Storing structured sparse memories in a multi-modular cortical network model

Alexis M. Dubreuil, Nicolas Brunel

Storing structured sparse memories in a multi-modular cortical network model

Alexis Dubreuil[1,2], Nicolas Brunel[1]

[1] Laboratoire de Physique Théorique, Ecole Normale Supérieure
[2] Laboratoire Jean Perrin, UPMC
[3] Departments of Statistics and Neurobiology, University of Chicago

Abstract

We study the memory performance of a class of modular attractor neural networks, where modules are potentially fully-connected networks connected to each other via diluted long-range connections. On this anatomical architecture we store memory patterns of activity using a Willshaw-type learning rule. P patterns are split in categories, such that patterns of the same category activate the same set of modules. We first compute the maximal storage capacity of these networks. We then investigate their error-correction properties through an exhaustive exploration of parameter space, and identify regions where the networks behave as an associative memory device. The crucial parameters that control the retrieval abilities of the network are (1) the ratio between the number of synaptic contacts of long- and short-range origins (2) the number of categories in which a module is activated and (3) the amount of local inhibition. We discuss the relationship between our model and networks of cortical patches that have been observed in different cortical areas.

1 Introduction

Attractor neural networks have been used extensively to model memory phenomena in the brain (e.g. Hopfield, 1982; Amit, 1989; Wang, 2001; Brunel, 2005). In these models a memory is represented by a pattern of activity that is able to self-sustain, thanks to recurrent connectivity, in the absence of any external stimulus. They naturally possess associative properties. For instance if an incomplete pattern of activity representing part of a memory is presented to the network, it will use recurrent connectivity to reconstruct the complete memory pattern. Moreover, attractor dynamics naturally account for the phenomenon of persistent activity observed in electrophysiological recordings across different cortical areas during working memory tasks (Fuster and Alexander, 1971; Miyashita, 1988; Funahashi et al., 1989; Miller et al., 1996; Romo et al., 1999).

Most modeling studies have focused on densely connected networks (where a large fraction of the connections between neurons can be shaped by the stored memories), which are appropriate for small cortical networks at the scale of a few hundred microns (Hellwig, 2000; Kalisman et al., 2005). Other studies have introduced a probability of connection that depends on the distance between neurons (Roudi and Treves, 2004, 2006), which allows to take into account the fact that the connection probability decreases with distance as has been shown for networks of larger sizes (Hellwig, 2000). When observed at larger scale, cortical connectivity is not randomly distributed with a distance-dependent parameter, but rather shows a non-trivial structure. For instance, Pucak et al.

(1996) have shown that connectivity from/to patches of pre-frontal cortex of monkeys of a few hundred microns send/receive connections from other discrete patches of cortex that have the shape of stripes of sizes of a few hundred microns. One patch is connected to about 15-20 other patches in the same or neighboring areas via grey matter connections, and at least 15-20 other patches connected via white matter connections. Other experimental studies have identified such a patchy connectivity in sensory cortices (DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Bosking et al., 1997). A few modeling studies have investigated associative memory properties of networks that implement a dichotomy between dense local connectivity and sparse long-range connectivity (O'Kane and Treves, 1992; Mari and Treves, 1998; Kropff and Treves, 2005; Johansson and Lansner, 2007).

In the present work, we study modular networks whose modules are fully connected networks (short-range connections) connected through long-range diluted connections. In order to match available experimental data, we impose that the numbers of short-range and long-range connections onto a neuron are of the same order (Braitenberg and Schütz, 1991; Stepanyants et al., 2009). We model these networks with binary neurons and binary synapses, for which the storage properties of fully-connected networks have already been extensively characterized (Willshaw et al., 1969; Knoblauch et al., 2010; Dubreuil et al., 2014). For such models, the distribution of the total synaptic input onto neurons can be expressed analytically, which allows to probe their associative memory properties. During a learning phase, the storage of patterns of activity is implemented using a Willshaw learning rule, that potentiates a fraction of the pre-existing synapses. The patterns of activity, or memories, that are stored reflect the modular architecture, in the sense that each of the memory consists in the activation of only a subset of the modules. Moreover, patterns are split in categories: two patterns in the same category share the same active modules. This is consistent with MRI studies which show that visually perceived objects that are semantically close to each other are represented in a similar manner on the cortical surface (Huth et al., 2012).

After defining our model, we describe the method that we use to quantify its storage capacity. We then apply this method to compute the maximal storage capacity the networks can reach when no associative properties are required. We then define three canonical error-correction tasks to be performed by the network, delimit the parameter regions where satisfactory associative properties are reached, and quantify the storage capacity in these regions. Last, we discuss in more detail the link between our model and cortical networks with patchy connectivity.

2 Network model and methods

We consider a network of M modules of N binary neurons connected through a binary connectivity matrix. Below we describe the dynamics of the network, specify the characteristics of the patterns of activity that are stored by the network, and describe the connectivity matrix that underlie the storage of these patterns.

2.1 Dynamics

The activity of neuron i in module m ($i = 1...N; m = 1...M$) is described by a binary variable $S_{i,m} = 0, 1$, that evolves in time according to

$$S_{i,m}(t+1) = \Theta\left[h_{i,m}(t) - \theta f N\right], \tag{1}$$

where

$$h_{i,m}(t) = h^l_{i,m}(t) + h^e_{i,m}(t) = \sum_{j=1}^{N} W^{m,m}_{ij} S_{j,m} + \sum_{n \neq m} \sum_{j=1}^{N} W^{m,n}_{ij} S_{j,n} \tag{2}$$

is the total synaptic input on neuron (i, m), i.e. the sum of a local field $h^l_{i,m}(t)$ resulting from the activity of neurons belonging to the same module and an external field $h^e_{i,m}(t)$ resulting from the activity of neurons belonging to other modules. $\theta = O(1)$ is an activation threshold, Θ is the Heaviside function, $W^{m,m}_{ij}$ is the efficacy of the synapse from neuron j to neuron i (both belonging to the same module m), while $W^{m,n}_{ij}$ is the the efficacy of the synapse from neuron j in module n to neuron i in module m.

2.2 Structured sparse memories

The learning process is assumed to have led to the storage of P patterns of activity (network states). A given network state, or pattern, is said to be stored if it is a fixed point of the dynamics (1). The activity of neuron i in module m in pattern μ, $\Xi^\mu_{i,m}$, is a binary 0,1 variable given by the product of two binary variables, a macroscopic term Ξ^μ_m and a microscopic term $\xi^\mu_{i,m}$

$$\Xi^\mu_{i,m} = \Xi^\mu_m \xi^\mu_{i,m} \in \{0, 1\} \tag{3}$$

At the macroscopic scale, a fraction F (macroscopic coding level) of the modules are active in each pattern,

$$\sum_{m=1}^{M} \Xi^\mu_m = FM \tag{4}$$

At the microscopic scale, a fraction f (microscopic coding level) of the neurons are active in any active module m

$$\sum_{i=1}^{N} \xi^\mu_{i,m} = fN \tag{5}$$

Patterns are split in $\mathscr{P}$ categories, with p patterns in each category ($P = p\mathscr{P}$). Patterns belonging to the same category have the same set of active modules.
We further impose that each module m is involved in the same number

$$c = \mathscr{P}F \tag{6}$$

of categories, this leads to the constraint that $\mathscr{P}F$ should be an integer. Fixing this number for each module rather than letting it fluctuate from module to module, is the optimal choice in terms of storage. In section 4.3 we discuss the effects on storage capacity of fluctuations in numbers of active neurons/modules per pattern, and numbers of categories encoded per module.

2.3 Connectivity

The connection W_{ij}^{mn} from neuron (j,n) to neuron (i,m) is described by a binary variable $\in \{0,1\}$. The storage capacity of fully-connected networks of binary synapses has been extensively characterized (Willshaw et al., 1969; Knoblauch et al., 2010; Dubreuil et al., 2014). We can thus use these studies as a benchmark to which the storage capacity of modular networks can be compared. The connectivity between these two neurons is determined by two factors, learning and architectural constraints. As a result, W_{ij}^{mn} is a product of two terms,

$$W_{ij}^{mn} = w_{ij}^{mn} d_{ij}^{mn} \text{ with } w_{ij}^{mn}, d_{ij}^{mn} \in \{0,1\} \tag{7}$$

The learning term w_{ij}^{mn} follows a Willshaw type learning rule (Willshaw et al., 1969): the learning variables are initialized with $w_{ij}^{mn} = 0$ and after a learning phase where patterns are imposed on the network, they are switched to $w_{ij}^{mn} = 1$ if there exists at least one pattern μ such that $\Xi_{i,m}^{\mu} = \Xi_{j,n}^{\mu} = 1$.

The architectural constraint described by d_{ij}^{mn} is an asymetric dilution term (d_{ij}^{mn} is not necessarily equal to d_{ji}^{mn}). We consider networks with potentially fully connected modules ($d_{ij}^{mm} = 1$ for all i,j,m), that is every local synapse W_{ij}^{mm} can be potentiated during the learning phase. This models the fact that local cortical circuits could be potentially fully connected, in the sense that the axon of any neuron touches (i.e. passes $<2\mu$m by) the dendritic tree of all nearby neurons (Kalisman et al., 2005; Hellwig, 2000). For the connectivity between modules, we distinguish two cases. If two modules m and n are never co-activated in a pattern, then long-range connections do not exist between these modules ($d_{ij}^{mn} = 0$ for all i,j). If there exists one category in which m and n are co-activated,

$$d_{ij}^{mn} = \begin{cases} 1 & \text{with probability } \frac{D}{N} \\ 0 & \text{with probability } 1 - \frac{D}{N}. \end{cases} \tag{8}$$

This implies that neuron (i,m) receives on average D connections from module n. In order to match available experimental data (Braitenberg and Schütz, 1991; Stepanyants et al., 2009), the amount of dilution $\frac{D}{N}$ is chosen such that a given neuron in module m receives a number of contacts from axons originating in remote modules ($n \neq m$) that is of the same order than the number of contacts from axons originating in the local module m. To quantify this, we introduce the parameter γ, defined as the ratio between the number of long-range contacts, and the number of short-range contacts. With the previously introduced parameters,

$$\gamma = \frac{rMD}{N}, \tag{9}$$

where $r = 1 - (1 - F^2)^{\mathscr{P}}$ is the fraction of pairs of modules that have been co-activated at least once during the learning phase. The numerator is then the number of contacts that originate from long-range connections, while the denominator N is the number of contacts that originate from short-range connections. The architecture of the model is described schematically in Fig. 1A-B.

