the decoding pattern (dynamic transition) can be due to different neuronal populations being active at different time epochs (cue, delay) during the memory condition. When we use neurons from all three populations in the model to train a decoder, we obtain a similar result to the one obtained when training the decoder on the stimulus-selective (or all excitatory) neurons in the Markowitz *et al.* (2015) data set (Figure 4.15). Training and testing on the same time bins for the memory condition (Figure 4.21a, *red*), the decoder's accuracy increases from chance level to above 80% when the cue is presented, and it remains above chance during the delay period.

The cross-temporal decoding patterns (Figure 4.21b,c) also resemble the ones obtained for the experimental data (Figure 4.15), highlighting in the same way the changes in the dynamics of the neural code. The memory pattern of the model shares the main properties with the corresponding pattern obtained from the data: the decoding accuracy is high during the stimulus presentation, lower for the across epochs (train cue test delay and vice-versa), and increasingly high for train-test time combinations spanning the middle and late delay ($\sim t > 1$ s). The lower decoding accuracy when training and testing across cue and delay epochs indicates again that the code changes substantially from one epoch to the other, while the capacity the decoder has to generalize (accurately decoding when trained and tested at different times) during the middle and late delay, is related to the stability the code achieves during this epoch. An additional feature the memory pattern shares with the corresponding pattern from the neuronal recordings is the asymmetry in the accuracy of the across-epoch periods. As with the experimental data, the decoder performs better when trained during the delay and tested on the cue period than the other way around.

We can interpret the features observed in the cross-temporal decoding (Figure 4.21) in

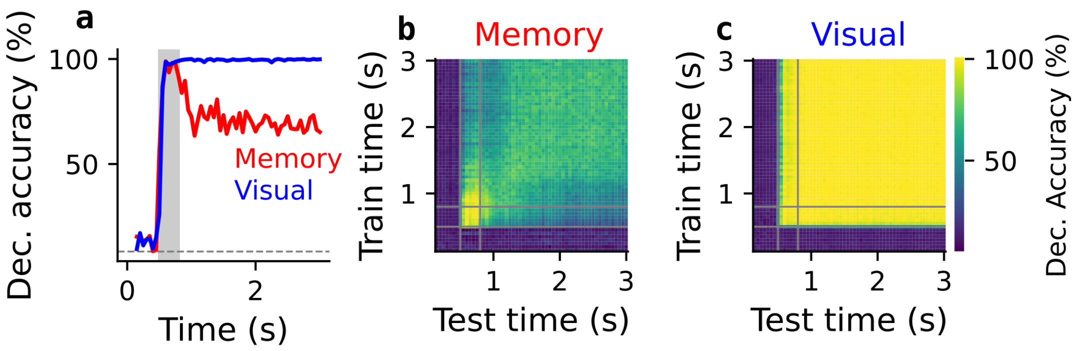

Figure 4.21: Decoder performance across time informs about the dynamics of the WM code **a** Decoding accuracy when training and testing at the same time point (50 ms time bins), memory (red), visual (blue) conditions. The time of stimulus presentation is indicated with gray shading. **b,c** Cross-temporal decoding patterns for the memory and visual conditions. Gray lines delimit the time of stimulus presentation. A selection of 615 neurons out of 6144 was used. 80 trials (10 trials for eight different stimulus locations)

terms of the circuit mechanisms of the model. We performed the same cross-temporal analysis for neurons taken from the different model sub-networks separately (Figure 4.22). Once again, the results parallel the corresponding patterns obtained for the experimental data (Figure 4.16): In the memory condition, the stimulus identity is encoded by the *encoding* neurons while the stimulus cue is present (Figure 4.22). Accordingly, the decoder's performance is high when trained and tested on the *encoding* neurons within the cue epoch (Figure 4.22a,b). During the delay, the activity of the *encoding* neurons is significantly lower, and the decoding accuracy drops. However, due to the feedback from the *storage* network, there is still stimulus-selective activity in the *encoding* ring, which allows for above-chance decoding. When training the decoder on the *encoding* neurons during the cue epoch, decoding the stimulus identity during the delay is worse because the same neurons are still active, but their firing rates are lower (decrease in signal-to-noise ratio). The opposite occurs when training the decoder on the *encoding* neurons during the delay; when tested during the cue, the same neurons are active at a higher firing rate, and the decoding accuracy increases (increase in signal-to-noise).

When decoding from the *storage* ring in the memory condition, the accuracy is close to chance until it increases to $\sim 75\%$ around $t = 1$ s in the delay. (Figure 4.22d). As with the *encoding* ring, the decoding accuracy follows the time course as the bump's amplitude (see Figure 4.18b,h). When the bump is formed, the *storage* ring becomes reliable for decoding the stimulus. Since the bump is stably sustained, training and testing the decoder across the same and different time points during the delay gives a similarly accurate result (cross-temporal generalization in Figure 4.22e).

Decoding the stimulus identity from the *readout* neurons is possible at any combination of train and test times since the presentation of the stimulus (Figure 4.22g,h). During the stimulus presentation, the *readout* units have stimulus-selective activity due to the selective excitation provided by the *encoding* ring, and during the delay, this stimulus-selective activity is maintained by the selective excitation supplied by the *storage* units. In the model, decoding from the *readout* neurons is always the best policy; the decoding is stable throughout the entire task duration, and the accuracy is high because this network reflects a sum of the selective activities in the other two networks.

The dynamics of the code during the visual are more straightforward. Maintaining the external stimulation ensures that a bump of activity is maintained in the *encoding* neurons while the *storage* neurons remain inactive. The *encoding* to *readout* excitation produces a similar bump in the latter ring, with the result that these two populations have a stable code throughout the time course of a trial. Such a stable code allows the decoder to generalize from any time point to another when decoding from the *encoding* and *readout* populations, while the accuracy remains close to chance when decoding from the *storage* neurons.

The cross-temporal analysis for separate populations (Figures 4.16 and 4.22) shows that the change in the WM code from cue to delay epoch can be attributed to the presence of groups of neurons that are mainly active at cue or delay. Our model proposes a possible mechanistic realization of such a structured population arrangement. We will now present some limitations of the possible formulation of the model and show how its initially proposed architecture can be modified to account for further experimental observations.

4.2.2.3 Limitations of the model: adaptation to different delay lengths

The presented 3-ring model (*bistable network*) captures essential qualitative aspects of the experimental data. The time courses of the *encoding*, *storage*, and *readout* model neurons qualitatively match the ones of the *perceptual*, *mnemonic*, and *persistent* neurons in the data. On the population level, using cross-temporal decoding on both the experimental data and the model, we also obtain similar results.

However, with the presented circuit architecture, the activity in the *storage* network cannot adapt to different delay lengths. Gradually increasing activity during the memory delay has been observed in similar task paradigms for different species (Emmons *et al.*, 2017; Finkelstein *et al.*, 2021; Funahashi *et al.*, 1989; Fuster & Alexander, 1971; Inagaki *et al.*, 2019). This ramping activity could be tracking the passage of time (Durstewitz, 2003; Paton & Buonomano, 2018; Simen *et al.*, 2011) or reflecting urgency or motor preparation (Carland *et al.*, 2019; Thura, 2020). In our presented model architecture, this phase of increasing activity corresponds to the transition from a stable homogeneous state of baseline activity to a stable structured activity of bump state. The dynamics of this transition depend on the level of excitation provided by the *encoding* ring to the *storage* ring and on the internal dynamics of the *storage* ring. Stronger feedforward connections, as well as stronger recurrence in the *storage* ring contribute to faster transitions (shorter ramping period). The transition of the *storage* neurons can span a delay length of ~ 1 s (Figure 4.18), but it is challenged when delays of different lengths are to be taken into account (Figure 4.23). Adapting this transitory dynamic to span different delay lengths would imply a high encoding error, as a slower bump formation would be necessarily associated with a higher encoding error (the bump would form in an already biased position and additionally diffuse). the slowly forming bump is more vulnerable to diffusion (Esnaola-Acebes *et al.*, 2022; Kilpatrick, 2013), and also to include plasticity mechanism that would adapt the synaptic weights to the task condition, so the feed-forward and the recurrent dynamics are adequately modified.

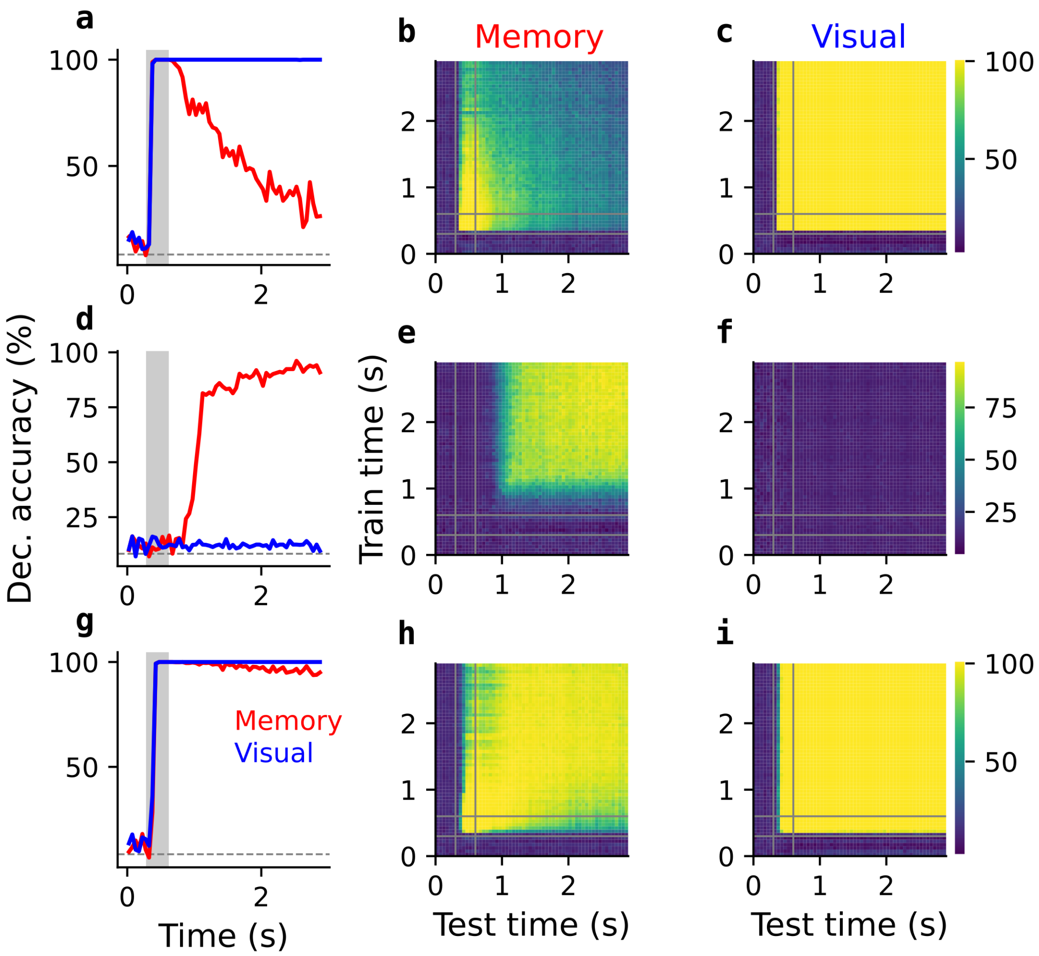

Figure 4.22: Cross-temporal decoding for the three rings separately (*bistable network*) a-c *encoding* ring separately: **a** testing and training at the same time point for the two task conditions. **b,c** Cross-temporal patterns for the memory (**b**) and visual (**c**) conditions. **d-f** same for the *storage* ring. **g-i** same for the *readout* ring.

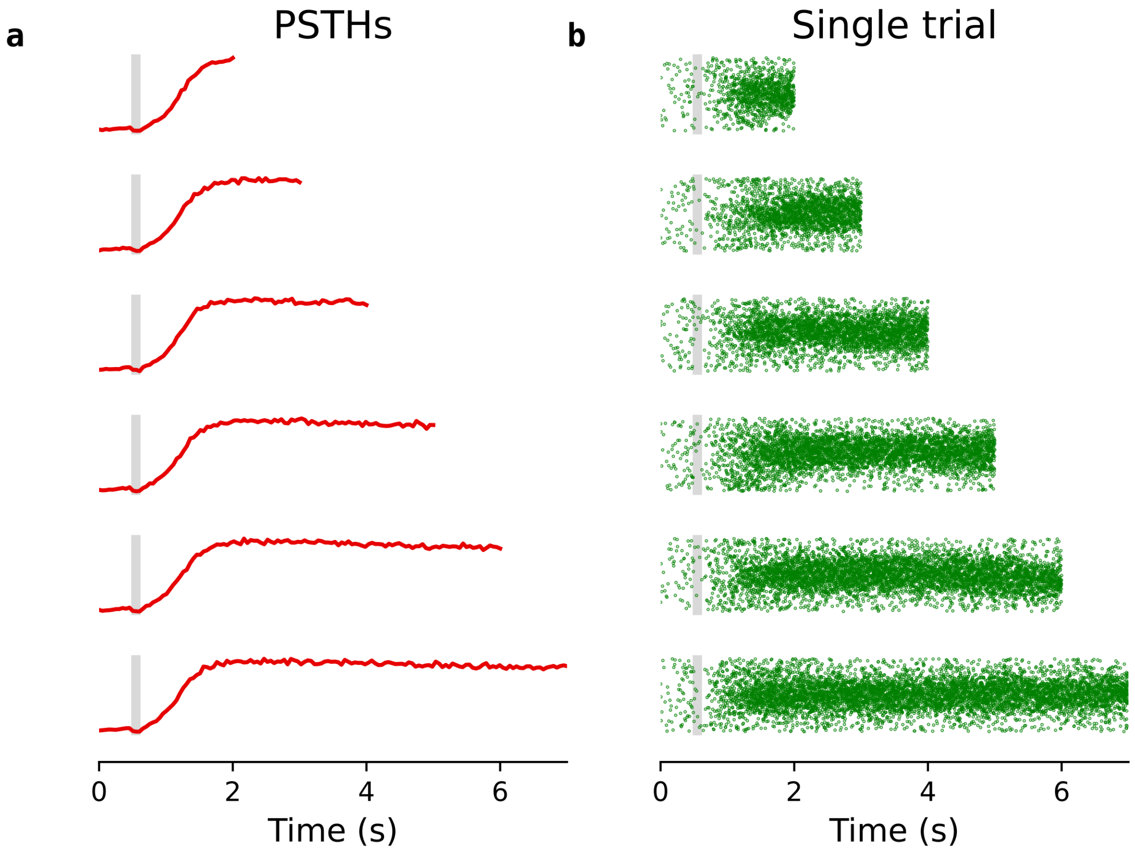

Figure 4.23: *storage* neurons in the 3-ring model cannot adapt their activity to variable delay lengths. Activity of the *storage* neurons in the 3-ring *bistable network* model for trials with different delay duration. **a** average activity for neurons aligned with stimulus location for 100 trials. **b** Raster plot showing the spiking activity of the *storage* ring for a trial example of each delay duration. Top to bottom, $2, 3, 4, 5, 6, 7$ s delay.

This shortcoming of the model motivated us to find an alternative mechanism to explain the dynamics of the *mnemonic* neurons. We introduce a variant of the 3-ring model with an additional external input, which can flexibly adapt the gradual response of the *storage* neurons to delays of different lengths. We expose both models as alternative circuit hypotheses because although they share the same 3-population structure, they implement memory maintenance in two conceptually different ways.

4.2.3 A three-population model with external ramping signal

4.2.3.1 Ramping activity: evidence and origins

Neural ramping activity, as the one exhibited by the *mnemonic* neurons (Figure 4.5) during the delay of the memory, has been observed in experimental paradigms with predictable delay time (Emmons *et al.*, 2017; Finkelstein *et al.*, 2021; Hamilos *et al.*, 2021). Some authors suggest that subcortical regions such as the basal ganglia (Chiba *et al.*, 2015; Jin *et al.*, 2009; Paton & Buonomano, 2018), or the locus coeruleous (LC)(Berridge & Waterhouse, 2003) could be supplying frontal cortical areas with information about elapsed time, motor response preparation, or urgency (Carland *et al.*, 2019; Thura, 2020). For example, the LC produces norepinephrine (NE) that eventually interacts with the receptors in cortical neurons, modifying their synaptic and leak current conductances (Berridge & Abercrombie, 1999; Berridge & Waterhouse, 2003). The LC, located in the pons (brainstem), is known to respond to arousal, stress, and panic (Berridge & Abercrombie, 1999; Ross & Van Bockstaele, 2021), which makes it a good candidate to provide an urgency or response anticipation signal.