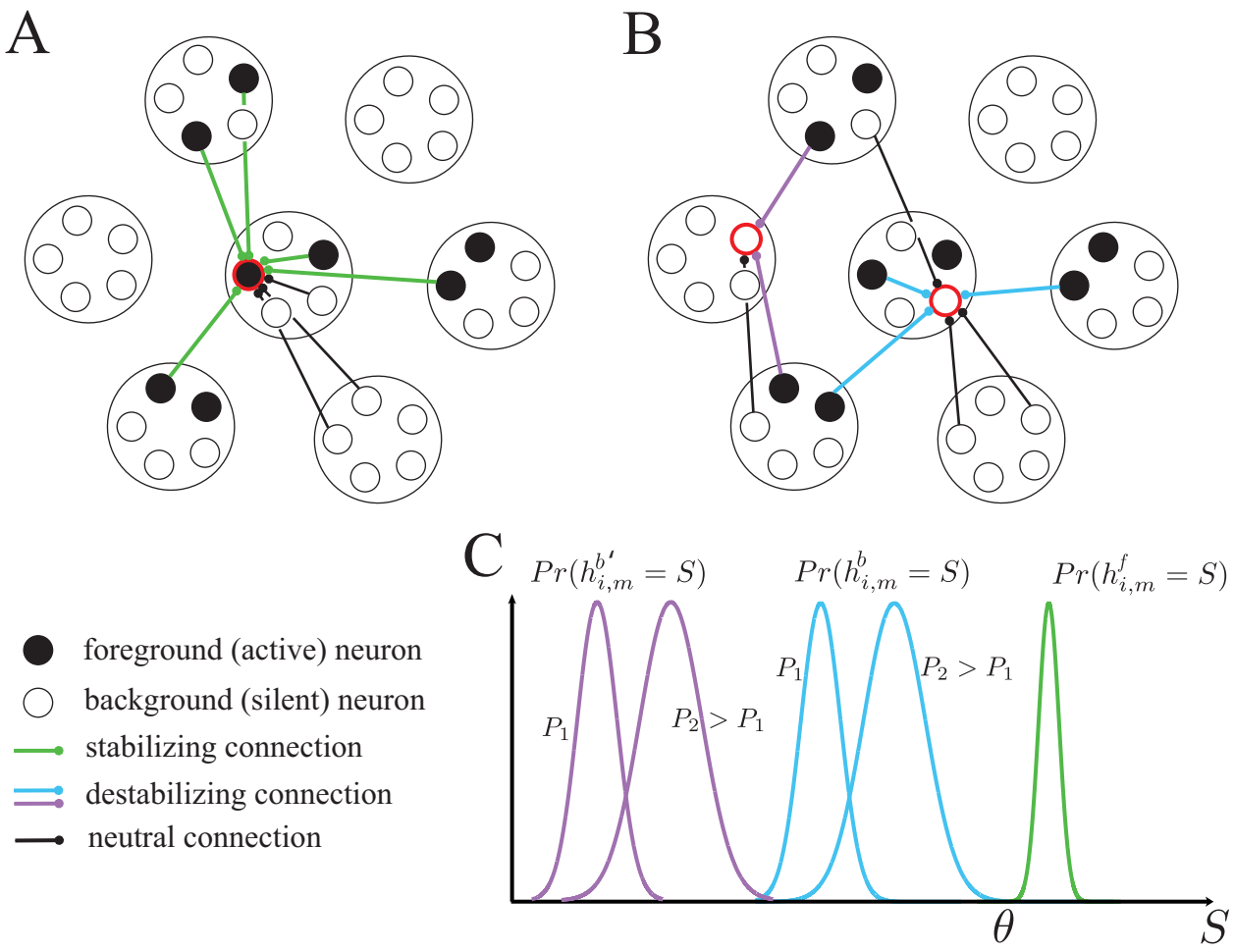

Figure 1: *Computing the number of patterns that can be stored in modular networks. After the learning phase using the Willshaw learning rule, the network is set in one of the stored patterns $\vec{\Xi}^{\mu_0}$, which is said to be stored if it is a stable fixed point of the network dynamics. A. Connectivity onto a foreground neuron for the memory μ_0 (red circle). Connections shown in green create feedback loops between foreground neurons which stabilizes pattern $\vec{\Xi}^{\mu_0}$. Black connections from silent neurons do not influence the stability of this network state, they have been potentiated during the presentation of a pattern $\mu \neq \mu_0$. B. Connectivity onto background neurons (red circles), one that belong to a module active in pattern $\vec{\Xi}^{\mu_0}$ and another one that belong to an inactive module. Connections onto these neurons result from the presentation of other patterns $\mu \neq \mu_0$ in which these neurons are active. The blue and magenta connections from foreground neurons provide excitation to this neuron, which can potentially destabilize $\vec{\Xi}^{\mu_0}$. C. The stability of a pattern can be assessed by evaluating the probability distributions of the inputs on background neurons (blue and magenta) and on foreground neurons (green). A tested pattern is stable if the probability that inputs of all background neurons to be above the activation threshold θ, as well as the probability that inputs of all foreground neurons to be below the activation threshold, are vanishingly small in the large N, M limits. When more patterns are stored in the synaptic matrix (from P_1 to $P_2 > P_1$), the distribution of inputs on background neurons shifts its mean towards θ and gets wider. Evaluating storage capacity consists in computing the largest P for which a tested pattern is stable.*

2.4 Scaling of parameters

In the large N limit networks of binary synapses have non-vanishing storage capacity if the microscopic coding level scales as $f \propto \frac{\ln N}{N}$ (Willshaw et al., 1969; Nadal, 1991; Knoblauch et al., 2010; Dubreuil et al., 2014). To work in this regime we introduce the parameter β

$$f = \beta \frac{\ln N}{N} \text{ with } \beta = O(1) \tag{10}$$

In a previous study of modular attractor networks, Mari and Treves (1998) have shown that if the macroscopic coding level F goes to zero when the number of modules becomes large, the number of patterns stored in the network increases with M. We will also study our networks in this regime. We will show that similar storage capacities are reached in both cases $FM \to \infty$ and $FM = O(1)$. Furthermore, for the following calculations to be valid, we assume that F is sufficiently small to have $FM \ll N$.

Finally, we have to specify how patterns are split in categories. An inspection of the terms quantifying the fraction of synaptic contacts activated during the learning phase allows us to anticipate which of the different regimes ($p = O(1), \mathscr{P} \to \infty$; $p \to \infty, \mathscr{P} = O(1)$ or $p \to \infty, \mathscr{P} \to \infty$) leads to largest storage capacity. After the learning phase, the fraction of short-range synaptic weights w_{ij}^{mm} that have been switched to 1 during the learning phase is $1 - (1 - f^2)^{Fp\mathscr{P}}$, and the fraction of long-range synaptic weights w_{ij}^{mn} (describing a pair of modules that have been co-activated at least once) that have been switched to 1 during the learning phase is $1 - (1 - f^2)^{p(1+\mathscr{P}F^2)}$. Both fractions should tend to a value $\in]0, 1[$, otherwise storage will be suboptimal: if one of this fraction is 0, a vast majority of the corresponding synapses are in their silent state, and therefore the connectivity matrix contains a vanishing information about the set of stored patterns as they do not participate in shaping the synaptic currents of specific patterns ; the same is true when the fraction is 1. Inspection of the expressions above with the scalings $f \propto \frac{\ln N}{N}$ and $F \to 0$ leads to the conclusion that the optimal situation is reached for $\mathscr{P} \propto \frac{1}{F} \to \infty$, and $p \propto \frac{1}{f^2} \to \infty$, i.e. each module is involved in a finite number of categories (see equation (6)), with a number of patterns in each category that can be measured by what we define as the **storage load** α

$$\alpha = pf^2 \tag{11}$$

With $\mathscr{P}F = c = O(1)$ and $F \to 0$ the expression of γ introduced in equation (9) is

$$\gamma = cFM\frac{D}{N}, \tag{12}$$

which sets the amount of dilution of the long-range connections, since we keep γ of order 1.

We thus focus on the regimes:

- $N \gg M \to \infty$;
- $f = \beta \frac{\ln N}{N}$ and $F \to 0$, with $FM = O(1)$ or $FM \to +\infty$ and $FM \ll N$;

- $pf^2 = \alpha = O(1)$ and $\mathscr{P}F = c = O(1)$
- $\gamma = O(1)$ with $\frac{D}{N} = O(\frac{1}{FM})$

2.5 Analytical methods

Our method consists in computing the distribution of inputs to different types of neurons of the network. From these distributions, we assess the stability of learned patterns and compute the storage capacity. An intuitive picture is given in figure 1C.

After the learning phase, we choose one of the P presented patterns $\vec{\Xi}^{\mu_0}$, set the network in a state corresponding to this pattern $\vec{S} = \vec{\Xi}^{\mu_0}$, and test whether it is a fixed point of the dynamics (1).

The stability of pattern $\vec{\Xi}^{\mu_0}$ is assessed by computing the probability $\mathbb{P}_{ne}$ that the fields on all neurons are on the right side of the activation threshold $\theta f N$ (see figure 1C for an illustration). To do so, we have to distinguish between three types of neurons: foreground neurons (neurons (i,m) such that $\Xi^{\mu_0}_{i,m} = 1$, figure 1A), background neurons that belong to an active module ((i,m) such that $\Xi^{\mu_0}_{i,m} = 0$ but $\Xi^{\mu_0}_{m} = 1$, figure 1B) and background neurons that belong to an inactive module ((i,m) such that $\Xi^{\mu_0}_{i,m} = 0$ and $\Xi^{\mu_0}_{m} = 0$). The probability of $\vec{\Xi}^{\mu_0}$ being a fixed point of (1) can be written

$$\begin{aligned}
\mathbb{P}_{ne} &= \left(1 - \mathbb{P}(h_{i,m} \leq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 1)\right)^{FMfN} \\
&\quad \times \left(1 - \mathbb{P}(h_{i,m} \geq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 0,\, \Xi^{\mu_0}_{m} = 1)\right)^{FM(1-f)N} \\
&\quad \times \left(1 - \mathbb{P}(h_{i,m} \geq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 0,\, \Xi^{\mu_0}_{m} = 0)\right)^{(1-F)MN}
\end{aligned} \tag{13}$$

In the limit of large networks and for a sparse microscopic coding level, these probabilities take the form (see Appendix)

$$\begin{aligned}
\mathbb{P}(h_{i,m} \leq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 1) &= \exp\left[-fN\Phi^f + o(fN)\right] \\
\mathbb{P}(h_{i,m} \geq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 0,\, \Xi^{\mu_0}_{m} = 1) &= \exp\left[-fN\Phi^b + o(fN)\right] \\
\mathbb{P}(h_{i,m} \geq fN\theta \,|\, \Xi^{\mu_0}_{i,m} = 0,\, \Xi^{\mu_0}_{m} = 0) &= \exp\left[-fN\Phi^{b'} + o(fN))\right]
\end{aligned} \tag{14}$$

where the Φ's are rate functions that describe the behavior of the tails of the relevant probability distributions (Φ^f: foreground neurons ; Φ^b: background neurons in active modules ; $\Phi^{b'}$: background neurons in silent modules). These rate functions depend on network parameters, pattern parameters and the number of stored patterns P. Eqs. (14) allow to rewrite P_{ne} as

$$\mathbb{P}_{ne} = \exp\left[-\exp(X_s) - \exp(X_n) - \exp(X_{n'})\right] \tag{15}$$

with

$$\begin{aligned}
X_s &= -\beta\Phi^f \ln N + o(\ln N) + O(\ln FM) \\
X_n &= -\beta\Phi^b \ln N + \ln N + o(\ln N) + O(\ln FM) \\
X_{n'} &= -\beta\Phi^{b'} \ln N + \ln N + o(\ln N) + O(\ln(1-F)M)
\end{aligned} \tag{16}$$

For $\mathbb{P}_{ne}$ to go to 1 in the large N limit, we need all X's to go to $-\infty$ in that limit. Given $FM \ll N$, this will be satisfied provided that

$$\Phi^f(\theta) > 0 \tag{17}$$

and

$$\Phi^b(\theta) > \frac{1}{\beta}. \tag{18}$$

Inequality (17) is illustrated schematically in figure 1C: the activation threshold θ has to be chosen such that inputs to the $FMfN$ foreground neurons drawn from the green distribution are above θ. Inequality (18) is also illustrated in figure 1C: the activation threshold θ has to be chosen large enough, such that inputs to the $FM(1-f)N$ background neurons belonging to active modules drawn from the blue distribution are below θ. There is no inequality involving the rate function related to errors in modules that are silent because the probability to activate a neuron in these modules is much lower than the one to activate neurons in active modules that receive local noise on top of external noise, as can be seen from equation (67) in the Appendix, which implies that we always have $\beta\Phi^{b'} \ln N \gg \ln M$ for $FM \ll N$. This is schematically represented in figure 1C by the fact that the average of magenta distributions of inputs (on background neurons in inactive modules) are further away from the activation threshold θ than are the blue distributions of inputs (on background neurons in active modules).

For a given set of parameters (β, c, γ, F), one can thus find the maximal number of patterns P_{max} that can be imprinted in the synaptic matrix while keeping pattern $\vec{\Xi}^{\mu_0}$ a fixed point of the dynamics. To do so, we saturate the two above inequalities. Inequality (17) is saturated by taking an activation threshold θ that goes to $\langle h_{i,m} \,|\, \Xi_{i,m}^{\mu_0} = 1 \rangle$ in the large N limit. The threshold can be chosen in this way because the number of foreground neurons scales as $\ln N$ (Dubreuil et al., 2014). Inequality (18) is saturated by choosing a storage load $\alpha = \alpha_{max}$ such that $\Phi^b \simeq \frac{1}{\beta}$. Then from equations (6),(11),

$$P_{max} = \frac{\alpha_{max}}{\beta^2} \frac{c}{F} \left(\frac{N}{\ln N} \right)^2 \tag{19}$$

We define the storage capacity of the network (see e.g. Nadal (1991); Knoblauch et al. (2010)) as

$$I = \frac{P_{max} I_{pattern}}{N_p} \tag{20}$$

where

$$\begin{aligned} I_{pattern} &= MFN \left(-f \ln_2 f - (1-f)\ln_2(1-f)\right) \\ &\quad + \frac{\mathscr{P}}{P_{max}} M \left(-F \ln_2 F - (1-F)\ln_2(1-F)\right) \end{aligned} \tag{21}$$

is the information content (entropy) of each pattern. The term on the first line is the contribution from the microscopic structure, and the other term is the contribution from the macroscopic structure. The factor $\frac{\mathscr{P}}{P_{max}}$ is due to the fact that there exists only $\mathscr{P}$, not P_{max}

macroscopic patterns (i.e. categories). This expression is a generalization of the entropy of the distribution of patterns $N\left(-f\ln_2 f-(1-f)\ln_2(1-f)\right)$ used to define storage capacity in networks storing unstructured patterns. The denominator of (20),

$$\begin{aligned} N_p &= MN^2+M(M-1)N^2r\frac{D}{N} \\ &\simeq MN^2(1+\gamma) \end{aligned} \tag{22}$$

is the number of modifiable synapses, i.e. the amount of physical substrate used to store patterns. I can be thought of as the total amount of information stored in the network, in bits per modifiable synapse (but see Discussion).

With parameters scaling as described in section 2.4, and using the definitions (5)-(6), the storage capacity can be rewritten

$$I=\frac{1}{\ln 2}\frac{\alpha_{max}c}{\beta(1+\gamma)} \tag{23}$$

where α_{max} can be computed for a fixed set of parameters (c, β,γ,FM) using the method described in 3.1.

3 Results

We now apply the method described above to quantify the storage properties of these modular networks for different values of the parameters. We first compute the maximal storage capacity that can be reached by modular networks, and then study their error-correction properties.

3.1 Maximal storage capacity

In this section, the network state is initialized in one of the stored pattern $\vec{\Xi}^{\mu_0}$ and we find the largest P such that this network state is stable under the dynamics given by eq. (1). From this value of P we compute the storage capacity I. The stability of a pattern is assessed by expressing the fields $h^f_{i,m}$ and $h^b_{i,m}$ on foreground and background neurons. The averages of these fields are

$$\begin{aligned} \langle h^f_{i,m}\rangle &= fN+FMfD=fN(1+\frac{\gamma}{c}) \\ \langle h^b_{i,m}\rangle &= fNg+FMfDG=fN(g+\frac{\gamma}{c}G) \end{aligned} \tag{24}$$

where

$$g=1-(1-f^2)^{p\mathscr{P}F}\underset{f\to 0}{\simeq}1-\exp(-\alpha c) \tag{25}$$

is the fraction of short-range synapses w^{mm}_{ij} that have been switched to 1 during learning and

$$G=1-(1-f^2)^{p(1+\mathscr{P}F^2)}\underset{f,F\to 0}{\simeq}1-\exp(-\alpha) \tag{26}$$

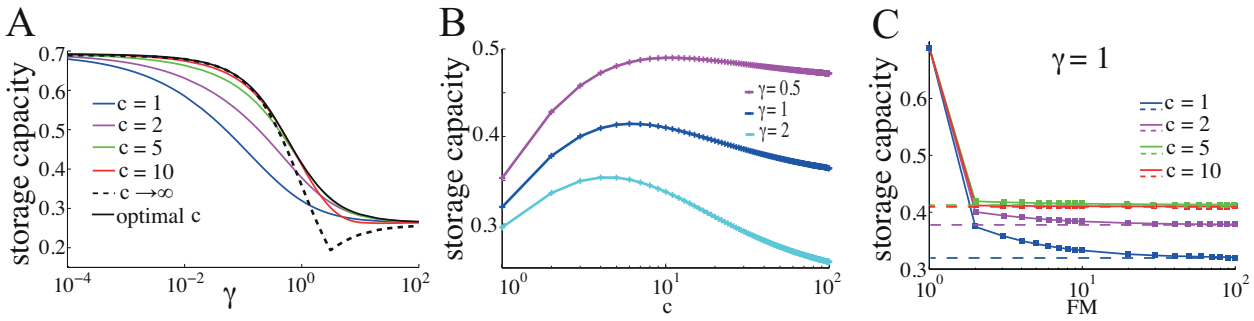

Figure 2: *Storage capacity of modular networks. A. Storage capacity in the regime* $FM \to \infty$ *as a function of* γ*, the ratio between the numbers of synaptic contacts from long-range and short-range origins. It is plotted for different values of* c*, the number of categories in which a module is activated. B. Storage capacity as a function of* c*, the number of categories in which each module is involved. C. Storage capacity as a function of* $FM = O(1)$*, the number of active modules in memory patterns. For each value of* c *dashed lines mark the capacity in the limit* $FM \to +\infty$ *for* $\gamma = 1$*.*