As an alternative explanation, some authors consider the hypothesis that ramping activity might be intrinsically generated by the group of cortical neurons (or region) where it is observed (see Finkelstein *et al.* (2021); Hansel & Mato (2013); Inagaki *et al.* (2019))

We wondered whether the ramping of the *mnemonic* neurons is intrinsically produced by the region where the recordings were made (PFC) or if it is instead generated by a different area that projects to PFC (see Finkelstein *et al.* (2021) for a discussion on this question). In the first version of the 3-ring model shown above, we mimicked the ramping of the *mnemonic* neurons with the increase in firing rate of the *storage* neurons as the stable bump forms. In light of the discussion about the origins of the ramping activity, our first approach proposes an intrinsic mechanism of ramping activity generation. However, we have already commented on some limitations of this mechanism, as it does not flexibly adapt to diverse delay lengths (see subsection above) In this section, we will assume that a given area in the brain provides PFC with a stimulus-agnostic signal that increases linearly with time. We will not describe the mechanisms that produce the linearly increasing signal in the first place (see Berridge & Abercrombie (1999); Durstewitz (2003); Hamilos *et al.* (2021) for the discussion of different possible mechanisms), but assume it is provided. We will show that such a signal can cause a linear increase in the amplitude of the *storage* bump that resembles the *mnemonic* neuron profile in the memory condition.

We implemented the ramping signal through a modulation of the NMDA synaptic conductance. We make similar assumptions as in Eckhoff *et al.* (2009). First, the signal is provided by a subcortical area that projects to PFC and increases the concentration of a neurotransmitter that modifies the prefrontal synaptic conductances. [1] We assume that neurotransmitter release

[1] A documented example of this type of modulation is the effect of the Locus Coeruleus (situated in the

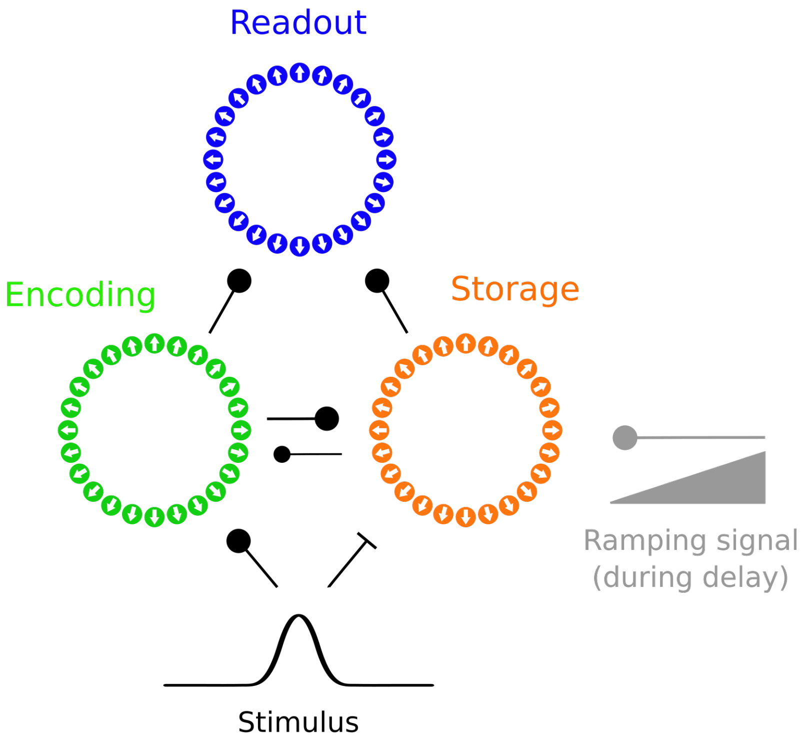

Figure 4.24: 3-ring model with external ramping signal: The model's architecture similar to the one of the *bistable network* model. An additional ramping input is provided to the *storage* network in the shape of positive modulation of the NMDA-meditated synaptic gain. In this model, the *storage* network operates in the monostable regime.

depends linearly on the activity of the subcortical area (as has been reported experimentally (Berridge & Abercrombie, 1999)) and that synaptic conductance (in our case, NMDA-mediated synapses) changes linearly with the neurotransmitter concentration. Particular to our network implementation is the assumption that the subcortical area providing the modulation undergoes a linear increase in activity during the delay in the memory condition. This last assumption makes our proposal compatible with ramping as an urgency or response preparation mechanism (Berridge & Waterhouse, 2003; Cisek, 2019; Finkelstein *et al.*, 2021; Thura, 2020). A schematic view of the model is shown in Figure 4.24

Besides including external stimulation, changing the dynamic working regime of the *storage* neurons was necessary to capture more precisely the time course of the *mnemonic* neurons. In the bistable regime, the ring model reaches the bump state (high-activity heterogeneous state) through a subcritical bifurcation. In this regime, the bump forms with higher (finite) amplitude, and further increases in amplitude due to external input (gain modulation in our case) lead to relatively small increases in activity and eventually to the instability of the bump (fig, Figure 3.6). For lower values of the positive recurrent connections, the ring is monostable, and upon stimulation, the network transitions to the bump state through a supercritical bifurcation. In this monostable regime (*encoding* and *readout* rings are monostable), the bump is formed with an arbitrarily small amplitude (see figure Figure 3.6), and its amplitude can increase

Pons, brainstem) on the concentration of norepinephrine in the prefrontal cortex, which is known to underlie changes in glutamatergic synapses, GABA, and membrane leak conductances (Berridge & Abercrombie, 1999; Berridge & Waterhouse, 2003; Eckhoff *et al.*, 2009).

proportional to the gain without losing stabilit.

This change in the operation regime of the *storage* network is of conceptual importance since it implies that WM can depend on the interaction of PFC with subcortical regions. In the first version of the 3-ring model, presented in section Section 4.2.2, the *storage* ring sustains the stimulus identity during the memory delay thanks to its bistability. With the *storage* ring operating in a monostable regime, our PFC circuit can no longer sustain the memory autonomously. In this regime, the external *stimulus-independent* gain modulation needed to produce the ramping activity is also necessary to maintain stimulus-selective activity (the bump) during the delay. So stimulus identity is provided to the *storage* ring by the *encoding* neurons, but maintenance of stimulus information is ensured by the external gain modulation, which itself does not depend on the stimulus.

In contrast to the models that describe WM as a function relying primarily on the dynamics of a single network (Barak *et al.*, 2013; Compte *et al.*, 2000; Murray *et al.*, 2017a; Stroud *et al.*, 2023), our model with external gain modulation aligns with circuits proposals in which memory maintenance also depends on the interaction between different non-bistable networks (Ashby *et al.*, 2005; Feng *et al.*, 2023; Jaramillo *et al.*, 2019; Mejías & Wang, 2022). These models are also closer to the spreading view of WM as a function relying on distributed areas across the brain (see Christophel *et al.* (2017) and Chapter 5).

4.2.3.2 Dynamics during the visual and memory conditions

Since the gain modulation (ramping input) is only applied during the delay (only present in the memory condition), the model activity during the visual condition is the same as for the 3-ring model without ramping (Section 4.2.2, Figure 4.18). During the memory condition, the external gain modulation is introduced after removing the stimulus. At the beginning of the delay, already under the effects of the neuromodulation, the *storage* population undergoes an initial increase in firing (Figure 4.25b,h) that corresponds to the formation of the bump (through a supercritical bifurcation). Until the intrinsic dynamics of the *storage* ring have produced a structured bump (see Figure 3.6) the gain modulation causes a similar increase in firing rate to all the neurons in the network (Figure 4.25b,h). Once the bump has formed, its amplitude keeps increasing due to the linear modulation of the NMDA conductance, while the bump tails are kept at a low firing rate thanks to the surround suppression connectivity structure (Figure 3.5). It is worth noting that the intrinsic dynamics of the *storage* network, even though not strong enough to sustain the bump without stimulation, are essential for proper stimulus encoding. A network with lower recurrency would not maintain a structured bump activity upon stimulus removal. It would eventually transition to a spatially flat activity profile under the effect of the modulation.

4.2.3.3 Cross-temporal decoding in the *ramping network* model

The results of the cross-temporal decoding analysis are similar to the ones obtained for the model with no ramping input (Figures 4.21 and 4.22). The patterns corresponding to the visual condition are identical (Figure 4.26c, Figure 4.27c,f,i), since the activity of the three populations is the same. Likewise, the activity of the *encoding* is also the same in the two circuit architectures, which produces the same cross-temporal decoding pattern. During the delay of the memory condition, the linearly ramping activity of the *storage* neurons produces a

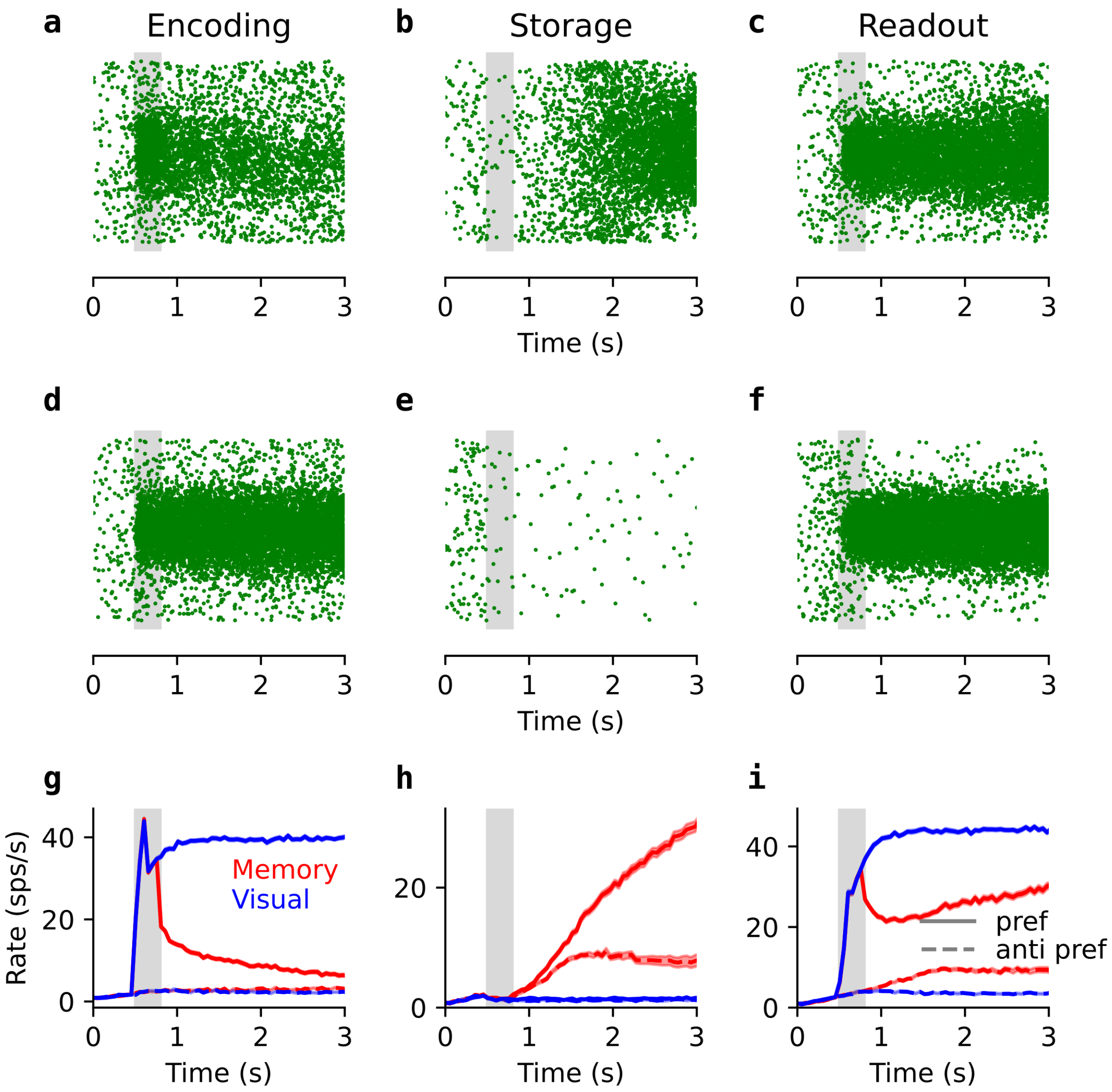

Figure 4.25: 3-ring model reproduces functionally different activity profiles observed in PFC neural recordings Average firing rates for the memory task in red and visual in blue. Response to preferred cue in solid lines, to anti-preferred (180 °) in dashed. **a-f**: raster plots, each dot represents a spike, the y-axis corresponds to the neurons arranged according to their preferred location ($[-\pi, \pi)$). **a-c** 3-ring activity during a memory trial, **d-f** 3-ring activity during a visual trial. Gray-shaded regions indicate the time during which the stimulus is presented. **g-i** Average firing rates of selected neurons from the different model networks. The PSTHs are averages over the PSTHs of 100 neurons for trials where their preferred direction was presented (solid lines) and trials where their *anti-preferred* (at 180° from the preferred direction) was presented (dashed lines). Activity for the memory and visual is shown in red and blue, respectively.

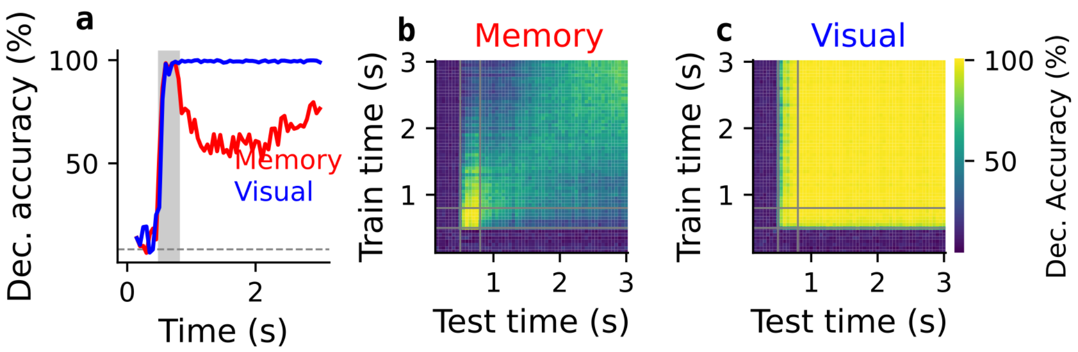

Figure 4.26: Decoder performance across time informs about the dynamics of the WM code **a** Decoding accuracy when training and testing at the same time point (50 ms time bins), memory (red), visual (blue) conditions. The time of stimulus presentation is indicated with gray shading. **b,c** Cross-temporal decoding patterns for the memory and visual conditions. Gray lines delimit the time of stimulus presentation. A selection of 615 neurons out of 6144 was used. 80 trials (10 trials for 8 different stimulus locations)

square pattern of increasing accuracy (Figure 4.26b,Figure 4.27e). This increase in accuracy during the delay, which was faster in the model with no ramping (Figure 4.21b, Figure 4.22e), is also observed in the decoding of experimental data we analyzed (Figure 4.16) as well as in other data sets (Spaak *et al.*, 2017).

Decoding from the *readout* units gives a qualitatively equivalent result in both circuit architectures (*bistable network*, *ramping network*). In this circuit, the transient decay in the activity of the *readout* ring during the memory is more noticeable than it is for the previous model proposal and becomes more apparent for longer delays. However, it does not alter the results qualitatively.

Notably, the orthogonal transition in the WM code from cue to delay epochs is explained by the same mechanism as in the previous circuit: *encoding* neurons are mostly active during stimulus presentation while *storage* neurons fire during the mnemonic delay. Our work still emphasizes the potential relevance of considering distinct functional subpopulations in the context of WM over the exact mechanistic details of the respective populations.

4.2.3.4 Ramping activity spanning different delay lengths

We finally show how the ramping activity can flexibly adapt to delays of different lengths (Figure 4.28). When increasing the delay duration, the rate of gain modulation increase (r_{NE} in Equation (3.20)) must be adjusted accordingly for the *storage* activity to span the duration of the delay. We do this by ensuring that the modulation level achieved by the end of the delay is the same for different delay lengths. The result is that the ramping slope or rate of activity increase depends on the delay length, and the *storage* bump always attains the same amplitude by the end of the delay (Figure 4.28). This adjustment of the modulatory input would be expected if the ramping signal was related to urgency or if the cortical activity was

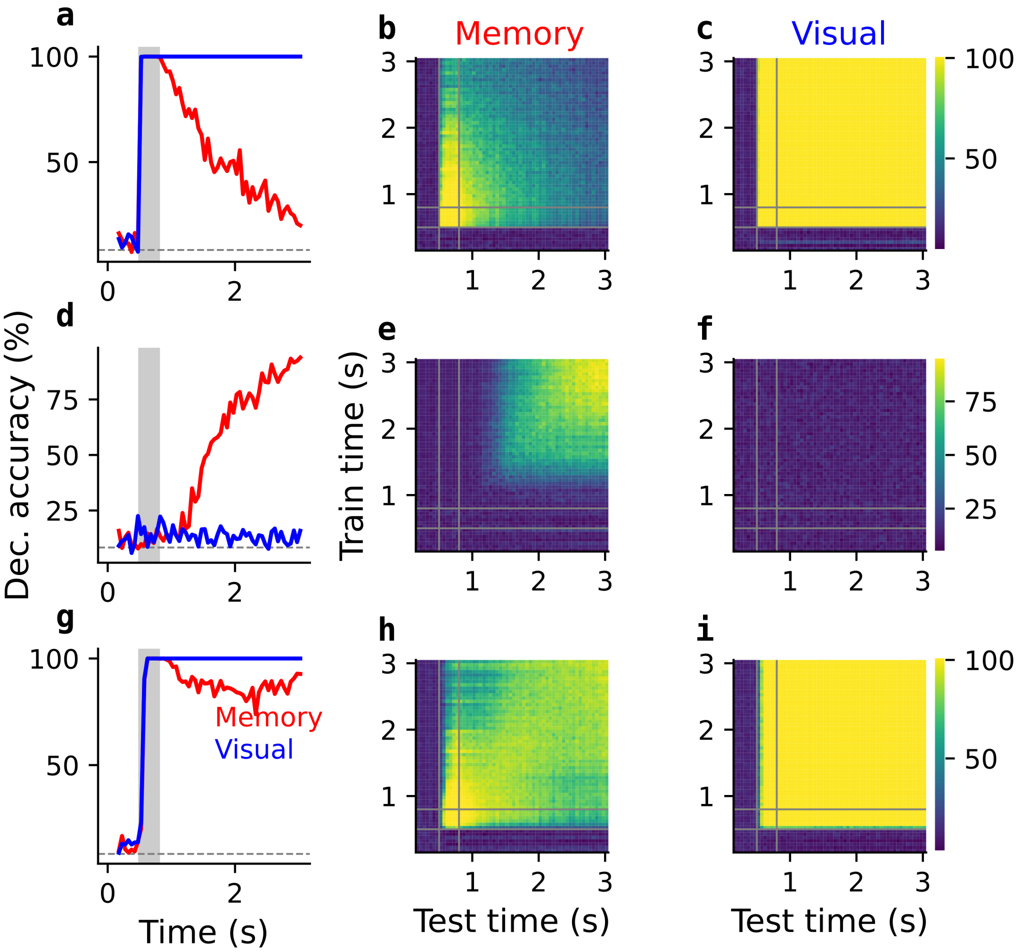

Figure 4.27: Cross-temporal decoding for the 3 rings separately (*ramping network*) a-c *encoding* ring separately: **a** testing and training at the same time point for the two task conditions. **b,c** Cross-temporal patterns for the memory (**b**) and visual (**c**) conditions. **d-f** same for the *storage* ring. **g-i** same for the *readout* ring.

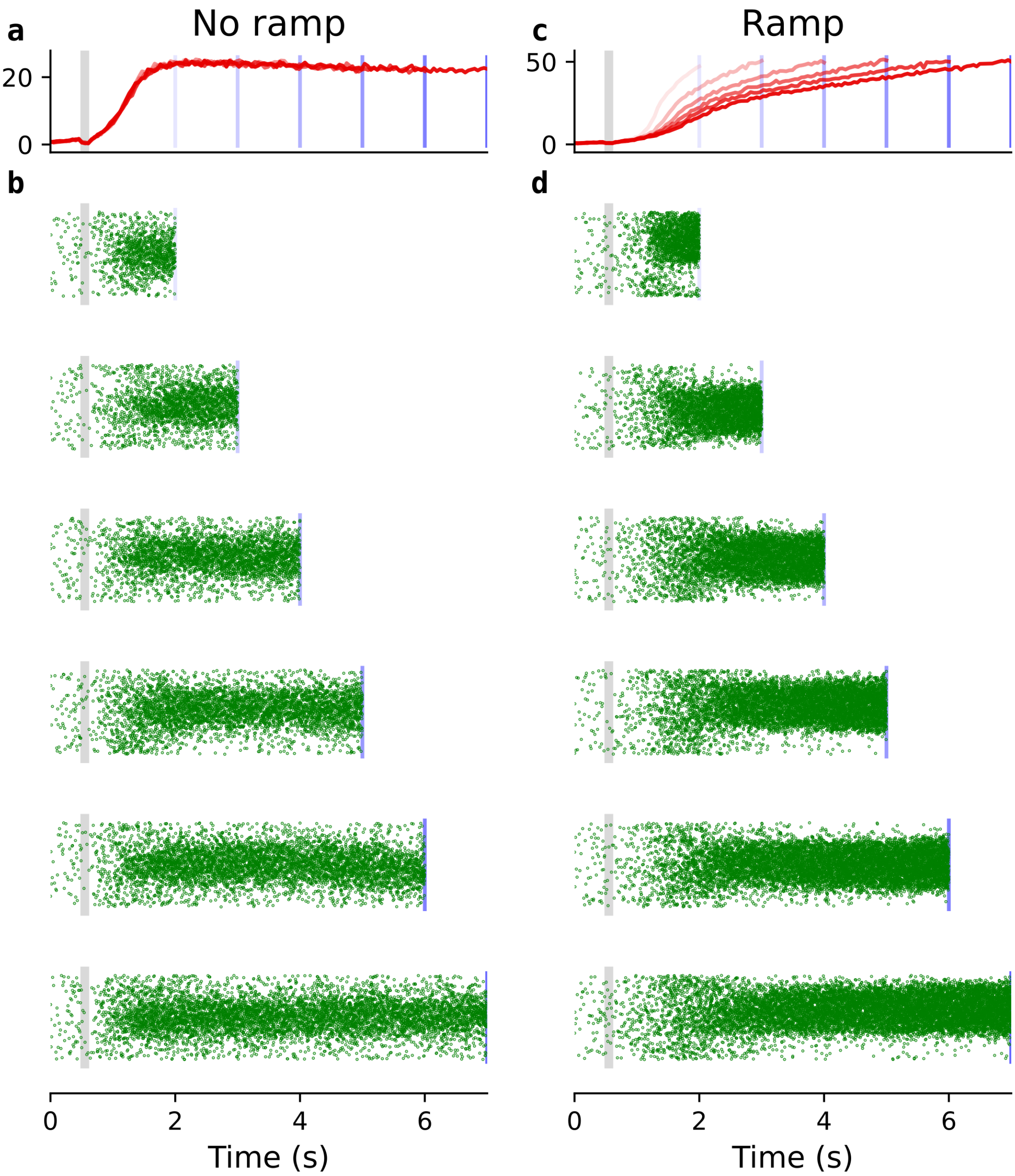

Figure 4.28: Model with external ramping signal adapts to variable delay length. a,b 3-ring model with no ramping (*bistable network*) for different lengths of the delay. **a** PSTHs and **b** single trial examples (raster plots) for. **c,d** 3-ring model with ramping signal (*ramping network*) **c** PSTHs and **d** single trial examples (raster plots). All PSTHs show the average activity of neurons aligned whose preferred location is aligned with the stimulus for 100 trials.

to reach a certain threshold to initiate the motor plan. If this were the case, different delay lengths coupled to the same motor plan should be associated with equal activity levels before the motor response (Emmons *et al.*, 2017).

Conclusions

Our three-network model relates the presence of subpopulations to the qualitative changes from cue to delay in the working memory code. In particular, the dynamics are associated with the activity of the *encoding* neurons, which fire during cue presentation, and the *storage* neurons, which fire during the delay. The model implements a feedforward structure (stimulus excites *encoding* neurons, which then excite *storage* neurons) that is proposed to be an essential property of the prefrontal network structure.

A network variation without bistable subnetworks, suggests that areas external to PFC (probaby subcortical areas such as Basal Ganglia or Locus Coeruleus) can provide non-selective stimulation and contribute to memory maintenance. This distributed interaction can increase the model's flexibility, allowing it to capture ramping activity that spans variable delay lengths.

Both model variations rely on attractor dynamics for appropriate stimulus-selective response and maintenance, proposing that the relevance of attractor dynamics is not at odds with the dynamic changes in the population code.

4.3 Distinct populations for encoding and maintenance improve the resistance against distracting stimuli

Section summary

In this section, we illustrate how our model behaves when distracting stimuli are presented. Our model naturally inherits some of the properties that bump-attractor networks exhibit in the presence of distractors. As a consequence, our model captures the observed dependence of the distractor impact on the distance between the target and the distractor stimulus. However, our model makes further predictions that are directly related to its subnetwork structure. In particular, we analyze how the time of distractor presentation and the amplitude of the top-down feedback connections (*storage* to *encoding*) affect the network's vulnerability. As observed in experimental studies (Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013), distractors have a greater impact on the model when presented shortly after target stimulus removal than when presented later during the delay. On the other hand, the feedback increases the overall robustness to distraction by enhancing the filtering capacity of the stimulus encoding network (*encoding* neurons).

We analyze the behavior of both network variations, *bistable network* and *ramping network*, in the presence of distracting stimuli and obtain qualitatively similar results.

4.3.1 Distracting stimuli

The ability to hold the memory in the presence of distractors is a crucial aspect of WM, which should be considered by any WM model (Lorenc *et al.*, 2021), see Chapter 1. We analyzed the effect of the impact of a presented distractor on the memory ($\Delta\theta = \theta_C - \theta_{\text{readout}}$, where θ_C is the angle at which the cue stimulus is presented and θ_{readout} is the readout at the end of the delay) depending on the time (t_D) and angle (θ_D) at which the distractor is presented. The distracting stimuli we used are modeled in the same way as the cue stimulus (excitatory projections to *encoding* network and excitatory projections to inhibitory neurons in the *storage* network) with the same Gaussian profile, and they are always presented after the cue stimulus (no presentation before nor coinciding with the time of cue presentation). The differences between a cue and a distracting stimulus in our network are their time (t_C, t_D) and the angle (θ_C,θ_D) of presentation. In all simulations, we have used distractors with the same amplitude as the stimulus.

In general, a distracting stimulation that does not affect the behavior must be processed differently from the way the task-relevant (cue) stimulus is processed. In the case of visual stimuli, the information that comes through the retina travels to the lateral geniculate nucleus (LGN) in the thalamus, before it gets to the primary visual cortex (V1) (Kandel *et al.*, 2000). At which point and how a distracting stimulus starts to be processed differently from the cue (filtered) might depend on the sensory modality and specific task requirements, and it is the focus of numerous studies (Finkelstein *et al.*, 2021; Lorenc *et al.*, 2018; Murray *et al.*, 2017b; Rademaker *et al.*, 2019; Yoon *et al.*, 2006).

In our model, we assume the distracting stimulus reaches our PFC networks unaffected (that is, as the cue stimulus), and the circuit's robustness depends entirely on the dynamics of the *encoding* and *storage* networks. We will show that if there is still a bump in the *encoding* ring when the distractor is presented, it produces a first averaging effect in favor of the original stimulus-triggered memory. The *storage* ring produces an additional averaging effect (on the already biased stimulation coming from the *encoding* ring), which will depend again on the existence and amplitude of its activity bump. Critical for the resistance to the distractor is whether a *storage* bump is already present or not at the moment of distractor presentation

All the results shown in this chapter were obtained with the mean-field firing rate model. Using the firing rate model allowed us to simulate all the different combinations of distractor timing and position and additional parameter searches with high efficiency. Based on preliminary simulations, we infer that equivalentcan be obtained with the LIF model.

4.3.2 Distractor impact decreases as a function of time and is reduced by the top-down feedback

We found that the distractor's impact depends on the presentation time, with early presentations during the delay having more impact, and the feedback (*storage* to *encoding*) makes the model more robust to distractors presented throughout the delay duration.

To understand these results, we need to consider a fundamental element of the dynamics described: how an activity bump behaves when a stimulation is presented. The interaction between the bump and the distracting stimulation will generally depend on the network properties, mainly specified by the shape of the connectivity footprint and the single neurons' transfer function. To understand the following results, we will briefly mention the dynamics of a single ring network (how a single network can behave) in the presence of a distracting stimulus, already studied elsewhere (Compte *et al.*, 2000). A distractor can produce the extreme responses of complete distraction (the bump moves to the distractor location or the original bump disappears and a new one forms at the distracted location) and complete robustness (the distractor does not displace the original bump) as illustrated in Compte *et al.* (2000). These cases are, however, restricted to distractors presented far ($\sim 120°$) from the original cue and to specific parameter values such as exceptionally high or low stimulus amplitude (see Compte *et al.* (2000)). In the general scenario we are considering, a bump can be biased towards (attraction) or away from (repulsion) an external stimulation (Almeida *et al.*, 2015). The attraction is caused by the overlap of the excitation-dominated regions of the original bump and the bump produced by the distraction. Repulsion, in turn, occurs when the distractor is presented at locations that are suppressed by the original bump (surround suppression, Figure 3.5). Consequently, attraction happens for nearby distractors and repulsion for far distractors (Almeida *et al.*, 2015). The ranges in which attraction and repulsion occur will depend on the width of the Mexican-hat synaptic profile and, eventually, on the width of the bump.

In our model, due to the width of the bump, distractors presented at an angle below $\sim 120°$ overlap with neurons active above baseline and have an attractive effect. Above $\sim 120°$ the distractors lie on the bump tails and are suppressed (Figure 3.5). In the attractive range, the effect is an “averaging”: the distractor moves the bump to a position between θ_C and θ_D. The bias in the bump position will be smaller the bigger the amplitude of the original bump (Esnaola-Acebes *et al.*, 2022; Kilpatrick, 2013).

4.3.3 Illustration on a single trial

We consider the activity of a network during an example trial during which a distractor is presented. For illustration, we first show the model's behavior when there is no feedback from the *storage* ring (Figure 4.29). If the distractor is presented shortly after the cue stimulus has been removed (Figure 4.29a), the bump produced by the stimulus is still active, producing an averaging effect. For later distractor presentations, in the absence of feedback, the *encoding* bump has faded and the distractor produces a bump of activity on the exact location where it is presented (Figure 4.29c). When the *encoding* ring perfectly encodes the distractor, the distractor resistance depends entirely on the *storage* ring. By itself, the *storage* ring behaves similarly; it will be biased by distracting stimulation (coming from the *encoding* ring) to a greater or lesser extent, depending on the amplitude of its bump. As with the *encoding* ring, the *storage* ring is most vulnerable when a distractor input arrives before the activity bump at the original cue position had the time to form (Figure 4.29b). In this case, the biased input provided by the *encoding* ring will strongly bias the position of the *storage* bump (Figure 4.29b). In contrast, an already stable *storage* bump offers resistance against the biased input of the *encoding* ring, producing a lesser impact (Figure 4.29d).