is the fraction of long-range connections w_{ij}^{mn}, $m \neq n$, that have been switched to 1 during learning for two modules m and n that have been co-activated at least once. In (26), α is the storage load defined in (11). As can be seen from equations (25),(26), and as illustrated on figure 1C, when the number of stored patterns is increased, the mean of the distribution of $h_{i,m}^b$ shifts towards the mean of the distribution of $h_{i,m}^f$ and it becomes more difficult to find an activation threshold separating these two inputs. In order to assess the stability of pattern $\vec{\Xi}^{\mu_0}$, we express the distribution of the inputs via the rate functions (65),(66) when $FM \gg 1$ and (68),(69) when $FM = O(1)$, and apply the method described in the 'Methods' section. The storage capacity is then expressed for given values of c, β, γ using equation (23) and choosing α that saturates inequality (18), while the activation threshold is chosen as large as possible to saturate inequality (17). Given the expression of the rate functions describing the field on selective neurons, we can choose $\theta \to \langle h_{i,m}^f \rangle$ for all neurons. Note that we have only discussed the fields on background neurons that belong to active modules. For background neurons in silent modules their average field is

$$\langle h_{i,m}^{b'} \rangle = FMfDG' = fN\frac{\gamma}{c}G' \tag{27}$$

where

$$G' = 1 - (1 - f^2)^{p\mathscr{P}F^2} \underset{f,F\to 0}{\simeq} \alpha c F \tag{28}$$

is the fraction of long-range connections w_{ij}^{mn}, $m \neq n$, that have been switched to 1 during learning for two randomly chosen modules. As mentioned in the 'Methods' section, these neurons do not constraint storage capacity because the field they receive is small due to the absence of local inputs and to the fact that G' vanishes in the limit $F \to 0$.

With the scaling introduced above, we find that I is non-zero, which means that the

maximal number of patterns P_{max} that can be stored scales as

$$P_{max} \propto \frac{1}{F}\left(\frac{N}{\ln N}\right)^2 \tag{29}$$

In figure 2A, we plot the value of the storage capacity (23) as a function of γ for different values of c, the number of categories in which a module is activated as defined in equation (6), and the scaling $FM \to +\infty$. For each (γ, c) we choose the value of β that maximizes I. In practice the optimization over β is performed in two steps: for a range of values of the storage load α, the value of β that saturates inequality (18) is chosen using the expression of Φ given in the Appendix ; and the pair (α, β) that maximizes I according to (23) is kept. For small γ (most of the connections are short-range) $I = 0.69$, which corresponds to the storage capacity of a fully-connected Willshaw network (Willshaw et al., 1969; Knoblauch et al., 2010; Dubreuil et al., 2014). For large values of γ (most of the connections are long-range), $I = 0.26$, the storage capacity of a highly-diluted Willshaw network as shown in the Appendix. In between these two limits, the storage capacity interpolates between the limits of a fully connected network and a highly diluted network, similar to a previous model of modular attractor network (O'Kane and Treves, 1992). The same trend is observed when different numbers of categories are involved in each module (i.e. different values of c). For $c \to +\infty$, the curve shows a discontinuity in its derivative, which is due to the fact that in this limit we have either $g = O(1)$ and $G \to 0$ (a vanishingly small fraction of information stored on long-range synaptic connections), or $g \to 1$ (a vanishingly small fraction of information stored on short-range synaptic connections) and $G = O(1)$. For low values of γ, storage capacity is optimized by using short-range connections ($g = O(1)$ and $G \to 0$), while for large values of γ it is optimized by using long-range connections ($g \to 1$ and $G = O(1)$).
Figure 2B shows the storage capacity as a function of c, the number of categories in which each module is active. Interestingly, storage capacity depends non-monotonically on c, with an optimum at $c = 10, 6, 4$ for $\gamma = \frac{1}{2}, 1, 2$.
We have also studied the case $FM = O(1)$ and found that on a broad range of values of FM, the storage capacity is very similar to the case $FM \to \infty$ as shown in figure 2C, where $\gamma = 1$. A similar behavior is obtained for other values of γ. Note that with this scaling, the maximal number of stored patterns is proportional to the number of modules $P_{max} \propto M\left(\frac{N}{\ln N}\right)^2$.

3.2 Error-correction capabilities of the modular network

We have shown in the previous section, that networks with $\gamma \to 0$ are optimal for storage capacity. The extreme case $\gamma = 0$ corresponds to a network with exclusively local connections. An important drawback of such a network is there is no communication between modules. This will be fine in the case where the network state is initialized in a state that exactly correspond to a stored pattern, but can be problematic if the network initial state differs from it. In the following section, we probe the error-correction properties of the modular network by studying how its storage properties are affected when a given level of error-correction is required, and how this depends on the values of γ and other parameters. Note that our analytical method can only tell us what happens after a single time

step of the dynamics, and therefore require that error correction should be performed on a single time step of the dynamics (1).

Below we consider three different kinds of errors to be corrected

- Microscopic pattern-completion where the initial state $\vec{S}^{\mu_0}$ is obtained from a particular memory $\vec{\Xi}^{\mu_0}$ by switching on or off a fraction of the neurons in all active modules (see figure 3A).
- Disambiguation where the initial state $\vec{S}^{\mu_0}$ is obtained from $\vec{\Xi}^{\mu_0}$ by setting a fraction of the active modules (m such that $\Xi_m^{\mu_0} = 1$) in a state that corresponds to other patterns $\mu \neq \mu_0$ (see figure 4A).
- Macroscopic pattern-completion where the initial state $\vec{S}^{\mu_0}$ is obtained from $\vec{\Xi}^{\mu_0}$ by silencing a fraction of the active modules and activating a fraction of the silent modules (see figure 5A).

In all three cases errors are introduced while keeping the overall activity level unchanged, i.e. the overall number of neurons active in $\vec{S}^{\mu_0}$ is the same as in $\vec{\Xi}^{\mu_0}$. We quantify the initial amount of errors by a parameter E, such that for a given value of E there is the same number of false-positives ($\simeq MFNfE$) and false-negatives ($\simeq MFNfE$). A precise definition of the initial states to be corrected is given in the following.

3.2.1 Microscopic pattern-completion

Here, we initialize the network in a state $\vec{S}^{\mu_0}$, which is obtained from $\vec{\Xi}^{\mu_0}$ by randomly flipping neurons in active modules. The state of a neuron (i, m) belonging to an active module m ($\Xi_m^{\mu_0} = 1$) becomes

$$S_{i,m}^{\mu_0} = \Xi_{i,m}^{\mu_0} \left(1 - X_{i,m}^{f}\right) + \left(1 - \Xi_{i,m}^{\mu_0}\right) X_{i,m}^{b} \tag{30}$$

where

$$X_{i,m}^{f} = \begin{cases} 1 & \text{with probability } (1-f)E \\ 0 & \text{with probability } 1-(1-f)E \end{cases} \tag{31}$$

while

$$X_{i,m}^{b} = \begin{cases} 1 & \text{with probability } fE \\ 0 & \text{with probability } 1-fE \end{cases} \tag{32}$$

In order for the network state to flow towards $\vec{\Xi}^{\mu_0}$ in a single time step of the dynamics (1), we need an activation threshold that is above the input received by 'false positive' neurons ($S_{i,m}^{\mu_0} = 1$ and $\Xi_{i,m}^{\mu_0} = 0$, neurons filled with black with no red circles in figure 3A), and below the input received by 'false negative' neurons ($S_{i,m}^{\mu_0} = 0$ and $\Xi_{i,m}^{\mu_0} = 1$, unfilled neurons circled in red). Considering the averages of the inputs to false negatives and false positives, the activation threshold has to satisfy

$$g + \frac{\gamma}{c}G < \frac{\theta}{fN} < (1-E) + Eg + \frac{\gamma}{c}(1-E) + \frac{\gamma}{c}GE \tag{33}$$

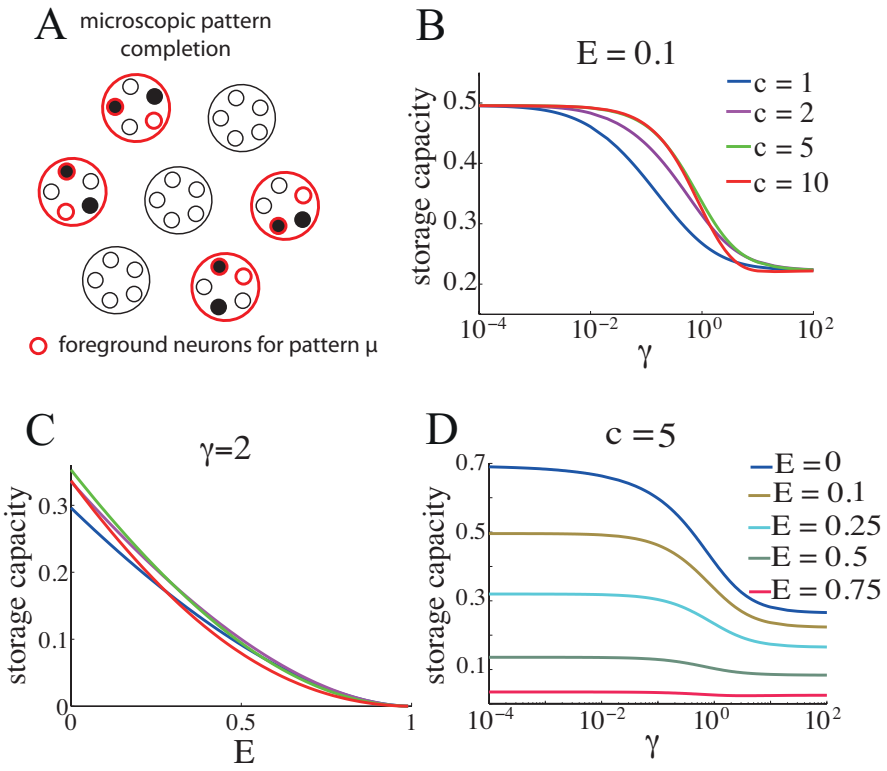

Figure 3: *Storage capacity with microscopic error-correction. A. The network is initialized at* $t = 0$ *in a state* $\vec{S}^{\mu_0}$ *which is obtained from the memory* $\vec{\Xi}^{\mu_0}$ *by randomly flipping a fraction* E *of the neurons state in active modules* m *(*$\Xi_m^{\mu_0} = 1$*). B. Storage capacity, as a function of* γ *(*$E = 0.1$*), in networks that can retrieve patterns with* $E = 0.1$*. Errors are corrected by choosing an appropriate activation threshold (see text for details). C. Same as B, but the storage capacity is plotted as a function of* E *for* $\gamma = 2$ *and various values of* c*, the number of categories in which a module is involved. D. Storage capacity as a function of* γ *for* $c = 5$ *and different values of* E*.*

where the four terms in the right hand side are the respective contribution of the correctly active neurons in the local module, the false-positive neurons in the local module, the correctly active neurons in remote modules, the false positive activated neurons in remote modules. Once errors are corrected, the larger the activation threshold we take, the more patterns we can store before background neurons are destabilized. And as before, we can take an activation threshold that tends in the large N limit towards the average of the fields on false negative neurons: $\theta \underset{N\to+\infty}{\to} fN(1-E+Eg+\frac{\gamma}{c}(1-E)+\frac{\gamma}{c}G)$. Using this activation threshold in the expression of the rate functions (65),(66) we can estimate the storage capacity of the networks when microscopic error correction is required. The result is shown in figure 3B where we plot the new storage capacity as a function of γ for different values of c and $E = 0.1$ (the optimization over the parameter β is performed as described in the previous section). In this case again, the optimal storage capacity is reached for small values of γ (short-range connections dominate), because microscopic errors correspond to small deviations around local attractors which do not require communication between modules to be corrected. Unsurprisingly, the storage capacity decreases monotonically with E, as illustrated in figure 3C.

3.2.2 Disambiguation

Now the network is initialized in a state $\vec{S}^{\mu_0} = \left(S_{i,m}^{\mu_0} = S_m^{\mu_0} s_{i,m}^{\mu_0}\right)_{i,m}$ such that $S_m^{\mu_0} = \Xi_m^{\mu_0}$, but a fraction E of the active modules are in a state that is not the correct one: $s_{i,m}^{\mu_0} = \xi_{i,m}^{\mu \neq \mu_0}$ (see figure 4A). In such a state, fields on false positive neurons have an average $fN + fN\frac{\gamma}{c}G$, where the first term describes local inputs and the second inputs from long range inputs. And fields on false negative neurons have an average $fNg + fN(\frac{\gamma}{c}(1-E)) + fN(\frac{\gamma}{c}EG)$, where again the first term describes local inputs, the second term describes long-range inputs from modules in a correct state and the third term describes long-range inputs from modules in a wrong state. In order to perform disambiguation the activation threshold has to be set between these average values:

$$1 + \frac{\gamma}{c}G < \frac{\theta}{fN} < g + \frac{\gamma}{c}\left(1 - E(1-G)\right) \tag{34}$$

Note that fields on correct foreground neurons (with average $1 + \frac{\gamma}{c}(1 - E * (1-G))$) and correct background neurons (with average $g + \frac{\gamma}{c}G$) do not constrain the choice of activation threshold to perform disambiguation as illustrated in figure 4B. Inspection of (34) at low storage ($g, G \simeq 0$) tells us that in order to find an activation threshold that performs disambiguation, the amount of long-range connections has to be large enough,

$$\gamma > \frac{c}{1-E}. \tag{35}$$

When the storage load increases ($\alpha = pf^2$ increases thus $g, G = O(1)$ increase), at fixed γ, c, the distance between the left-hand side and the right-hand side of (34) increases, and in fact it becomes easier to find an activation threshold separating the inputs to false positive and negative, as illustrated in figure 4B where we highlight the range of α where it is possible to find a relevant activation threshold. The figure shows that for these fixed

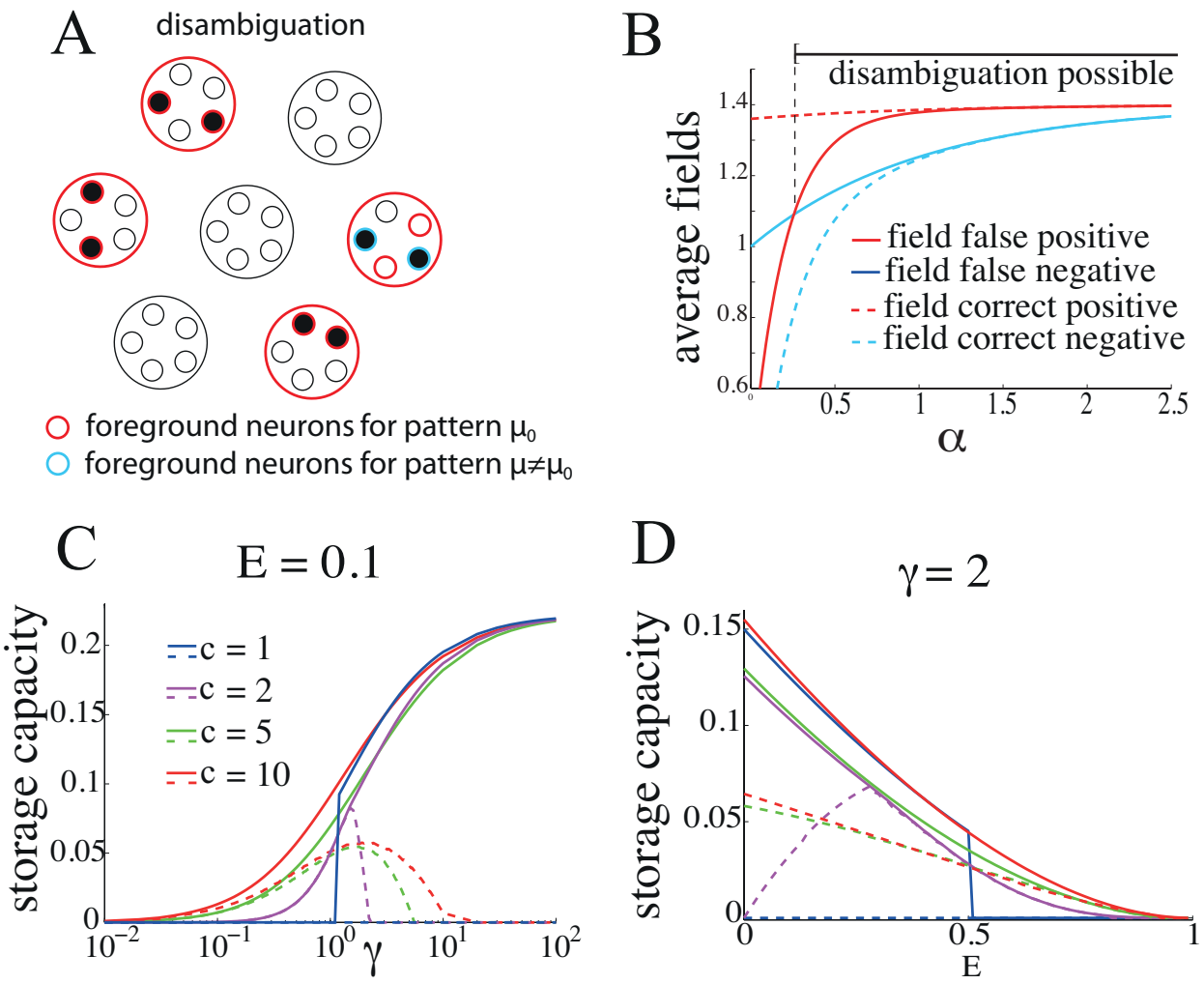

Figure 4: *Storage capacity with disambiguation. A. The network is initialized at $t = 0$ in a state $\vec{S}^{\mu_0}$ which is obtained from the memory $\vec{\Xi}^{\mu_0}$ by setting a fraction E of the active modules (a fraction FE of all the modules) in a local pattern that corresponds to another memory in the same category $\vec{\Xi}^{\mu}$ with $\mu \neq \mu_0$. B. The possibility to find an activation threshold that correct for that kind of errors depends on the amount of stored patterns $\alpha = pf^2$ that controls the variables g and G representing the amount of local and external noise. We plot the average fields onto neurons while the network state is $\vec{S}^{\mu_0}$. If the fields on correct positive and false-negative neurons are higher than the fields on correct negative and false-positive neurons, it is possible to find θ such that the network performs disambiguation. C- Quantification of storage capacity of networks that perform disambiguation as a function of γ, for $E = 0.1$. To get the full lines, we find the values of $(\beta_{max}, \alpha_{max})$ (for each value of γ) for which the storage capacity is maximal. Then at fixed β_{max}, we find the value α_{min} for which $\Delta\theta$ becomes negative (see B), inserting $(\beta_{max}, \alpha_{min})$ into the formula for storage capacity* (23) *we get the dotted lines. Thus the region of the plane between the full lines and the dotted lines delimits for which storage load the networks perform disambiguation. D- Same as C but the storage capacity is plotted as a function of E for $\gamma = 2$.*

values of β, γ, c, when $\gamma < \frac{c}{1-E}$ it is only from a minimal non zero value α_{min} that the network can perform disambiguation. The storage capacity of networks performing disambiguation is shown by the full lines in figure 4C, while the dotted lines show the storage capacity associated with the minimal amount of patterns (given by α_{min}) that need to be stored in order to find a value of θ that performs disambiguation. When γ is sufficiently large, i.e. (35) is satisfied, $\alpha_{min} = 0$. We show results for different values of c, the number of categories in which a module is involved. When the modules are active in a single category, $g = G$ and $\Delta\theta > 0$ can be satisfied if and only if (35) is satisfied, while when $c > 1$, g increases faster than G with α, and $\Delta\theta > 0$ if and only if

$$\gamma > \frac{c}{1-E} \exp\left(-\alpha(c-1)\right). \tag{36}$$

Thus the constraint (35) on γ can be relaxed, but the price to pay is that the network performs disambiguation only if a sufficient amount of patterns are stored in the network, and that the activation threshold $g + \frac{\gamma}{c}\left(1 - E(1-G)\right)$ should increase as more memories are stored. With disambiguation, the storage capacity increases monotonically with γ (figure 4C), which is the opposite of the behavior in the microscopic-error correction case. To explain why in the disambiguation case, storage capacity increases with γ, it is instructive to consider the case $E \ll 1$. Then the activation threshold that we choose to perform disambiguation is $\theta_{dis} = g + \frac{\gamma}{c}$, which we can compare to the activation threshold we choose when no error correction properties are required, $\theta_0 = 1 + \frac{\gamma}{c}$ (i.e. the average field on foreground neurons when the network is in the stored pattern). Increasing γ allows θ_{dis} to get closer to the maximal activation threshold θ_0.