The feedback input from the *storage* to the *encoding* ring increases the model's robustness against distracting input (Figure 4.30). As a consequence of the feedback, the *storage* bump will sustain the *encoding* bump throughout the remaining part of the delay (Figure 4.30a,c). The resistance of the *encoding* bump against the distracting stimulus will decrease the distractor impact (Figure 4.31c). In addition to ensuring this first filtering, the feedback indirectly increases the amplitude of the *storage* bump, improving its stability against the distracting

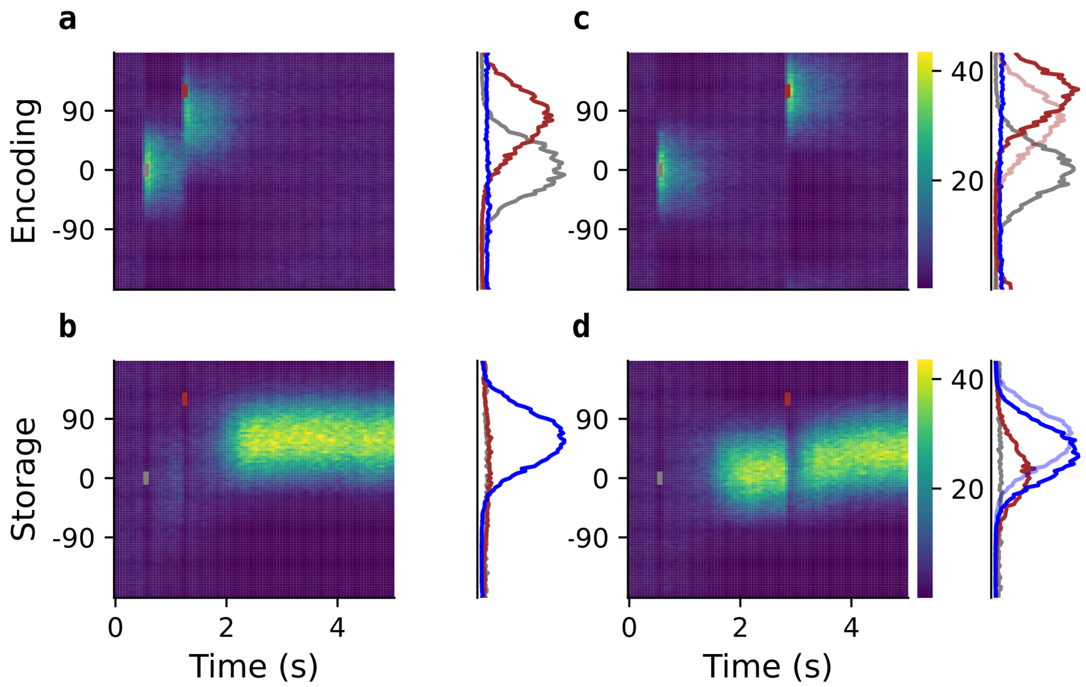

Figure 4.29: Bump dynamics in response to distracting stimuli for the *bistable network* model without feedback. Dynamics of the *encoding* and *storage* networks when a delay is presented at $\theta_D = 120°$. **a**,**b**, left: distractor presented to the circuit at $t = 1.2$ s. Network activity of the *encoding* ring (**a**) and *storage* ring (**b**). Colored rectangles indicate the time and position of the stimulus (gray) and distractor (brown). **a**,**b**, right: corresponding population activity profile after stimulus presentation (gray), after distractor presentation (brown), and at the end of the trial (blue). **c**,**d**: same as in **a**,**b** for a distractor presented at $t = 2.8$ s. **c** and **d**, right, additionally show the activity profile after distractor presentation (light brown), and at the end of the delay (light blue) corresponding to the example in the left ($t_D = 1.2$ s), for comparison. In **a**,**b**, the distractor impact is mildly attenuated by the fading activity of the *encoding* ring, and the *storage* ring forms at the already biased position indicated by the *encoding* bump. In **c**,**d** the *encoding* ring is completely biased by the distractor, but the *storage* bump is already formed at the cue location, and filters the distractor's effect

input. In this way, the *storage* bump contributes to the model's robustness directly by increasing the distractor filtering of the *storage* ring and indirectly by increasing the filtering done by the feedback-sustained *encoding* ring (Figure 4.30).

4.3.4 Distractor impact as a bias in the position of the bump

An alternative way of looking at these results is to observe the time course of the bump position of the *encoding* and *storage* rings(Figure 4.31).

During the cue presentation, the *encoding* ring forms a bump centered at the presented

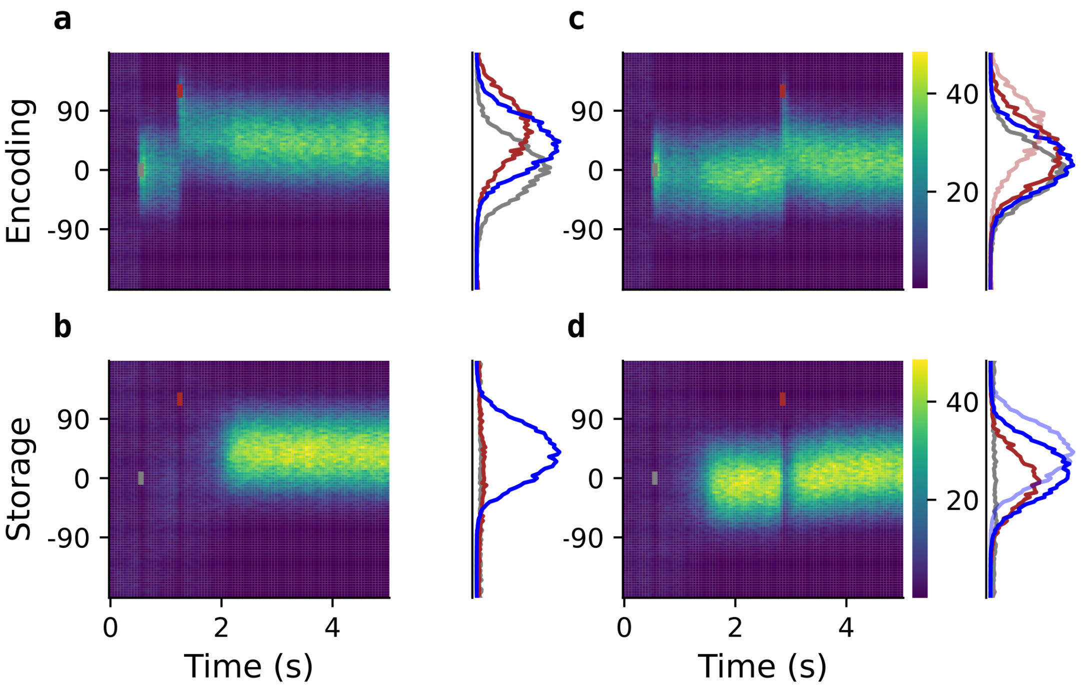

Figure 4.30: Bump dynamics in response to distracting stimuli for the *bistable network* model with feedback. Same examples as in Figure 4.29 for a circuit with feedback ($Fb = 3$. In **a,b** is similar as in Figure 4.29a,b because the feedback only has an effect when the *storage* bump emerges. In **c,d**, the bump in the *encoding* ring first filters the distractor, which is further filtered by the *storage*.

location (0 ° for all trials in Figure 4.31), while there is no clear readout from the stimulus-suppressed *storage*. When the distractor is presented, the *encoding* bump updates its position fast (Figure 4.31a, c). In the absence of feedback, later distractor presentations cause higher biases in the position of the *encoding* bump (Figure 4.31). In the *storage* ring, the increasing amplitude of the bump produces smaller biases for later distractor presentations (Figure 4.31b).

When feedback is included, the increasing bias effect on the *encoding* ring is reverted (Figure 4.31c). The feedback maintains the *encoding* bump, which makes the network more robust. At the same time, the feedback enhances the robustness of the *storage* network (Figure 4.31d), which benefits both from the less-biased *encoding* input and from the feedback-mediated increase in the amplitude of its bump.

The effect of the feedback drastically improves the robustness against distractors presented at increasingly distant angles (Figure 4.32c,d). The presence of the *encoding* bump reduces the impact of the distractor (Figure 4.32c), which is perfectly encoded in the absence of feedback (Figure 4.32a). Distractors presented farther away from the cue location have a decreasing impact, as they overlap less with *encoding* bump (Figure 4.32d).

Overall, the model is more vulnerable to distracting stimuli at the beginning of the delay,

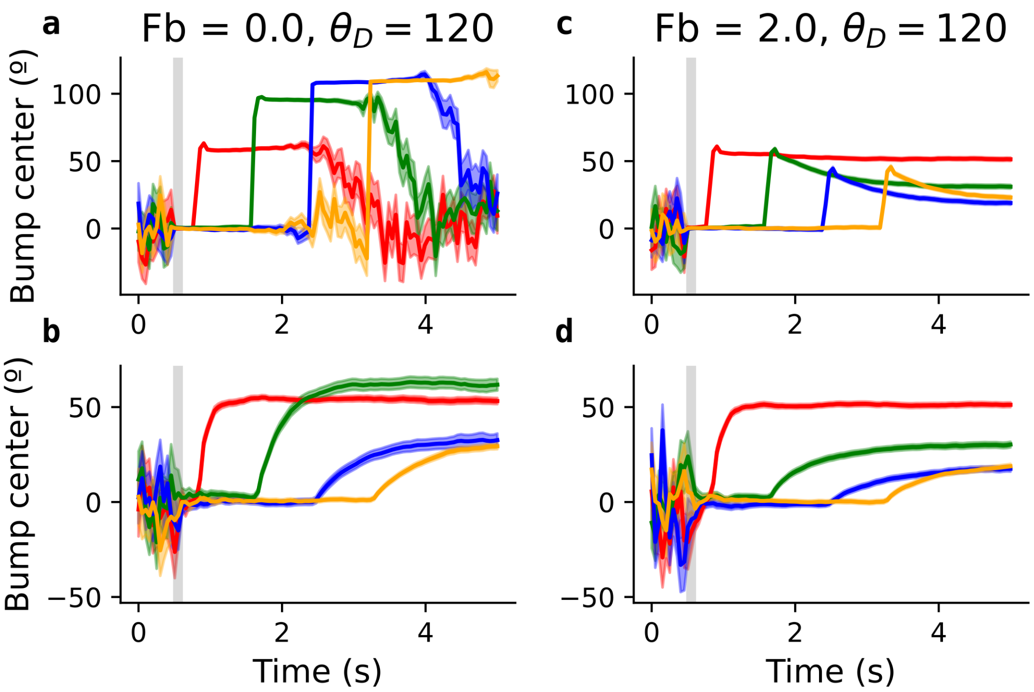

Figure 4.31: The *storage* bump, together with the feedback, increases the robustness against distracting stimuli presented at different times during the delay. Bump center during example trials corresponding to distractor presentations at different times during the task. The lines become noisy when there is no bump in the network (**a** during delay, **a-d** before cue presentation). **a,b** distractor presented at 120° to the circuit without feedback. **c,d** same for the circuit with feedback ($Fb = 2$).

and the vulnerability throughout the delay is reduced when the amplitude of the feedback connections is increased (see Figures 4.33 and 4.34 for summary). This time dependence of distractor impact agrees with experimental observations (Suzuki & Gottlieb, 2013; Zaksas & Pasternak, 2006).

The dependence of the distractor impact on the distractor location in our model is qualitatively equivalent to that of a single ring network (Compte *et al.*, 2000), which is why we have focused less on its analysis. The overall distractor impact first increases with the difference between cue and distractor location ($\Delta\theta_D$) and then decreases (from $\sim \Delta\theta_D > 120$) (Figure 4.34e,f). As explained above, this effect is related to the attraction and suppression of coexisting bumps. Suzuki & Gottlieb (2013) observe a lower impact on behavior for distractors far away ($> 135°$) from the cue location than for near distractors ($45°$). They also saw a lower impact on prefrontal firing rates for the far condition (Suzuki & Gottlieb (2013), Fig 1,3). Our model can be compatible with their results if we consider sufficiently distant angles ($\Delta\theta$ 150 °), but the comparison is not straightforward. (see Chapter 5, discussion).

Our results are a consequence, on the one hand, of the general dynamics of the interaction

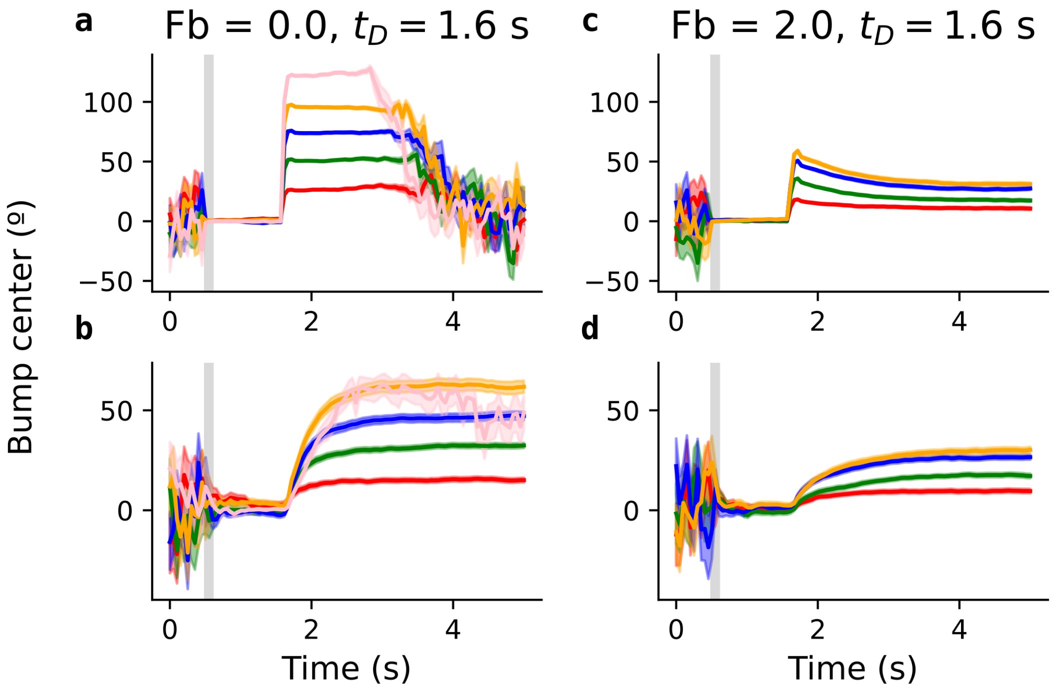

Figure 4.32: Feedback reduces significantly the effect of distractors presented at increasingly distant locations. **a,b** effect of distractors presented at $t = 1.6$ s at different distances from the cue location to the model with no feedback. **a** perfect distractor encoding, **b** displacement of the *storage* bump depending on the distractor's position. **c,d** same examples as in **a,b** for the model with feedback. **c** distractor impact first attenuated by the encoding ring, and **b** further attenuated by the *storage* ring.

between bumps in a ring network and, on the other hand, of the population architecture of our model. The model's vulnerability at the beginning of the delay relies strongly on the information relay from *encoding* to *storage* ring and on the fact that the *storage* neurons' activity is close to baseline at the beginning of the delay.

4.3.5 Effect of distractors in the *ramping network* model

Our circuit proposal with ramping (*ramping network*) exhibits the same qualitative results (Figures 4.35 and 4.36). The differences are related to the time course of the *storage* bump, which increases gradually in the *ramping network* model, producing an analogously gradual decay of the distractor impact (Figure 4.35b,d).

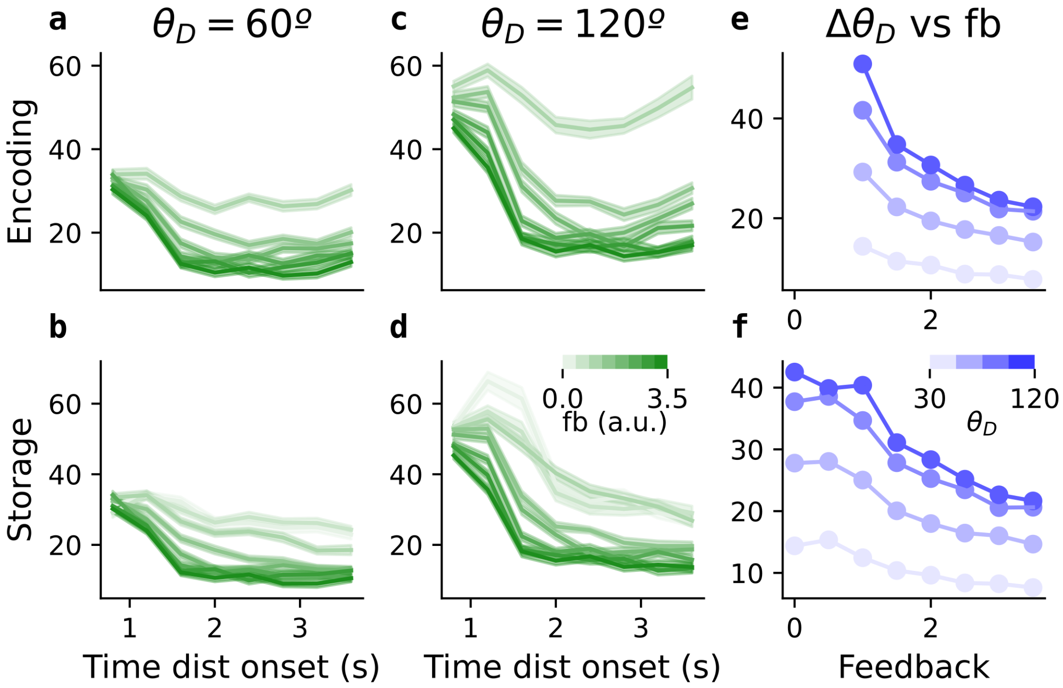

Figure 4.33: Distractor impact is lower for late presentation times and higher feedback values The impact of the distractor is measured as the average bump centroid position during the last 250 ms for 100 trials (cue stimulus at 0°. **a**-**d**: The distractor impact ($\Delta\theta$, $y-$axis) is plotted against the time of distractor presentation for the *encoding* (top) and *storage* ring (bottom) for distractors presented at $\theta = 30$ ° (left) and $\theta = 120$ ° (center). **e**,**f** Average impact over all times of distractor presentation ($y-$ axis) against feedback value. Values corresponding to $Fb = 0, 0.4$ are omitted for the *encoding* plots (**a**,**c**,**e**) since the absence of selective activity prevents a meaningful memory readout.