In figure 4D, we show the storage capacity as a function of E for $\gamma = 2$. Note the abrupt drop in storage capacity for $c = 1$ when E reaches the point where (35) is violated.

3.2.3 Macroscopic pattern-completion

We now consider the situation in which the initial macroscopic state of the network $\vec{S}^{\mu_0}$ is not consistent with $\vec{\Xi}^{\mu_0}$ (see figure 5A):

$$S_{i,m} = S_m s_{i,m} \text{ with } S_m = (1 - X_m^f)\Xi_m^{\mu_0} + X_m^b(1 - \Xi_m^{\mu_0}) \tag{37}$$

where

$$X_m^f = \begin{cases} 1 & \text{with probability } (1-F)E \\ 0 & \text{with probability } 1-(1-F)E \end{cases} \tag{38}$$

and

$$X_m^b = \begin{cases} 1 & \text{with probability } FE \\ 0 & \text{with probability } 1-FE \end{cases} \tag{39}$$

The microscopic activity is given by

$$s_{i,m} = \begin{cases} \xi_{i,m}^{\mu(m)\neq\mu_0} & \text{if } \Xi_m^{\mu_0} = 1 \text{ and } X_m^f = 1 \\ \xi_{i,m}^{\mu_0} & \text{otherwise} \end{cases} \tag{40}$$

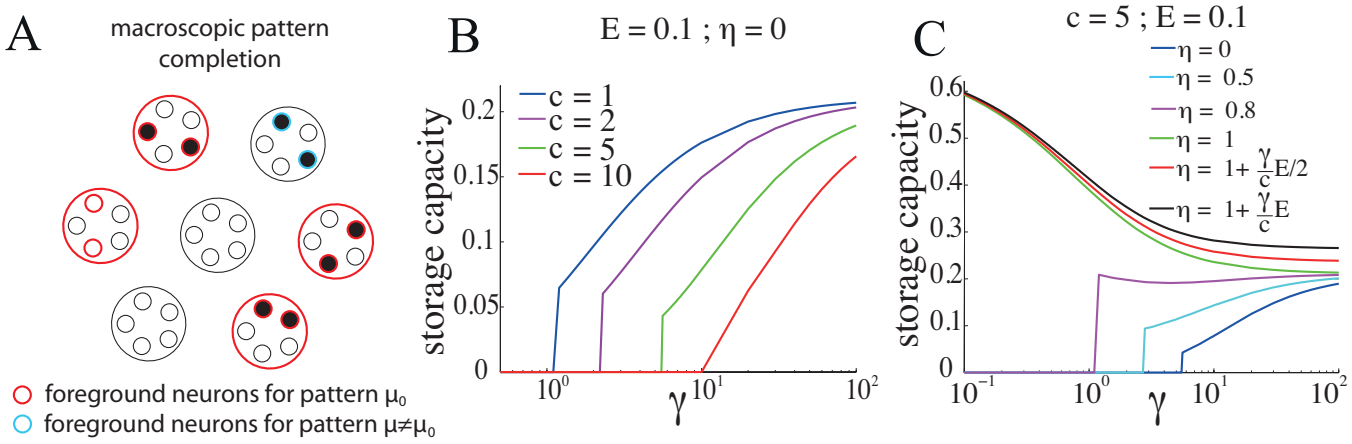

Figure 5: *Storage capacity with macroscopic pattern completion. A. The network is initialized at* $t = 0$ *in a state* $\vec{S}^{\mu_0}$ *which is obtained by turning on or off the activity of a fraction* FE *(*$E = 0.1$*) of the modules. Wrongly active modules are set in a state that corresponds to another memory* $\mu \neq \mu_0$*. B. Storage capacity as a function of* γ*. In order to perform pattern completion,* γ *has to satisfy* (43)*. C. Storage capacity with local inhibition. The constraint on* γ *can be relaxed by adding local inhibition (see* (44)*), the new constraint is given by* (46)*. For* $\eta \geq 1$*, macroscopic pattern completion do not set any constraint on* γ*.*

This differs from disambiguation where the state of modules is always correct ($S_m = \Xi_m^{\mu_0}$), here a fraction $(1-F)E$ of the FM foreground modules are silent, and a fraction FE of the $(1-F)M$ silent modules are active in a state supported by local connections as shown in figure 5A. Note the dependence of μ on m in (40) which implies that erroneously active modules do not interact with each other through long-range connections. In order to correct for this kind of error, the activation threshold has to separate fields on false positive and false negative neurons, which leads to the inequality

$$1 + \frac{\gamma}{c} G'(1-E) < \frac{\theta}{fN} < \frac{\gamma}{c}(1-E) + \frac{\gamma}{c} G' E \tag{41}$$

where G' is the fraction of co-activated synapses between two neurons taken in randomly chosen modules (as opposed to modules that are co-activated in a given category), defined in (28). G' goes to 0 in the regime we are considering, which leads to the constraint on the activation threshold

$$1 < \frac{\theta}{fN} < \frac{\gamma}{c}(1-E) \tag{42}$$

In order to find such a threshold, we should be in a parameter regime such that

$$\gamma > \frac{c}{1-E} \tag{43}$$

This is the same constraint as the one obtained for the disambiguation task at low storage load (see eq. (35)), which corresponds to the fact that currents coming from long-range

connections alone has to be able to activate silent neurons that are foreground neurons for the pattern to be retrieved (see empty red circles in figures 4-5A)

We have computed the storage capacity that can be reached by taking $\theta \to \frac{\gamma}{c}(1-E)$. Results are reported in figure 5B. From this plot, memory performance increases with γ, and the amount of long-range connections required to perform macroscopic pattern completion increases linearly with c the number of categories in which a module is involved. This is because when c increases, more pairs of modules are co-activated and the total amount of long-range connections, fixed by the constraint (9), is spread on more pairs of modules rendering the effect of long-range inputs less efficient. Note the similarity with the disambiguation task described in section 3.2.2.

3.2.4 Macroscopic pattern-completion with local inhibition

The amount of long-range connections required to perform macroscopic pattern completion can be decreased by using local inhibition. The introduction of inhibition is compensated by lowering the fixed activation threshold θ. Having a smaller fixed activation threshold, makes false-negative modules (m such that $S_m^{\mu_0}=0$ and $\Xi_m^{\mu_0}=1$) more sensitive to long-range inputs compare to the case where there is a large fixed threshold without local inhibition. Formally, the input to neurons becomes

$$h_{i,m} = \sum_{j=1}^{N} W_{ij}^{m,m} S_{j,m} - \eta \sum_{j=1}^{N} S_{j,m} + \sum_{n \neq m} \sum_{j=1}^{N} W_{ij}^{m,n} S_{j,n}, \tag{44}$$

and the constraint on θ to perform macroscopic pattern completion becomes

$$1 - \eta < \frac{\theta}{fN} < \frac{\gamma}{c}(1-E) \tag{45}$$

In figure 5C, we show, for $c=1$ and $E=0.1$, how the storage capacity evolves as a function of γ. The minimal amount of long-range connections required to perform macroscopic pattern completion can be dramatically reduced by increasing the strength of the local inhibition. The constraint on parameters is now

$$\gamma > (1-\eta)\frac{c}{1-E} \tag{46}$$

Storage capacity can decrease or increase with γ, depending on η, this reflects the two competing effects of the decrease of storage capacity with γ (cf figure 2A) in the absence of error correction properties, and the need to use long-range connections to perform error correction. For low inhibition, capacity increases with γ; for high inhibition, the opposite occurs. There is an intermediate range of values of η for which there exists an optimal value of γ (see the case $\eta = 0.8$ in Fig. 5C). The best performance are reached when the inhibition is taken as $\eta = 1 + \frac{\gamma}{c}E$, the value for which, once the pattern is retrieved, the effective threshold is close to the field on foreground neurons: $\theta + \eta \sum_{j=1}^{N} S_{j,m} \to 1 + \frac{\gamma}{c}$. For larger inhibition the field on foreground neurons is not strong enough to overcome inhibition and no patterns can be retrieved.

Note that for errors considered in the previous sections, replacing part of the threshold by inhibition does not improve pattern completion abilities. This is because both inputs on false-positive and false-negative neurons are affected the same way by inhibition since both of these types of neurons belong to modules with fN active neurons. Formally a term $-\eta$ is added to both sides of inequalities (33) and (34), which therefore remain unchanged.

3.2.5 Storage capacity of networks performing all forms of error correction

In the previous sections, the storage capacity was optimized with respect to the parameter β for each error-correction task separately. To synthesize the results of the three previous sections, here we set β to its value that optimizes I for all the three different tasks. Results are summarized in figure 6, where we plot the capacity (full lines) that can be reached while a network with inhibition performs microscopic pattern completion, disambiguation and macroscopic pattern completion, with the same error correction level E for all three types of error correction. In this figure we also show dotted lines (as in 4C,D) below which the storage load is not high enough to find an activation threshold satisfying (34) and therefore the network is unable to perform disambiguation. On panel A, we show how the maximal storage capacity behaves as a function of γ for $E = 0.1$ and $\eta = 1$ (the value of β being chosen to maximize storage capacity). Storage capacity increases with γ, as well as the range of storage load on which the model performs well (see dotted lines). The larger the value of c, the better performance is for values of $\gamma < \frac{c}{1-E}$. Although note that for values of $\gamma < \frac{c}{1-E}$, networks have to be loaded with a sufficient number of memories to perform well. If one considers values of $\gamma > \frac{1}{1-E}$, the value of maximal storage capacity is similar for all values of c.

In panel B, we study the dependence on E for $\gamma = 2$ and $\eta = 1$, as expected storage capacity decreases with the amount of errors to be corrected. We then inspect the effect of changing the amount of local inhibition η, which has been introduced to relax the constraint (43) on γ to perform macroscopic pattern completion. This is shown for $E = 0.1$ and $\gamma = 2$ (panel C) and $\gamma = 1$ (panel D). The more c increases, the closer η has to be to 1 for the network to have a finite storage capacity. For instance for $c = 10$, it is only between $\eta = 0.8$ and the maximal value $1 + \frac{\gamma}{c}E$ that storage capacity is non-zero. This is because the ability to perform macroscopic pattern completion requires $\gamma > (1-\eta)\frac{c}{1-E}$.

In summary, the overall performance of the network is mainly constrained by the disambiguation task that imposes $\gamma > \frac{1}{1-E}$ for $c = 1$. For $c \geq 2$, there are no strong constraints on γ (although the networks have larger storage capacity at larger γ) but disambiguation is possible only in a limited range of storage loads (dotted lines). For the particular value $\eta = 1$ used in figure 6A, the storage capacity of the network performing the three error-correction tasks is the storage capacity obtained when only disambiguation is required (compare figures 6A and 4C). The requirements of macroscopic pattern completion do not put a strong constraint on the values of γ, provided that local inhibition can be tuned such that $\gamma > (1-\eta)\frac{c}{1-E}$. Microscopic pattern completion does not particularly constrain the parameters of the network: even though storage capacity decreases with γ when only microscopic pattern-completion is required, it is the disambiguation task that imposes the increase of storage capacity with γ. However, if the amount of errors to be corrected for

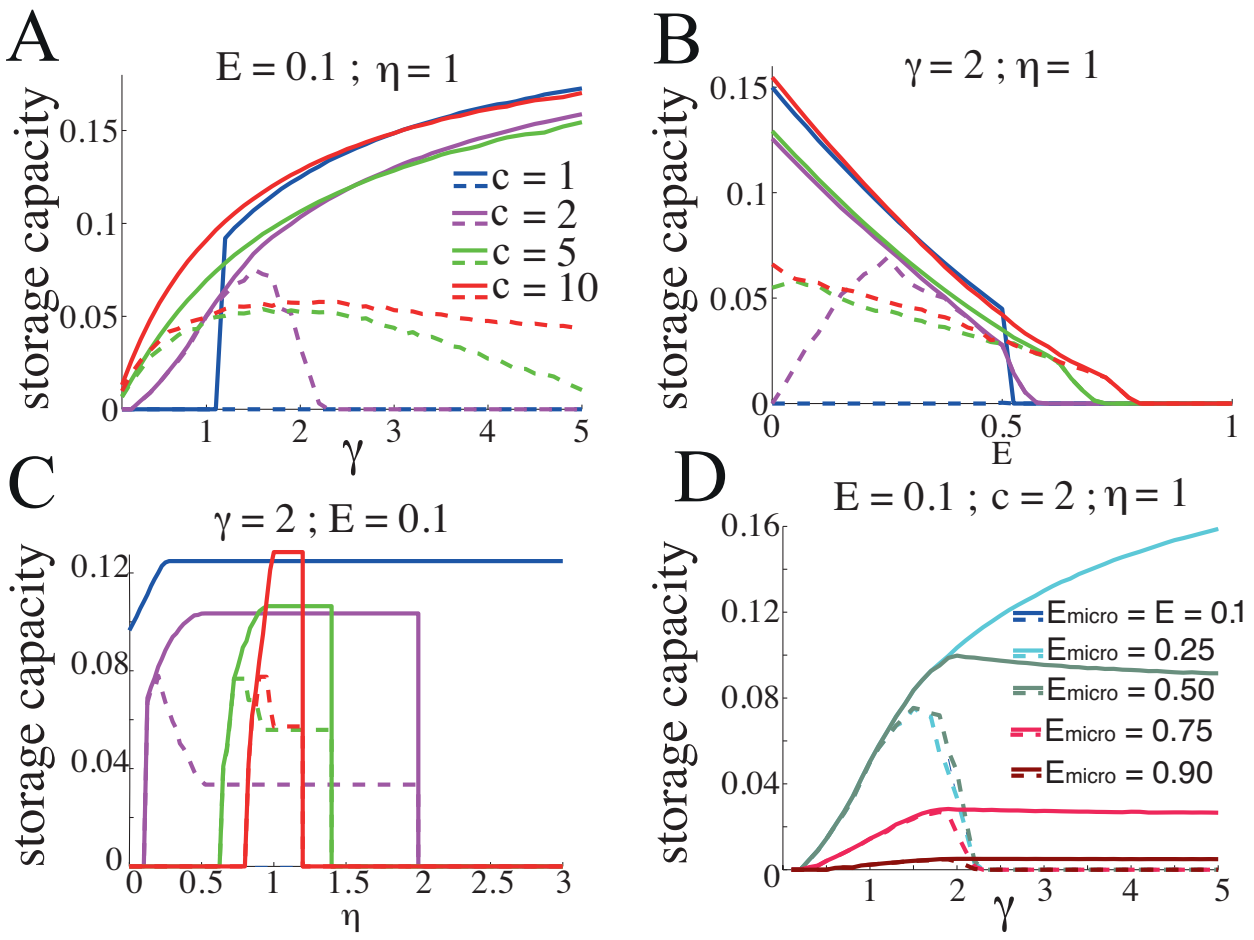

Figure 6: *Storage capacity of networks performing all forms of error-correction. A. Full lines show storage capacity as a function of γ, and dotted lines delimit the range of storage loads (parameter α) in which the network is able to perform disambiguation. B. Storage capacity as a function of E, the parameter quantifying the amount of error to correct, on the x-axis. C Storage capacity as a function of η, the parameter quantifying the amount of local inhibition for $\gamma = 2$ and $E = 0.1$. Different curves correspond to different values of c in panels A to C. D Storage capacity when the amount of errors in the microscopic error-correction task E_{micro} becomes larger than the amount of errors in the disambiguation and macroscopic error-correction tasks.*

the microscopic pattern-completion task becomes sufficiently larger than the amounts for the disambiguation and macroscopic pattern-completion tasks, the optimal storage capacity can be reached at an intermediate value of γ. This can be seen in figure 6D where we plot the storage capacity as a function of γ for an amount of error $E = 0.1$ for the disambiguation and macroscopic pattern completion tasks, and an amount $E_{micro} > E$ of errors for the microscopic error-correction task ($\eta = 1$, $c = 2$).

4 Discussion

4.1 Summary of results

In this article we have explored the parameter space of a class of multi-modular attractor neural networks and quantified their memory performance in this space. In order to perform this detailed characterization, we have considered large networks where the number of modules and the number of neurons per module is large ($M, N \to +\infty$), with the constraint $N \gg M$. We have first focused on the maximal storage capacity of these networks, i.e. the maximal number of patterns that can be stored without requiring error-correction properties. Our 'Willshaw type' networks can behave as an efficient memory storage device in the sense that their storage capacity is of order 1. In the limits of sparse patterns of activity we have considered, $F \to 0$ (sparseness at the level of module activation) and $f \propto \frac{\ln N}{N}$ (sparseness at the level of single neuron activation), this amounts to state that the number of stored patterns scales as $P_{max} \propto \frac{1}{F}\left(\frac{N}{\ln N}\right)^2$. We have shown that in the case where $F \propto \frac{1}{M}$, storage capacity also remains finite (see figure 2C), meaning that the number of stored patterns can be proportional to the number of modules in the network. In order to quantify how modules are interconnected, we have introduced a parameter γ, the ratio between the numbers of synaptic contacts of long-range and short-range origins a neuron receives. We have found that when this parameter γ is varied from small to large values, the storage capacity of modular networks varies between the one of a fully connected network and the one of a highly diluted network, which corresponds to the fact that when $\gamma \to 0$ patterns of activity are stored exclusively on short-range connections that are part of fully connected networks, and when $\gamma \to +\infty$ patterns of activity are stored exclusively on long-range connections that are highly diluted.

One particularity of our model resides in the structure of the set of stored patterns, namely that P stored patterns are split in $\mathscr{P} \propto \frac{1}{F}$ categories. Two patterns belonging to the same category have the same FM active modules. From the point of view of a single module, we have imposed it is activated in $c = \text{int}\left[\mathscr{P}F\right]$ categories. Interestingly, for each value of γ there exists a value of c that optimizes storage capacity (e.g. $c = 6$ at $\gamma = 1$, see figure 2B). In a model without categories (for each pattern the identity of active modules is chosen randomly and independently of the other stored patterns), storage capacity or error correction properties of the networks are similar to the one described here (Dubreuil, 2014), one difference is that the modular sparseness F can not be taken to scale as low as $\frac{1}{M}$, and the number of stored patterns can not scale linearly with the number of modules in the network. Such a pattern structure is a first step towards modeling the fact that semantically similar objects elicit similar cortical patterns of activity when observed

at the scale of $\backsim$ 1mm (Huth et al., 2012). It would be interesting to study different pattern structures, as e.g. a structure in which similarity across patterns varies continuously rather than in a discrete manner.