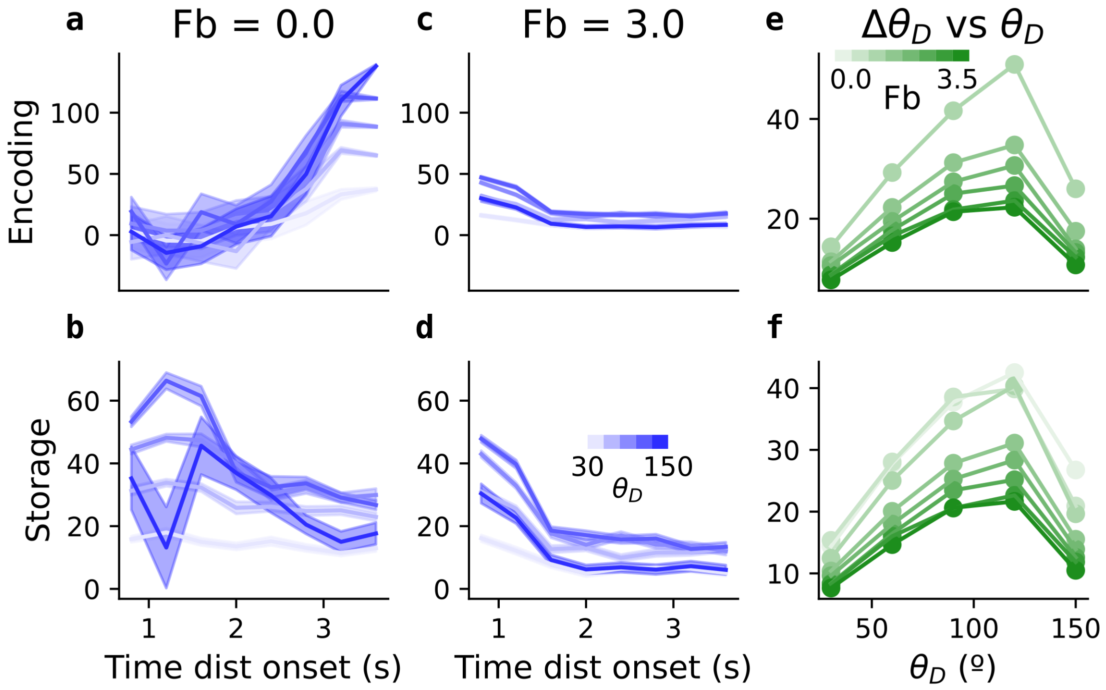

Figure 4.34: Distractor impact increases with the distance between cue and distractor position while the interaction is attractive and drops for farther distractor locations. a-d Distractor impact against time of distractor presentation for different angles of distractor presentation (different scales of blue within each plot) for no-feedback (left) and $Fb = 3.0$ (center). **e,f** average impact over different times of distractor presentation ($y-$axis) against angle of distractor ($x-$axis) for different values of feedback (different scales of green).

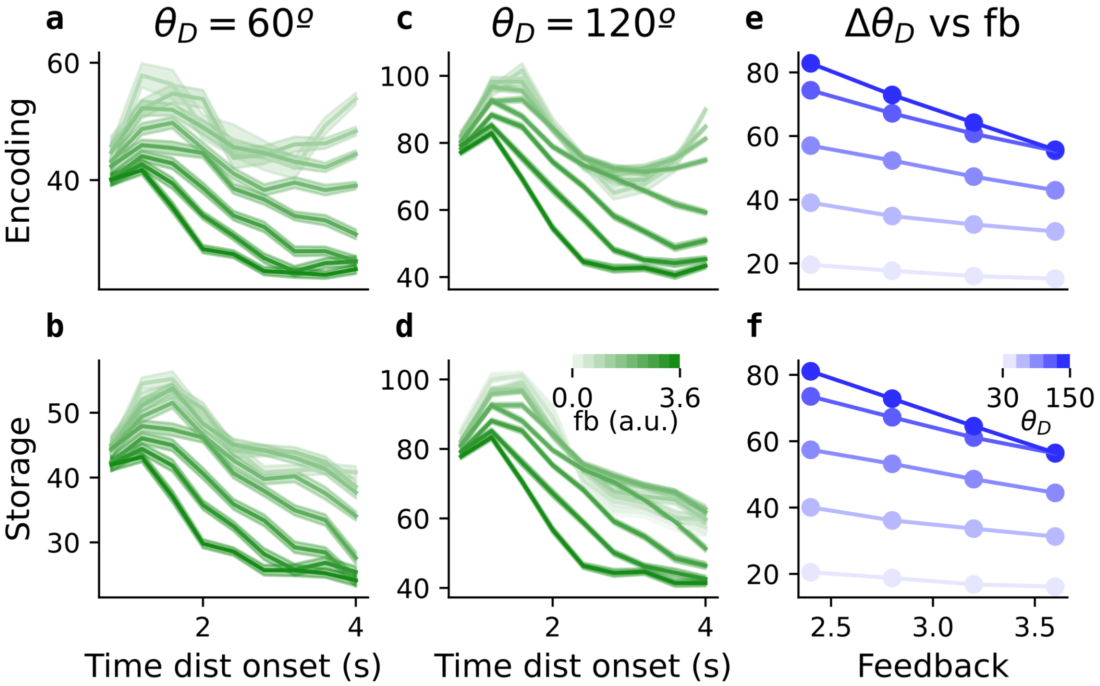

Figure 4.35: Distractor impact is lower for late presentation times and higher feedback values Same a Figure 4.33 for the *ramping network* model (with ramping signal). the

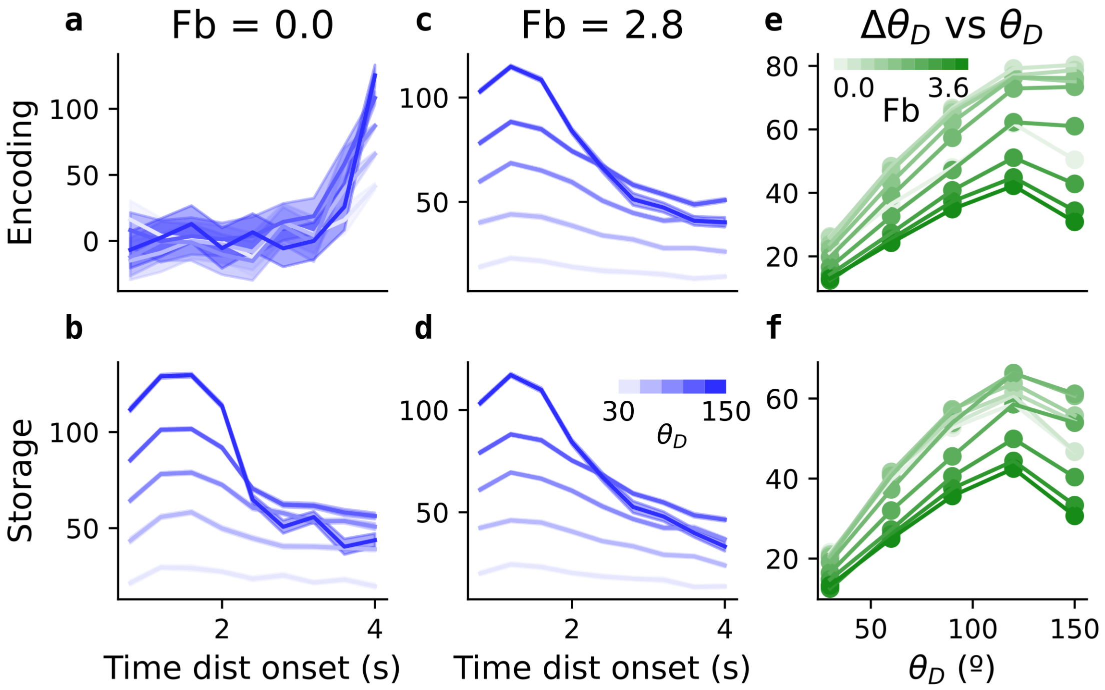

Figure 4.36: Distractor impact increases with the distance between cue and distractor position while the interaction is attractive and drops for farther distractor locations. same as Figure 4.34 for the *ramping network* model (with ramping signal)

Conclusion

We show that the subpopulation structure endows the prefrontal network model with enhanced robustness to distracting stimuli. In particular, our network is more vulnerable at the beginning of the delay period when the memory bump is still forming in the *storage* ring. This time dependence agrees with experimental observations (Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013). Additionally, the network robustness is proportional to the strength of the top-down feedback. This prediction highlights the possible functional advantage of having feedback projections in a prefrontal network with a prominent feedforward structure.

Altogether, these results provide indirect support for the proposed network structure. While the model is built intentionally to reproduce the results in Section 4.2, its behavior in the presence of distracting stimuli is naturally exhibited as a consequence of the network structure.

Chapter 5

Discussion

In this thesis, we investigated the possibility that functional subsets or subpopulations of prefrontal neurons underlie salient features of the working memory code. In particular, we investigated the origins of the well-characterized dynamic-to-stable transition in the population code (Cavanagh *et al.*, 2018; Parthasarathy *et al.*, 2019; Spaak *et al.*, 2017; Stokes *et al.*, 2013) and the impact of distracting stimuli on the prefrontal cortex (Suzuki & Gottlieb, 2013). We first used different tools to quantify the degree of structure (clusters) in a set of macaque prefrontal recordings and found a non-random distribution of feature selectivity among the prefrontal neurons. More specifically, the observed stable stimulus selectivity and stereotyped firing rate profiles (of distinct neuronal groups (Markowitz *et al.*, 2015)) are incompatible with overall random structure and non-linear mixed selectivity. Next, we proposed a computational model based on the bump attractor (Compte *et al.*, 2000) that illustrates how the distinct activity profiles can be related to the dynamic-to-stable transition of the population code. The model suggests that the delay code is orthogonal to the code during the cue presentation because two distinct sets of neurons (labeled as *encoding* and *storage* neurons in the model) are active during these epochs, respectively. We showed that different network realizations are compatible with the qualitative dynamics of the population code and investigated the possible dependency of prefrontal maintenance on subcortical inputs (Finkelstein *et al.*, 2021; Johnstone & Rolls, 1990; Sreenivasan & D'Esposito, 2019). Finally, we illustrated how the subpopulation structure of the model can increase the resistance to distracting stimuli. We found that the interaction between the *encoding* and the *storage* neurons accounts for higher vulnerability to distraction at the beginning of the delay (as observed in Suzuki & Gottlieb (2013)). Moreover, vulnerability to distractors is overall reduced by increasing the amplitude of the top-down feedback projections.

Structure and functional specialization in the prefrontal cortex

A departing point and motivation for our work was to investigate the origins of the dynamic transition between the cue and delay epochs in delayed-response working memory paradigms (Mendoza-Halliday & Martinez-Trujillo, 2017; Parthasarathy *et al.*, 2019; Spaak *et al.*, 2017; Stokes *et al.*, 2013). Since this transition is consistently observed in delay-response tasks, it likely reflects essential features of the working memory dynamics.

The discovery of dynamics (Stokes *et al.*, 2013) contributed to the questioning of persistent activity as the principal mechanism (phenomenon) behind short-term memory (Lundqvist *et al.*, 2018; Meyers, 2018; Stokes, 2015; Stokes *et al.*, 2013). This emphasis on dynamics motivated the proposal of mechanisms such as chaotic regimes (Barak *et al.*, 2013), oscillations (Bastos *et al.*, 2018; Lundqvist *et al.*, 2018; Miller *et al.*, 2018), and sequential activations (Ganguli & Latham, 2009; Goldman, 2009) as alternatives to attractor dynamics. Some of these mechanisms ensured that dynamics are compatible with a stable readout (Druckmann & Chklovskii, 2012; Goldman, 2009). Only recently, some models have tried to capture both dynamics and stability in the code (Bouchacourt & Buschman, 2019; Murray *et al.*, 2017a; Stroud *et al.*, 2023).

Our approach also focuses both on the dynamics (transition from cue to delay) and stability (late delay). As mentioned in the introduction, this dynamic transition implies a relay or transformation of the stimulus-related information from cue to delay. What mechanisms underlie this transformation?

We discussed two contrasting scenarios compatible with the code transition: non-linear mixed selectivity and functional populations. These scenarios can be seen as the extremes of a continuum, which implies that the real neurons should lie somewhere in between. We emphasize that these scenarios are not considered entirely mutually exclusive. On the opposite, non-linear mixed selectivity has been observed in the prefrontal cortex (Asaad *et al.*, 2000; Dang *et al.*, 2021; Fusi *et al.*, 2016; Rigotti *et al.*, 2013) as well as in some neurons in the data set we analyzed (Figure 4.4). The question we try to answer (in our analysis) is whether the degree of structure in the distribution of feature selectivity among neurons can be informative about network mechanisms. A negative answer to this question would imply that the observed dynamics in the data should be understood as the dynamics of a *single complex network* that exhibits mixed selectivity. On the other hand, a positive answer suggests that part of the dynamics are due to the differential behavior of distinct functional groups or *functional subpopulations*. We argue that the view of the prefrontal circuits related to working memory as a single network with complex dynamics is implicit in many modeling works (Compte *et al.*, 2000; Goldman, 2009; Murray *et al.*, 2017a; Wimmer *et al.*, 2014). Moreover, this view is especially suited for models that use artificial intelligence for the training of RNNs, which usually do not include structure in their weights previous to the training (Barak *et al.*, 2013; Stroud *et al.*, 2023) Explicitly modeling the prefrontal cortex as separate subnetworks has been a less popular approach (see the model proposed by Bouchacourt & Buschman (2019) for an example).

However, our analysis shows that the degree of functional clustering in the data significantly differs from what would be expected by chance, suggesting that specialized subpopulations underlie the transition. Similar results had been reported in rodent frontal cortex (Hirokawa *et al.*, 2019; Yang *et al.*, 2022). However, to our knowledge, there have not been comparable studies with primates.

We narrowed the focus of the population analysis, considering fewer Principal Components and classifying neurons according to some of the dPC weights, to bring it closer (to relate) to the single neuron results (Markowitz *et al.*, 2015; Mendoza-Halliday & Martinez-Trujillo, 2017). This approach placed the distinct functional activity profiles obtained by single neuron analysis (Markowitz *et al.*, 2015) in the context of the heterogeneity of the prefrontal cortex responses.

Our findings underscore the relevance of the two contrasting types of activity profiles (neurons active during cue, labeled as *perceptual*, and neurons active during delay, labeled as *mnemonic*). As a limitation of the segregated picture offered by the single neuron classification (Markowitz *et al.*, 2015), our results suggest that the observed heterogeneity is compatible with a continuum of activity profiles. However, *perceptual* and *mnemonic* neurons are more present than would be expected by chance. These activity profiles have been consistently observed in other studies (Finkelstein *et al.*, 2021; Inagaki *et al.*, 2019; Mendoza-Halliday & Martinez-Trujillo, 2017; Yang *et al.*, 2022), and we directly relate them to the dynamic transition in the code. In this regard, the cross-temporal decoding serves as a proof of principles, showing how the respective activation profiles of these neurons can be related to changes in decoding accuracy between the two epochs.

The analysis done in Markowitz *et al.* (2015) shows a higher density of *perceptual* neurons in the deeper layer and a higher density of *mnemonic* neurons in the superficial layers. These results align with other studies with macaques (Bastos *et al.*, 2018), rodents (Wu *et al.*, 2020) and humans (Finn *et al.*, 2019). From a computational point of view, having neurons that receive the stimulus from upstream areas and different neurons that maintain the stimulus when required (working memory) is advantageous because it offers a layer of protection to the maintained memory. We elaborate on this point in Section 4.3. Moreover, this functional distribution can also be regarded as a consequence of the evolution of the neural circuitry. Maybe biased by abstract thinking and mathematical formulations, we sometimes want to understand the biological circuits as if they had emerged from an optimization problem. This mindset finds its perfect playground in the use of artificial intelligence to understand brain function (Zador, 2019). However, the prefrontal cortex did not grow (through evolution and lifespan) only to solve the ocular-delayed response task. Considering that the representation of visual stimuli in the prefrontal cortex can have a separate history from that of the maintenance of these stimuli for small time spans (Fuster, 2015; Hussar & Pasternak, 2009; Martin-Cortecero & Nuñez, 2016; Zaksas & Pasternak, 2006) sheds light on the observed dynamics. This simple consideration makes the functional and anatomical division natural.

A three-population model explains salient features of the prefrontal activity during working memory

We explicitly modeled the different stereotypical responses to task condition (Markowitz *et al.*, 2015). The key ingredient of our model proposals are: the separate representation of different subpopulations and their defined working regimes. The model realizes what we summarize in the data analysis. The interaction between the stimulus responsive (labeled as *encoding*, corresponding to the *perceptual* neurons in the data) and delay responsive (labeled as *storage*, corresponding to the *mnemonic* in the data) ring constitute a fundamental aspect of the model's dynamic. Our network has an explicit feedforward structure (*encoding* neurons receive the stimulus, then they activate the *storage* neurons), which is only softened by the feedback projections (*storage* to *encoding*). Our *readout* network has its analogous counterpart in other models of dynamics in prefrontal working memory (Druckmann & Chklovskii, 2012; Goldman, 2009). It also serves the purpose of modeling explicitly neurons with persistent firing rate profiles. Since the dynamics in our model are highly structured (they consist of the sequential activation of the *encoding* and *storage* ring), including a network with constant readout (the *readout* ring in our model) is straightforward. Reading out from *encoding* and *storage* neurons (excitatory projections from these populations to the *readout* ring) ensures a constant stimulus representation.

Our network proposal can be contrasted with the models in Murray *et al.* (2017a), Stroud *et al.* (2023) and Bouchacourt & Buschman (2019). All these works explain the dynamic to stable transition of the code. In Murray *et al.* (2017a) and Stroud *et al.* (2023), the transition to a stable code happens due to the non-normal network dynamics. As highlighted above, non-normality is exhibited by almost any network with non-symmetric connectivity (Horn & Johnson, 1985) [1]. Our network proposal converges with the ones in Murray *et al.* (2017a); Stroud *et al.* (2023) in that they both require directionality in the information flow (feedforward structure). We argue that any network that can explain the dynamic to stable transition necessarily relies on feedforward mechanisms. In this regard, claiming that non-normal networks explain the dynamic transition is more a tautology than a new finding. However, Stroud *et al.* (2023) do contribute to a discussion of the mechanisms by providing a measure of energetic efficiency (see supplementary material in Stroud *et al.* (2023)). In particular, they prove that the information loading with non-normal dynamics is more optimal than the information loading in standard attractor models, which have normal connectivity (Compte *et al.*, 2000; Wimmer *et al.*, 2015). Their insight into information loading efficiency provides a perspective that can be complemented by our specific network implementation. To sum up, on the one hand, we could say that prefrontal networks exploit the advantages of feedforward motifs for memory loading (the discussion in Stroud *et al.* (2023)). On the other hand, we keep looking for concrete network realizations that would implement feedforward elements (our work deals with this question) that are compatible with the experimental observations.

A specific shortcoming of the *bistable network* network is the degree to which it depends on fine-tuning to produce an activity compatible with the ramping of the *mnemonic* neurons in the Markowitz *et al.* (2015) data. Too much feedforward excitation leads to fast bump formations that span a small fraction, not the entire delay. We emphasized the plausibility of this architecture because of its conceptual relevance (the bistability of the *storage* endows this prefrontal network with autonomous memory sustain function). Additionally, the main qualitative hallmark of the dynamic transition (orthogonality between cue and delay) is captured regardless of the time course of the *storage* neurons.

The variation of our three-ring network architecture with an external ramping signal introduces a conceptual modification without changing the central aspects of the model: the different function profiles and their relation to the population dynamics. Ramping activity has been observed repeatedly in paradigms with predictable delay length (Finkelstein *et al.*, 2021; Hirokawa *et al.*, 2019; Inagaki *et al.*, 2019; Mendoza-Halliday & Martinez-Trujillo, 2017). Comparing experimental data to network simulations, Finkelstein *et al.* (2021) found evidence for the ramping being provided by an area external to the frontal cortex rather than intrinsically generated, which inspired our approach.

The external input led us to modify the dynamic regime of the *storage* network from bistable to monstable. This change ensures that the modulation significantly increases the bump's amplitude without producing a loss in stability. The conceptual implication of this model modification (monostable memory circuit) justified the account of the two network proposals

[1] Skewed-symmetric matrices ($A = -A^{\star}$) are a special case of non-symmetric matrices that are normal. A minimal example network would be two units where the first excites the second while the second inhibits the first with the same intensity. Such a connectivity can describe a balanced regime between excitation and inhibition. Still, it cannot include any element of directionality that is needed to explain the working memory data

separately. In the context of attractor networks, memory is sustained through bistability at the network level (Compte *et al.*, 2000; Wimmer *et al.*, 2014). Bistability is the basic element of any system capable of autonomous memory (Zylberberg & Strowbridge, 2017). The absence of bistable circuits directly implies that maintenance relies on the external input source. Recent models of working memory, closely inspired by the experimental data, have proposed that maintenance can be achieved by the interaction of monostable cortical networks (Feng *et al.*, 2023; Mejías & Wang, 2022). Our network with ramping signal aligns with this view, which questions the role of the prefrontal cortex as an autonomous hub and focuses instead on the distributed nature of working memory (Christophel *et al.*, 2017). Regarding the dynamics at the population level, the network with ramping produces a ramping pattern in the cross-temporal decoding, which is a feature encountered in the data (Parthasarathy *et al.*, 2019; Spaak *et al.*, 2017; Stokes *et al.*, 2013). However, when short delay lengths are considered, the network with bistable dynamics produces a similar increasing pattern.

A common aspect of both network formulations that deserves a separate comment is the implementation of the suppression mechanism. The non-activation of the *mnemonic* cells (Markowitz *et al.*, 2015; Mendoza-Halliday & Martinez-Trujillo, 2017) is an important characteristic of the dynamics. Note that this behavior is manifested throughout the visual condition, where there is no memory delay, but also during the time of the stimulus presentation in the memory task. This memory-dependent activation, together with the stimulus-dependent activation of the *perceptual* neurons, underlies the dynamics observed in the cross-temporal decoding. How to model this memory selective activation was not obvious. We did not find any significant modulation in the activity of the inhibitory neurons during stimulus presentation, which would have given a direct cue of a suppressing mechanism. We eventually resolved to a phenomenological implementation where the stimulus directly projects to the inhibitory neurons. We additionally explored a mechanism that implied excitatory to inhibitory AMPA-mediated projections from the *encoding* to the *storage* ring. With this implementation, we achieved qualitatively equivalent results.

Regarding how we implemented the stimulus-dependent suppression, it can be argued that we did not base our modeling choices on direct experimental evidence. This is true as far as the exact details are concerned. However, we emphasize again that our focus was more on illustrating how structure (functional subpopulations) can underlie the observed dynamics than on the specificities of connectivity.

In future analyses, we plan to train an RNN on a dual task paradigm as the one used in Markowitz *et al.* (2015); Mendoza-Halliday & Martinez-Trujillo (2017). Contrasting our network solution with the one proposed by a trained network should shed light on our discussion on the implementation of feedforward structure, functional specialization and mixed-selectivity.

Distinct populations for encoding and maintenance during working memory improve the resistance against distracting stimuli

Our model naturally inherits many of the properties that bump attractor models exhibit in the presence of distracting stimuli. These properties have been discussed in previous works (Almeida *et al.*, 2015; Compte *et al.*, 2000; Edin *et al.*, 2009). The bump attractor faithfully captures the experimentally observed dependence of the distractor impact on the distraction-target similarity (Nemes *et al.*, 2012; Rademaker *et al.*, 2015). Within a certain range specified by the bump's width, the impact increases as a function of the difference between the target and the distractor. Out of the attractive region (no overlap between the bump-mediated positive modulations of the target and the distractor), the distractor has no effect on the target bump. The model's behavior for distractors presented far from the target changes for high amplitude distracting stimulus, which can produce a complete distraction (Compte *et al.*, 2000). The extreme cases of complete distraction and complete filtering (both observed experimentally (Nemes *et al.*, 2012; Rademaker *et al.*, 2015)) are explained by the effects of the surrounding suppression of the distractor and the target bumps, respectively.

The study of the impact of distractors at different positions is motivated by investigating the working memory capacity of a network (Edin *et al.*, 2009). Working memory capacity and its underlying mechanisms have been on the focus of experimental (Bays *et al.*, 2009; Bouchacourt & Buschman, 2019; Vogel & Machizawa, 2004; Vogel *et al.*, 2005; Watanabe & Funahashi, 2014) and modeling (Bouchacourt & Buschman, 2019; Edin *et al.*, 2009; Oberauer *et al.*, 2012) works in the last two decades. In contrast, the importance of distractor presentation time has been overlooked in both experiments and models, with a few exceptions (Murray *et al.*, 2017b; Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013). However, as our work highlights, understanding how distractors impact memory at different task epochs can give cues about working memory function and underlying circuit structure. In particular, the time course of the distractor impact in our model is directly related to the presence of distinct functional subpopulations.

The bistable bump attractor, usually the circuit proposed to be responsible for memory encoding and storage (Compte *et al.*, 2000; Wimmer *et al.*, 2014), loads the stimulus fast into the stable memory representation. This fast loading implies that the system has stable activity shortly after the presentation of the stimulus (time of bump formation ~ 50 ms). Consequently, distractors presented at different times will have the same impact. A distractor can have a smaller or higher impact when presented before the cue is withdrawn. This effect will depend on whether the bump's amplitude is below (due to weak stimulation) or above (due to strong stimulation) its stable value during the delay. This efffect, which we do not show, is confined to distractors presented during the cue period. Moreover, its dependence on the stimulus intensity makes it vulnerable to parameter changes and more difficult to interpret.

In our model, on the contrary, the dependence on the time of distraction presentation is an intrinsic feature that does not depend on stimulus intensity and is not restricted to a specific parameter combination. Our network becomes increasingly resistant to distractors as the activity in the *storage* ring increases. The behavior is exhibited by the both proposed network architectures (*bistable network*, *ramping network*). In both cases, the *storage* bump starts to form after the withdrawal of the cue, always leaving an initial time of increased vulnerability. Our three-ring networks also show enhanced robustness when the top-down feedback is increased. Interestingly, this effect also does not depend on whether our frontal network sustains memory autonomously (*bistable network*) or with the help of external input (*ramping network*). This result contributes to the distributed picture of working memory (Christophel *et al.*, 2017; Lara & Wallis, 2015). It suggests that even if there are areas that can maintain memory intrinsically, the interaction between circuits within these areas can be transcendental for behavior.

From a quantitative perspective, the model predicts overall greater memory biases than observed in the behavioral experiments (Nemes *et al.*, 2012; Rademaker *et al.*, 2015; Suzuki & Gottlieb, 2013). Values closest to the behavioral biases ($\sim 10°$) are obtained with high values of the top-down feedback. Whereas this result can be seen as a shortcoming of the model, two important aspects should be taken into account. The first one is that we do not model any filtering of the distractor that happens previous to its arrival to the prefrontal cortex. Distractors filtering by regions upstream from the prefrontal cortex have been reported in several studies (Lorenc *et al.*, 2018; Murray *et al.*, 2017a; Suzuki & Gottlieb, 2013; Yoon *et al.*, 2006). The second aspect to consider is that we also do not model the motor output, making the quantitative relation between the memory and response bias arbitrary. These points, together, emphasize that our analysis aims at capturing qualitative aspects of the data and not at fitting them quantitatively.

Our model has parallelisms with the work of (Murray *et al.*, 2017b). These authors model the PFC and posterior parietal cortex (PPC) as separate circuits with reciprocal projections, with the stimulus projecting only to PPC. Among other properties, their model also exhibits increased robustness for late distractor presentations, and its vulnerability is reduced when PFC to PPC feedback is included. However, there are conceptual differences between our works. In the first place, they used two-population modules responsive (capable of representing binary stimuli) to represent each area (PFC, PPC), while we used a ring attractor network (suited to represent continuous stimuli) for each population (*encoding*, *storage*, *readout*). Although this distinction is orthogonal to the emphasis placed on distinct functional subpopulations, discrete and continuous networks make different testable predictions. The continuous bump attractor has successfully captured aspects that discrete models cannot explain, such as the increase in the memory error as a function of the delay length, the correlation between rate and behavior variability, and correlations between single neuron firing rates (Wimmer *et al.*, 2014). Moreover, the discrete two-population modules only allow to model distractors that either overlap completely or are orthogonal to the target. Accordingly, these networks can only capture the extreme cases of complete bias or perfect filtering. Finally, it can be argued that a continuous representation is more realistic (natural stimuli usually vary in a continuous fashion), and should therefore be prioritized. A possible extension of our work could involve building a network equivalent to our three-ring model, composed of two-population modules instead of ring attractors. The interest would be in highlighting the predictions that would remain unchanged. We can anticipate that the dynamic-to-stable code transition could be equally explained, as well as the overall time dependence of distraction impact and the role of the top-down feedback.

A second difference between our works is that Murray *et al.* (2017b) do not consider a network realization without bistable circuits, unlike us. The emphasis we make on the proposed architecture with external ramping input and its conceptual difference from the network with bistability was commented on above.

Finally, the circuit modules in Murray *et al.* (2017b) represent different areas (PFC,PPC), while ours represent different functional populations within the same area (PFC). Both works highlight the potential relevance of having different networks supporting working memory. By considering different populations within the same area (in the model and data (Markowitz *et al.*, 2015)), we translate the distributed view of working memory from the interaction between areas (as illustrated in Edin *et al.* (2009); Murray *et al.* (2017b)) to the interaction between regions within the same area. In this regard, we emphasize that despite the evidence for layer-specific activity found in the literature (Bastos *et al.*, 2018; Finn *et al.*, 2019; Wu *et al.*, 2020), our analysis does not articulate the concrete relation between functional specialization and anatomical distinction.

Based on the previous discussion, we emphasize our intention of proposing a simple way of rethinking prefrontal circuitry based on experimental information about functional structure. From a conceptual point of view, thinking of memory as distributed across areas or subareas does not make a big difference. However, there is still a lot of focus on finding network models whose emergent dynamics can explain the most relevant aspects of working memory (Murray *et al.*, 2017a; Stroud *et al.*, 2023). These efforts often lead to network solutions whose dynamics are hard to relate to specific neural mechanisms (Barak *et al.*, 2013; Murray *et al.*, 2017a; Stroud *et al.*, 2023). We emphasize again the possible contribution of artificial intelligence to the view of the brain as a machine that has evolved to solve a certain task. This formulation, which we believe is more detrimental than beneficial, is improved when several tasks are considered instead of a single one (Masse *et al.*, 2019; Yang *et al.*, 2019). However, we believe that explicitly including *different areas* (Feng *et al.*, 2023; Mejías & Wang, 2022; Yang & Molano-Mazón, 2021) is also essential to get a better understanding of the brain circuitry. Exhibiting dynamic and stable code, linear-increasing delay activity, and being robust against distractors may be too many requirements to be satisfied by a single neuronal network. However, these requirements might be easily met by integrating the dynamics of different networks whose respective dynamics are simple a nd interpretable.

Chapter 6

Conclusions

4.1 Structural and functional specialization in the prefrontal cortex

1 We analyzed the activity of regular spiking (putative excitatory) cells in the macaque prefrontal cortex, finding that the heterogeneity observed at the level of single neurons is compatible with structured featured selectivity. This finding contradicts the view that prefrontal computations rely primarily on non-linear mixed selectivity.

2 The structured feature selectivity can be related to distinct functional neuron groups. Among the stimulus-selective neurons, three types of profiles are modulated differently according to the memory or visual task condition. These neurons overlap with the groups highlighted in Markowitz *et al.* (2015).

3 The two contrasting neuronal profiles that are most modulated by the task condition (*perceptual* and *mnemonic* neurons) are active at different task epochs (cue presentation and delay). Cross-temporal decoding from these groups separately shows their close to orthogonal contribution to the population code. This finding indicates that the presence of these contrasting response profiles underlies the experimentally observed dynamic to stable transition in the working memory code.

4.2 A three-population model explains salient features of the prefrontal recordings

1 A model composed of three interconnected ring attractors captures the diverse functional profiles and the dynamic to stable transition in the prefrontal code during working memory.

2 The model highlights the relevance of the contrasting types of activation profiles among the neurons, labeled as *encoding* and *storage* in the model, that correspond to the groups labeled as *perceptual* and *mnemonic* in the data. In particular, it suggests that these functional profiles relate to stimulus encoding and maintenance. These contrasting (encoding/maintenance) task epoch selective responses have been observed in other studies with different species (Bastos *et al.*, 2018; Finn *et al.*, 2019; Van Kerkoerle *et al.*, 2017; Wu *et al.*, 2020), which suggests this structured architecture can be a common feature of prefrontal working memory networks.

3 We present two network variations with the same subnetwork structure that make equivalent predictions regarding profile heterogeneity and population code. The network labeled as *bistable network*, which has intrinsic bistability, illustrates how the relevance of subpopulations can be compatible with the traditional view of the prefrontal cortex as a memory-sustaining hub. In contrast, the network labeled as *ramping network* depends on an external input to sustain memory. This external input is not stimulus-selective and can be interpreted as an urgency or ramping signal that tracks elapsed time or response anticipation. This network illustrates how prefrontal dynamics could interact with other areas (likely, subcortical areas such as basal ganglia or locus coeruleus) to sustain memory. Additionally, the *ramping network* network illustrates a way in which the delay activity can naturally adapt to span different delay lengths. This feature shows the capacity to flexibly adapt to different environmental demands can be enhanced through

the interaction between areas or distribution of the code.

4 The two networks (*bistable network* and *ramping network*) offer equivalent explanations for the dynamic to stable transition in the population code: *encoding* (representing the *perceptual* neurons in the data) neurons are active during cue presentation, and *storage* neurons (*mnemonic* neurons in the data) are increasingly engaged during the delay.

4.3 Distinct populations for encoding and maintenance improve the resistance against distracting stimuli

1 Having the memory encoding and maintenance distributed across the *encoding* and *storage* rings makes the system more robust to distracting stimuli. This enhanced robustness is due to the *encoding* ring, which receives the excitation from the stimulus and acts as a first filtering stage.

2 The model is most vulnerable to distractors short after cue withdrawal when the *storage* bump has not formed yet. As the *storage* bump amplitude increases, the model becomes more robust (the impact of the distractor becomes smaller). This vulnerability to early distractor presentations agrees with behavioral (Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013) and neurophysiological Suzuki & Gottlieb (2013) observations. Such behavior cannot be obtained with a model of stable persistent activity.

3 The model's vulnerability to distracting stimuli is reduced by the top-down feedback (*storage* to *encoding*). Increasing feedback values sustain higher amplitude *encoding* bumps during the delay, which is less biased by the distracting input.

4 The enhanced robustness to distracting stimuli of the proposed subpopulation-based model indirectly supports the plausibility and purpose of such a network structure.

Bibliography

Almeida, Rita, Barbosa, João, & Compte, Albert. 2015. Neural circuit basis of visuo-spatial working memory precision: A computational and behavioral study. *Journal of Neurophysiology*, **114**(3), 1806–1818.

Asaad, Wael F., Rainer, Gregor, & Miller, Earl K. 2000. Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, **84**(1), 451–459.

Ashby, F. Gregory, Ell, Shawn W., Valentin, Vivian V., & Casale, Michael B. 2005. FROST: A distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, **17**(11), 1728–1743.

Bae, Gi-Yeul, & Luck, Steven J. 2019. What happens to an individual visual working memory representation when it is interrupted? *British Journal of Psychology*, **110**(2), 268–287.

Barak, Omri, Sussillo, David, Romo, Ranulfo, Tsodyks, Misha, & Abbott, L. F. 2013. From fixed points to chaos: Three models of delayed discrimination. *Progress in Neurobiology*, **103**, 214–222.

Barbosa, Joao, Stein, Heike, Martinez, Rebecca L., Galan-Gadea, Adrià, Li, Sihai, Dalmau, Josep, Adam, Kirsten C.S., Valls-Solé, Josep, Constantinidis, Christos, & Compte, Albert. 2020. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience*, **23**(8), 1016–1024.

Barbosa, Joao, Lozano-Soldevilla, Diego, & Compte, Albert. 2021. Pinging the brain with visual impulses reveals electrically active, not activity-silent, working memories. *PLoS Biology*, **19**(10), e3001436.

Bartholomew, D. J. 2010. Principal components analysis. *International Encyclopedia of Education*, 374–377.

Bastos, André M., Loonis, Roman, Kornblith, Simon, Lundqvist, Mikael, & Miller, Earl K. 2018. Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(5), 1117–1122.

Batuev, A S, Pirogov, A A, Orlov, A A, & Sheafer, V I. 1980. Cortical mechanisms of goal-directed motor acts in the rhesus monkey. *Acta neurobiologiae experimentalis*, **40**(1), 27–49.

Bays, Paul M., Catalao, Raquel F.G., & Husain, Masud. 2009. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, **9**(10), 1–11.

Berridge, C W, & Abercrombie, E D. 1999. Relationship between locus coeruleus discharge rates and rates of norepinephrine release within neocortex as assessed by in vivo microdialysis. *Neuroscience*, **93**(4), 1263–1270.

Berridge, Craig W, & Waterhouse, Barry D. 2003. The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain research. Brain research reviews*, **42**(1), 33–84.

Bouchacourt, Flora, & Buschman, Timothy J. 2019. A Flexible Model of Working Memory. *Neuron*, **103**(1), 147–160.e8.

Brunel, Nicolas, & Van Rossum, Mark C.W. 2007. Lapicque's 1907 paper: From frogs to integrate-and-fire. *Biological Cybernetics*, **97**(5-6), 337–339.

Burkitt, A. N. 2006. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biological Cybernetics*, **95**(1), 1–19.

Camperi, Marcelo, & Wang, Xiao Jing. 1998. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience*, **5**(4), 383–405.

Carland, Matthew A., Thura, David, & Cisek, Paul. 2019. The Urge to Decide and Act: Implications for Brain Function and Dysfunction. *Neuroscientist*, **25**(5), 491–511.

Cavallari, Stefano, Panzeri, Stefano, & Mazzoni, Alberto. 2014. Comparison of the dynamics of neural interactions between current-based and conductance-based integrate-and-fire recurrent networks. *Frontiers in Neural Circuits*, **8**(MAR), 1–23.

Cavanagh, Sean E., Towers, John P., Wallis, Joni D., Hunt, Laurence T., & Kennerley, Steven W. 2018. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature Communications*, **9**(1), 1–16.

Chai, Wen Jia, Abd Hamid, Aini Ismafairus, & Abdullah, Jafri Malin. 2018. Working memory from the psychological and neurosciences perspectives: A review. *Frontiers in Psychology*, **9**(MAR), 1–16.

Chang, Chih-Chung, & Lin, Chih-Jen. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, Wenqing, Li, Chen, Liang, Wanmin, Li, Yunqi, Zou, Zhuoheng, Xie, Yunxuan, Liao, Yangzeng, Yu, Lin, Lin, Qianyi, Huang, Meiying, Li, Zesong, & Zhu, Xiao. 2022. The Roles of Optogenetics and Technology in Neurobiology: A Review. *Frontiers in Aging Neuroscience*, **14**(April), 1–12.

Chiba, Atsushi, Oshio, Ken Ichi, & Inase, Masahiko. 2015. Neuronal representation of duration discrimination in the monkey striatum. *Physiological Reports*, **3**(2), 1–17.

Christophel, Thomas B, Klink, P Christiaan, Spitzer, Bernhard, Roelfsema, Pieter R, & Haynes, John-Dylan. 2017. The Distributed Nature of Working Memory. *Trends in cognitive sciences*, **21**(2), 111–124.

Chung, Sue Yeon, & Abbott, L. F. 2021. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, **70**(November), 137–144.

Cisek, Paul. 2019. Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, and Psychophysics*, **81**(7), 2265–2287.

Compte, A., Brunel, N, Goldman-Rakic, P S, & Wang, X J. 2000. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex (New York, N.Y. : 1991)*, **10**(9), 910–23.

Constantinidis, C, Franowicz, M N, & Goldman-Rakic, P S. 2001a. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **21**(10), 3646–55.

Constantinidis, C., Franowicz, M. N., & Goldman-Rakic, P. S. 2001b. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience*, **4**(3), 311–316.

Constantinidis, Christos, & Goldman-Rakic, Patricia S. 2002. Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *Journal of Neurophysiology*, **88**(6), 3487–3497.

Constantinidis, Christos, & Wang, Xiao-Jing. 2004. A neural circuit basis for spatial working memory. *The Neuroscientist*, **10**(6), 553–565.

Crammond, Donald J., & Kalaska, John F. 1996. Differential relation of discharge in primary motor cortex and premotor cortex to movements versus actively maintained postures during a reaching task. *Experimental Brain Research*, **108**(1), 45–61.

Cueva, Christopher J, Marcos, Encarni, Saez, Alex, Genovesio, Aldo, Jazayeri, Mehrdad, Romo, Ranulfo, Salzman, C Daniel, Shadlen, Michael N, & Fusi, Stefano. 2018. Low dimensional dynamics for working memory and time encoding. *bioRxiv*, 504936.

Cueva, Christopher J., Saez, Alex, Marcos, Encarni, Genovesio, Aldo, Jazayeri, Mehrdad, Romo, Ranulfo, Salzman, C. Daniel, Shadlen, Michael N., & Fusi, Stefano. 2020. Low-dimensional dynamics for working memory and time encoding. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(37), 23021–23032.

Dang, Wenhao, Jaffe, Russell J., Qi, Xue Lian, & Constantinidis, Christos. 2021. Emergence of non-linear mixed selectivity in prefrontal cortex after training. *Journal of Neuroscience*, **41**(35), 7420–7434.

Darshan, Ran, & Rivkind, Alexander. 2022. Learning to represent continuous variables in heterogeneous neural networks. *Cell Reports*, **39**(1), 110612.

Dehaene, Stanislas, Meyniel, Florent, Wacongne, Catherine, Wang, Liping, & Pallier, Christophe. 2015. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, **88**(1), 2–19.

Druckmann, Shaul, & Chklovskii, Dmitri B. 2012. Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology*, **22**(22), 2095–2103.

Dubreuil, Alexis, Valente, Adrian, Beiran, Manuel, Mastrogiuseppe, Francesca, & Ostojic, Srdjan. 2022. The role of population structure in computations through neural dynamics. *Nature Neuroscience*, **25**(6), 783–794.

Durstewitz, Daniel. 2003. Self-organizing neural integrator predicts interval times through climbing activity. *Journal of Neuroscience*, **23**(12), 5342–5353.

Eckhoff, Philip, Wong-Lin, K. F., & Holmes, Philip. 2009. Optimality and robustness of a biophysical decision-making model under norepinephrine modulation. *Journal of Neuroscience*, **29**(13), 4301–4311.

Edin, Fredrik, Klingberg, Torkel, Johansson, Pär, McNab, Fiona, Tegnér, Jesper, & Compte, Albert. 2009. Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(16), 6802–6807.

Emmons, Eric B., De Corte, Benjamin J., Kim, Youngcho, Parker, Krystal L., Matell, Matthew S., & Narayanan, Nandakumar S. 2017. Rodent medial frontal control of temporal processing in the dorsomedial striatum. *Journal of Neuroscience*, **37**(36), 8718–8733.

Engel, Tatiana A., Chaisangmongkon, Warasinee, Freedman, David J., & Wang, Xiao Jing. 2015. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications*, **6**, 1–12.

Esnaola-Acebes, Jose M., Roxin, Alex, & Wimmer, Klaus. 2022. Flexible integration of continuous sensory evidence in perceptual estimation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, **119**(45), 1–11.

et. all. Hastie, Trevor. 2009. Statistics The Elements of Statistical Learning. *Springer Series in Statistics*, **27**(2), 745.

Feng, Mengli, Bandyopadhyay, Abhirup, & Mejias, Jorge F. 2023. Emergence of distributed working memory in a human brain network model. *bioRxiv*, 2023.01.26.525779.

Finkelstein, Arseny, Fontolan, Lorenzo, Economo, Michael N, Li, Nuo, Romani, Sandro, & Svoboda, Karel. 2021. Attractor dynamics gate cortical information flow during decision-making. *Nature Neuroscience*, **24**(6), 843–850.

Finn, Emily S., Huber, Laurentius, Jangraw, David C., Molfese, Peter J., & Bandettini, Peter A. 2019. Layer-dependent activity in human prefrontal cortex during working memory. *Nature Neuroscience*, **22**(10), 1687–1695.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, **61**(2), 331–349.

Funahashi, Shintaro. 2017. Working memory in the prefrontal cortex. *Brain Sciences*, **7**(5).

Fusi, Stefano, Miller, Earl K., & Rigotti, Mattia. 2016. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, **37**, 66–74.

Fuster, J M. 2015. *The Prefrontal Cortex*. Elsevier, Academic Press.

Fuster, J M, & Alexander, G E. 1971. Neuron Activity Related to Short-Term Memory. *Science*, **173**(3997), 652–654.

Ganguli, Surya, & Latham, Peter. 2009. Feedforward to the Past: The Relation between Neuronal Connectivity, Amplification, and Short-Term Memory. *Neuron*, **61**(4), 499–501.

Glover, Gary H. 2011. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, **22**(2), 133–139.

Goldman, Mark S. 2009. Memory without Feedback in a Neural Network. *Neuron*, **61**(4), 621–634.

Goldman-Rakic, P. 1988. Topography Of Cognition: Parallel Distributed Networks In Primate Association Cortex. *Annual Review of Neuroscience*, **11**(1), 137–156.

Goldman-Rakic, P. S. 1992. Working memory and the mind. *Scientific American*, **267**(3), 111–117.

Gonzalez-Burgos, G., Kroener, S., Seamans, J. K., Lewis, D. A., & Barrionuevo, G. 2005. Dopaminergic modulation of short-term synaptic plasticity in fast-spiking interneurons of primate dorsolateral prefrontal cortex. *Journal of Neurophysiology*, **94**(6), 4168–4177.

Guest, Olivia, & Love, Bradley C. 2017. What the success of brain imaging implies about the neural code. *eLife*, **6**, 1–16.

Hamilos, Allison E., Spedicato, Giulia, Hong, Ye, Sun, Fangmiao, Li, Yulong, & Assad, John A. 2021. Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements. *eLife*, **10**, 1–38.

Hansel, David, & Mato, German. 2013. Short-term plasticity explains irregular persistent activity in working memory tasks. *The Journal of Neuroscience*, **33**(1), 133–149.

Harlow, J M. 1848. Passage of an iron rod through the head. *The Journal of neuropsychiatry and clinical neurosciences*, **11**(2), 281–283.

Harlow, J M. 1869. Recovery from the passage of an iron bar through the head. *Massachusetts Medical Society.*, **2**, 327–346.

Henderson, Margaret, Rademaker, Rosanne L, & Serences, John T. 2021. Flexible utilization of spatial- and motor-based codes for the storage of visuo- spatial information. *bioRxiv*.

Hennequin, Guillaume, Vogels, Tim P., & Gerstner, Wulfram. 2012. Non-normal amplification in random balanced neuronal networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **86**(1), 1–12.

Hirokawa, Junya, Vaughan, Alexander, Masset, Paul, Ott, Torben, & Kepecs, Adam. 2019. Frontal cortex neuron types categorically encode single decision variables. *Nature*, **576**(7787), 446–451.

Horn, Roger A., & Johnson, Charles R. 1985. *Matrix Analysis*. 2nd edn. Cambridge University Press.

Hussar, Cory R., & Pasternak, Tatiana. 2009. Flexibility of Sensory Representations in Prefrontal Cortex Depends on Cell Type. *Neuron*, **64**(5), 730–743.

Inagaki, Hidehiko K., Fontolan, Lorenzo, Romani, Sandro, & Svoboda, Karel. 2019. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, **566**(7743), 212–217.

Jacobsen, C F. 1935. Functions of frontal association area in primates. *Archives of Neurology & Psychiatry*, **33**(3), 558–569.

Jacobsen, C F. 1936. Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. *Comparative Psychology Monographs*, **13, 3**, 1–60.

Jahr, E, Health, Oregon, & Diego, San. 1990. Voltage Dependence of NMDA-Activated Predicted by Single-Channel Kinetics Macroscopic Conductances. *The Journal of Neuroscience*.

Jaramillo, Jorge, Mejias, Jorge F., & Wang, Xiao Jing. 2019. Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. *Neuron*, **101**(2), 321–336.e9.

Jha, Amishi P., Fabian, Sara A., & Aguirre, Geoffrey K. 2004. The role of prefrontal cortex in resolving distractor interference. *Cognitive, Affective and Behavioral Neuroscience*, **4**(4), 517–527.

Jin, Dezhe Z., Fujii, Naotaka, & Graybiel, Ann M. 2009. Neural representation of time in cortico-basal ganglia circuits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(45), 19156–19161.

Johnstone, Syren, & Rolls, Edmund T. 1990. Delay, discriminatory, and modality specific neurons in striatum and pallidum during short-term memory tasks. *Brain Research*, **522**(1), 147–151.

Jolliffe, Ian T, Cadima, Jorge, & Cadima, Jorge. 2016. Principal component analysis : a review and recent developments Subject Areas. *Phil.Trans.R.Soc.A*, **374**(20150202), 1–16.

Kandel, Eric R, Schwartz, James H, & Jessell, Thomas M. 2000. *Principles of neural science*. 4th ed edn. New York SE -: McGraw-Hill, Health Professions Division New York.

Kao, Ta Chu, & Hennequin, Guillaume. 2019. Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. *Current Opinion in Neurobiology*, **58**, 122–129.

Kilpatrick, Zachary P. 2013. Interareal coupling reduces encoding variability in multi-area models of spatial working memory. *Frontiers in Computational Neuroscience*, **7**(July), 1–14.

Kim, S., Rouault, H., Seelig, J.D., Druckmann, Shaul, & Jayaraman, Vivek. 2016. Ring attractor dynamics in the Drosophila central complex. *Society for Neuroscience*, **853**(San Diego, CA), 849–853.

King, J. R., & Dehaene, S. 2014. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, **18**(4), 203–210.

Knight, Bruce W. 1972. Dynamics of encoding in a population of neurons. *Journal of General Physiology*, **59**(6), 734–766.

Kobak, Dmitry, Brendel, Wieland, Constantinidis, Christos, Feierstein, Claudia E, Kepecs, Adam, Mainen, Zachary F, Qi, Xue-Lian, Romo, Ranulfo, Uchida, Naoshige, & Machens, Christian K. 2016. Demixed principal component analysis of neural population data. *eLife*, **5**, 1–36.

Lapicque, Louis. 2007. Quantitative investigations of electrical nerve excitation treated as polarization. 1907. *Biological cybernetics*, **97**(5-6), 341–349.

Lara, Antonio H., & Wallis, Jonathan D. 2015. The role of prefrontal cortex in working memory: A mini review. *Frontiers in Systems Neuroscience*, **9**(DEC), 1–7.

Latimer, Kenneth W, Yates, Jacob L, Meister, Miriam L R, Huk, Alexander C, & Pillow, Jonathan W. 2015. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 184–188.

Lawrence, Samuel J.D., van Mourik, Tim, Kok, Peter, Koopmans, Peter J., Norris, David G., & de Lange, Floris P. 2018. Laminar Organization of Working Memory Signals in Human Visual Cortex. *Current Biology*, **28**(21), 3435–3440.e4.

Leavitt, Matthew L., Mendoza-Halliday, Diego, & Martinez-Trujillo, Julio C. 2017. Sustained Activity Encoding Working Memories: Not Fully Distributed. *Trends in Neurosciences*, **40**(6), 328–346.

Lewinsohn, P M, Zieler, R E, Libet, J, Eyeberg, S, & Nielson, G. 1972. Short-term memory: a comparison between frontal and nonfrontal right- and left-hemisphere brain-damaged patients. *Journal of comparative and physiological psychology*, **81**(2), 248–255.

Lorenc, Elizabeth S., Sreenivasan, Kartik K., Nee, Derek E., Vandenbroucke, Annelinde R.E., & D'Esposito, Mark. 2018. Flexible coding of visual working memory representations during distraction. *Journal of Neuroscience*, **38**(23), 5267–5276.

Lorenc, Elizabeth S., Mallett, Remington, & Lewis-Peacock, Jarrod A. 2021. Distraction in Visual Working Memory: Resistance is Not Futile. *Trends in Cognitive Sciences*, 1–12.

Lundqvist, Mikael, Herman, Pawel, & Miller, Earl K. 2018. Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *The Journal of Neuroscience*, **38**(32), 7013–7019.

Mante, Valerio, Sussillo, David, Shenoy, Krishna V., & Newsome, William T. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, **503**(7474), 78–84.

Markowitz, David A., Curtis, Clayton E., & Pesaran, Bijan. 2015. Multiple component networks support working memory in prefrontal cortex. *Proceedings of the National Academy of Sciences*, **112**(35), 11084–11089.

Martin-Cortecero, Jesus, & Nuñez, Angel. 2016. Sensory responses in the medial prefrontal cortex of anesthetized rats. Implications for sensory processing. *Neuroscience*, **339**(October), 109–123.

Masse, Nicolas Y., Yang, Guangyu R., Song, H. Francis, Wang, Xiao Jing, & Freedman, David J. 2019. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, **22**(7), 1159–1167.

Mejías, Jorge F., & Wang, Xiao Jing. 2022. Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *eLife*, **11**, 1–33.

Mendoza-Halliday, Diego, & Martinez-Trujillo, Julio C. 2017. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nature Communications*, **8**, 1–13.

Meyers, Ethan M. 2018. Dynamic population coding and its relationship to working memory. *Journal of Neurophysiology*, **120**(5), 2260–2268.

Miles, Richard, & Wong, Robert K S. 1987. Latent synaptic pathways revealed after tetanic stimulation in the hippocampus. *Nature*, **329**(6141), 724–726.

Miller, Earl K., Lundqvist, Mikael, & Bastos, André M. 2018. Working Memory 2.0. *Neuron*, **100**(2), 463–475.

Milner, B. 1982. Some cognitive effects of frontal-lobe lesions in man. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **298**(1089), 211–226.

Mongillo, Gianluigi, Barak, Omri, & Tsodyks, Misha. 2008. Synaptic theory of working memory. *Science*, **319**(5869), 1543–1546.

Murphy, Brendan K., & Miller, Kenneth D. 2009. Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. *Neuron*, **61**(4), 635–648.

Murray, John D., Bernacchia, Alberto, Roy, Nicholas A., Constantinidis, Christos, Romo, Ranulfo, & Wang, Xiao Jing. 2017a. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(2), 394–399.

Murray, John D., Jaramillo, Jorge, & Wang, Xiao-Jing. 2017b. Working Memory and Decision-Making in a Frontoparietal Circuit Model. *The Journal of Neuroscience*, **37**(50), 12167–12186.

Nemes, Vanda A., Parry, Neil R.A., Whitaker, David, & McKeefry, Declan J. 2012. The retention and disruption of color information in human short-term visual memory. *Journal of Vision*, **12**(1), 1–14.

Oberauer, Klaus. 2009. Chapter 2 Design for a Working Memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, **51**, 45–100.

Oberauer, Klaus, Lewandowsky, Stephan, Farrell, Simon, Jarrold, Christopher, & Greaves, Martin. 2012. Modeling working memory: An interference model of complex span. *Psychonomic Bulletin and Review*, **19**(5), 779–819.

Oberauer, Klaus, Lewandowsky, Stephan, Awh, Edward, Brown, Gordon D.A., Conway, Andrew, Cowan, Nelson, Donkin, Christopher, Farrell, Simon, Hitch, Graham J., Hurlstone, Mark J., Ma, Wei Ji, Morey, Candice C., Nee, Derek Evan, Schweppe, Judith, Vergauwe, Evie, & Ward, Geoff. 2018. Benchmarks for models of short-term and working memory. *Psychological Bulletin*, **144**(9), 885–958.

Parthasarathy, Aishwarya, Tang, Cheng, Herikstad, Roger, Cheong, Loong Fah, Yen, Shih Cheng, & Libedinsky, Camilo. 2019. Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex. *Nature Communications*, **10**(1), 1–11.

Pasternak, Tatiana, & Zaksas, Daniel. 2003. Stimulus specificity and temporal dynamics of working memory for visual motion. *Journal of Neurophysiology*, **90**(4), 2757–2762.

Paton, Joseph J., & Buonomano, Dean V. 2018. The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, **98**(4), 687–705.

Pertzov, Yoni, Manohar, Sanjay, & Husain, Masud. 2017. Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, **43**(4), 528–536.

Rademaker, Rosanne L., Bloem, Ilona M., De Weerd, Peter, & Sack, Alexander T. 2015. The impact of interference on short-term memory for visual orientation. *Journal of Experimental Psychology: Human Perception and Performance*, **41**(6), 1650–1665.

Rademaker, Rosanne L., Chunharas, Chaipat, & Serences, John T. 2019. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, **22**(8), 1336–1344.

Raposo, David, Kaufman, Matthew T., & Churchland, Anne K. 2014. A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, **17**(12), 1784–1792.

Renart, Alfonso, Song, Pengcheng, & Wang, Xiao Jing. 2003. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, **38**(3), 473–485.

Rigotti, Mattia, Rubin, Daniel Ben Dayan, Wang, Xiao Jing, & Fusi, Stefano. 2010. Internal representation of task rules by recurrent dynamics: The importance of the diversity of neural responses. *Frontiers in Computational Neuroscience*, **4**(October), 1–29.

Rigotti, Mattia, Barak, Omri, Warden, Melissa R., Wang, Xiao Jing, Daw, Nathaniel D., Miller, Earl K., & Fusi, Stefano. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature*, **497**(7451), 585–590.

Romo, Ranulfo, Brody, Carlos D., Hernández, Adrián, & Lemus, Luis. 1999. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, **399**(6735), 470–473.

Ross, Jennifer A., & Van Bockstaele, Elisabeth J. 2021. The Locus Coeruleus- Norepinephrine System in Stress and Arousal: Unraveling Historical, Current, and Future Perspectives. *Frontiers in Psychiatry*, **11**(January), 1–23.

Sarma, Arup, Masse, Nicolas Y, Wang, Xiao-Jing, & Freedman, David J. 2016. Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nature neuroscience*, **19**(1), 143–149.

Shin, Hongsup, Zou, Qijia, & Ma, Wei Ji. 2017. The effects of delay duration on visual working memory for orientation. *Journal of Vision*, **17**(14), 1–24.

Simen, P., Balci, F., DeSouza, L., Cohen, J. D., & Holmes, P. 2011. A Model of Interval Timing by Neural Integration. *Journal of Neuroscience*, **31**(25), 9238–9253.

Spaak, Eelke, Watanabe, Kei, Funahashi, Shintaro, & Stokes, Mark G. 2017. Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *The Journal of Neuroscience*, **37**(27), 6503–6516.

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

Stokes, Mark G. 2015. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, **19**(7), 394–405.

Stokes, Mark G., Kusunoki, Makoto, Sigala, Natasha, Nili, Hamed, Gaffan, David, & Duncan, John. 2013. Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, **78**(2), 364–375.

Stroud, Jake P., Watanabe, Kei, Suzuki, Takafumi, Stokes, Mark G., & Lengyel, Máté. 2023. Optimal information loading into working memory in prefrontal cortex explains dynamic coding. *bioRxiv*.

Suzuki, Mototaka, & Gottlieb, Jacqueline. 2013. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*, **16**(1), 98–104.

Thura, David. 2020. Decision urgency invigorates movement in humans. *Behavioural brain research*, **382**(mar), 112477.

Traub, Roger D., & Jefferys, John G.R. 1994. Are there unifying principles underlying the generation of epileptic afterdischarges in vitro? *Progress in Brain Research*, **102**(C), 383–394.

Traub, Roger D, & Wong, Robert K S. 1982. Cellular Mechanism of Neuronal Synchronization in Epilepsy. *Science*, **216**(4547), 745–747.

Van Kerkoerle, Timo, Self, Matthew W., & Roelfsema, Pieter R. 2017. Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature Communications*, **8**.

Vogel, Edward K., & Machizawa, Maro G. 2004. Neural activity predicts individual differences in visual working memory capacity. *Nature*, **428**(6984), 748–751.

Vogel, Edward K., McCollough, Andrew W., & Machizawa, Maro G. 2005. Neural measures reveal individual differences in controlling access to working memory. *Nature*, **438**(7067), 500–503.

Wang, Xiao Jing. 1999. Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory. *Journal of Neuroscience*, **19**(21), 9587–9603.

Wang XJ. 2002. Probabilistic decision making by slow reverberation in cortical circuits. - PubMed - NCBI. *Neuron*, **36**(5), 955–68.

Watanabe, Kei, & Funahashi, Shintaro. 2014. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience*, **17**(4), 601–611.

Wei, Z., Wang, X.-J., & Wang, D.-H. 2012. From Distributed Resources to Limited Slots in Multiple-Item Working Memory: A Spiking Network Model with Normalization. *Journal of Neuroscience*, **32**(33), 11228–11240.

Wilson, Hugh R, & Cowan, Jack D. 1972. Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, **12**(1), 1–24.

Wimmer, Klaus, Nykamp, Duane Q, Constantinidis, Christos, & Compte, Albert. 2014. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, **17**(3), 431–439.

Wimmer, Klaus, Compte, Albert, Roxin, Alex, Peixoto, Diogo, Renart, Alfonso, & de la Rocha, Jaime. 2015. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications*, **6**(1), 6177.

Wolff, Michael J., Ding, Jacqueline, Myers, Nicholas E., & Stokes, Mark G. 2015. Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, **9**(september), 1–12.

Wong, Kong-Fatt, & Wang, Xiao-jing. 2006. A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *Journal of Neuroscience*, **26**(4), 1314–1328.

Wu, Zheng, Litwin-Kumar, Ashok, Shamash, Philip, Taylor, Alexei, Axel, Richard, & Shadlen, Michael N. 2020. Context-Dependent Decision Making in a Premotor Circuit. *Neuron*, **106**(2), 316–328.e6.

Xu, Yaoda. 2020. Revisit once more the sensory storage account of visual working memory. *Visual Cognition*, **28**(5-8), 433–446.

Yang, Guangyu Robert, & Molano-Mazón, Manuel. 2021. Towards the next generation of recurrent network models for cognitive neuroscience. *Current Opinion in Neurobiology*, **70**, 182–192.

Yang, Guangyu Robert, Joglekar, Madhura R., Song, H. Francis, Newsome, William T., & Wang, Xiao Jing. 2019. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, **22**(2), 297–306.

Yang, Weiguo, Tipparaju, Sri Laasya, Chen, Guang, & Li, Nuo. 2022. Thalamus-driven functional populations in frontal cortex support decision-making. *Nature Neuroscience*, **25**(10), 1339–1352.

Yoon, Jong H., Curtis, Clayton E., & D'Esposito, Mark. 2006. Differential effects of distraction during working memory on delay-period activity in the prefrontal cortex and the visual association cortex. *NeuroImage*, **29**(4), 1117–1126.

Zador, Anthony M. 2019. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, **10**(1).

Zaksas, Daniel, & Pasternak, Tatiana. 2006. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *Journal of Neuroscience*, **26**(45), 11726–11742.

Zhou, Xin, Katsuki, Fumi, Qi, Xue-Lian, & Constantinidis, Christos. 2012. Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex. *Journal of Neurophysiology*, **108**(1), 31–38.

Zylberberg, Joel, & Strowbridge, Ben W. 2017. Mechanisms of Persistent Activity in Cortical Circuits: Possible Neural Substrates for Working Memory. *Annual Review of Neuroscience*, **40**(1), 603–627.