We have also quantified the storage capacity of modular networks with error-correction (or associative) properties. We have considered three canonical types of errors, namely microscopic pattern completion, disambiguation and macroscopic pattern completion (see figures 3A, 4A and 5A). We have seen that if the network is required to perform well on all these error-correction tasks on a **full-range of storage loads**, the constraint $\gamma > \frac{c}{1-E}$ needs to be satisfied where E quantifies the fraction of errors to be corrected. In qualitative terms, it implies to have more long-range contacts than short-range ones, as $\frac{c}{1-E} > 1$. We have also seen that for values of γ smaller than $\frac{c}{1-E}$, if $c \geq 2$, and if the amount of local inhibition is high enough to satisfy $\gamma > (1-\eta)\frac{c}{1-E}$ the networks can reach a reasonable storage capacity. The ability to correct for the three kinds of errors in this regime of low γ is paid by the existence of a region of storage load where the networks can not perform disambiguation (represented by dotted lines in the figures), moreover in the region of storage load where it can, the activation threshold has to be adjusted according to the storage load. Note that the ability to perform disambiguation and macroscopic pattern completion strongly depends on the ratio $\frac{\gamma}{c}$, which scales the strength of the recurrent input from long range origin when the network is in one of the stored patterns (see (24) for instance). The inverse dependence on c comes from the fact that if c increases, the number of pairs of modules that get connected during the learning phase increases, and the amount of long-range connections between two given modules is effectively reduced.

The storage capacity of modular networks remain below the storage capacities of networks with no structure: a fully-connected Willshaw network has a capacity of 0.69 (Willshaw et al., 1969; Knoblauch et al., 2010; Dubreuil et al., 2014), while modular networks studied here have a lower capacity. Thus, in terms of storage capacity, it would be more efficient to store patterns of activity in multiple distinct unstructured networks. However, such networks would be unable to perform error correction at the macroscopic level.

4.2 Comparison with previous models

Modular networks have also been studied by other authors. O'Kane and Treves (1992) have studied a model with threshold linear units and synaptic weights that can take a continuum of values. They also found an storage capacity of order 1. Similarly to our study, they found that when γ is varied from 0 to large values, the storage capacity varies from the one of a fully-connected networks to the one of a highly diluted networks. In their study they found that together with stored patterns, other states ('memory-glass states'), are minima of the energy function describing their network. These states correspond to modules sustaining one of their local pattern, with the global combination of modules activity inconsistent with the stored patterns. In our model such states are destabilized when the activation threshold is high enough, such that only states that are correct combination of local activities are stable. We believe that these states could also be destabilized in their model, by taking into account an appropriate activation threshold. In a subsequent study, Mari and Treves (1998) introduced a parameter controlling the macroscopic sparseness and showed that it allowed to store a number of patterns that

increases with the number of modules in the network. In this model, they also introduced a bias in the statistics of the patterns to store, such that pairs of modules between which long-range connections exist tend to be co-activated in a number of global patterns that is larger than expected by chance. This improved the model previously studied by O'Kane and Treves (1992) by reducing the size of the region of storage load in which stored patterns are stable together with memory-glass states. Our model, where patterns are organized in categories, has a similar feature since two connected modules (that are co-activated in at least one category) are likely to be involved in more patterns of activity than a pair of modules taken at random. However as our model does not exhibit memory glass-state, such a feature is not crucial for its performance and in fact a model with non-categorized patterns can have similar behavior both in terms of storage capacity and associative properties (Dubreuil, 2014).

In a following study, Mari (2004) focused on the dynamics of pattern retrieval and proposed an oscillatory mechanism that allowed to get rid of these memory-glass states. It should be mentioned that in all these models (Mari, 2004; O'Kane and Treves, 1992) the same local pattern is used in μ several global patterns, while in our case one local pattern is used in only one global pattern. They have found (O'Kane and Treves, 1992) that the number of patterns stored in the network increases linearly with μ, while the storage capacity decreases with μ.

Johansson and Lansner (2007) explored the storage capacity of a modular attractor network with three levels of spatial organization (neurons, mini-columns, hyper-columns) while in our model we have only two levels of spatial organization (neurons and modules). They focused on finite-size networks and studied the storage capacity using a signal to noise ratio analysis. Doing so they were able to derive constraints on the number of neurons composing a mini-column and on the number of mini-columns composing an hyper-column for the network to have a reasonable storage capacity. In this work, long-range connectivity is not patchy in the sense that connections from two pre-synaptic neurons in the same column do not necessarily end up in the same ensemble of modules. In a subsequent work Meli and Lansner (2013) added the constraint of patchy connectivity and found similar storage capacities. This result also holds for our model, it is shown elsewhere that a similar storage capacity can be obtained in a model with non-patchy diluted long-range connections (Dubreuil, 2014).

While it is difficult to give quantitative comparisons of the storage capacity of our model with others given the specificities of each model and the differences in analysis, we arrive at the similar conclusion than Mari and Treves (1998); Mari (2004); Johansson and Lansner (2007) that modular attractor networks have reasonable storage performance. An interesting feature of our model is the absence of spurious states which spares us from introducing additional mechanisms to destabilize these states. The main novelty of our study is the detailed study of the error-correction abilities of the modular network, which allowed us to quantitatively discuss the important parameters (ratio of long/short range connections, local inhibition, number of categories) that allow large storage capacities together with associative properties.

We have defined the storage capacity (20) as the total number of stored patterns, multiplied by the entropy of the distribution of patterns, divided by the amount of ressources used for storage. This measure can be interpreted as the information stored in the system

in bits/synapse. It is a generalization, for modular networks, of the capacity measure used in studies of fully-connected associative memories with binary neurons (see e.g. Nadal (1991); Knoblauch et al. (2010)). However, identifying this measure as an 'information' leads to the following paradox: In the standard Willshaw autoassociative network, the storage capacity is 0.69. However, since the connectivity matrix is symmetric, there are only $N(N-1)/2$ independent binary elements that can be used to store the patterns. If the storage capacity is identified as information stored per independent binary plastic element, this leads to a stored information of 1.39 bits per independent binary storage device, which violates the bound from information theory that states that it should be impossible to store more than 1 bit of information per binary storage element. For this reason we have refrained from using the word 'information' which was used in previous studies (Nadal, 1991; Knoblauch et al., 2010).

4.3 Extension to more realistic models

Our model and analysis could be modified to include more realistic features. For instance studies of storage capacity of modular networks have been carried in the limit where M the number of modules and N the number of neurons are infinite, or in finite networks of small sizes. In a previous study, we have quantified finite-size effects in fully connected networks (Dubreuil et al., 2014). The leading order corrections due to finite size effects scale as $\frac{\ln(\ln N)}{\ln N}$ and lead to a decrease in storage capacity of a factor around 3 for fixed size patterns as the one considered here (in each pattern exactly $FMfN$ neurons are active). For modular networks, the first order correction term is proportional to $\frac{\ln(\ln N)}{\ln N} + \frac{\ln(FM)}{\ln N}$, which decays extremely slowly as N and FM become large (it is ~ 0.6 for $N = 10^5$ and $FM = 100$). We thus expect a drop in capacity of the same order as the one observed for fully connected networks. Besides this drop in capacity, we expect that for finite size networks, the regions of parameters in which networks perform error-correction remain the same. Indeed, error correction is possible if the stability of patterns is achieved by a sufficient amount of long-range inputs compared to short-range inputs, and the ratio of these two quantities is quantified by $\frac{\gamma}{c}$ which is independent of network size.

All the patterns we store have the same number of active neurons spread in the same number of modules, one could imagine storing an ensemble of patterns where the numbers of active neurons/modules fluctuate from pattern to pattern. For such patterns to be fixed points of the dynamics, the activation threshold has to be lowered to ensure the stability of patterns with a small number of active neurons. In finite networks, lowering the threshold decreases the number of patterns that can be stored (Dubreuil et al., 2014). Another quantity which is susceptible to fluctuate is c, the number of categories encoded per module. If we would let c fluctuate, the storage capacity would be limited by modules that encode the largest number of categories since these modules are more densely connected, which increases the chance to activate background neurons. Such limitation would be attenuated for large values of γ where storage is supported mainly by long-range connections.

The learning rule used in this work requires that patterns are learned off-line, and does not allow the networks to have palimpsest properties, i.e. the ability to continuously

learn new patterns (at the expense of erasing old ones). A popular learning rule to learn patterns on-line in networks of binary neurons and binary synapses is the one proposed by Amit and Fusi (1994). For this rule plasticity is activity dependent on a stochastic manner, and the presentation of a pattern leads not only to the activation of synapses between pairs of 'active-active' neurons (LTP), but also to the inactivation of synapses between pairs of 'silent-active' neurons (LTD). If LTP and LTD are well balanced, the network is able to continuously learn new patterns by erasing old ones. Extending our calculations to such a rule should be straightforward, similar to what has been already done for local unstructured networks (Dubreuil et al., 2014).

Our networks are composed of binary neurons. A major challenge is to understand whether networks made of more realistic neurons could have similar performance. Networks of fully-connected spiking neurons in the balanced regime can not stabilize patterns of activity with coding levels smaller than $f \propto \frac{1}{\sqrt{N}}$, because the signal on foreground neurons would be wiped out by the activity of background ones (Brunel, 2003; van Vreeswijk and Sompolinsky, 2003). For modular networks, we expect a similar constraint, $f \propto \frac{1}{\sqrt{N}}$ since each neuron receives inputs from a number of neurons that scales with N (because of the dilution of long-range connections). Coding levels obeying this scaling seem too large to have reasonable storage capacity with the binary synapses used in our model, as it would require $f \propto \frac{\ln N}{N}$ (Willshaw et al., 1969). However in our previous study on fully-connected networks, we have seen that for finite networks of realistic sizes (e.g. $N \simeq 10^4$), the coding levels optimizing storage capacity are not far from $f = \frac{1}{\sqrt{N}}$. We thus expect that modular network of spiking neurons with binary synapses could also have reasonable storage capacities.

4.4 Relationship to experimental data

In attractor network models, patterns of activity are imprinted in the synaptic matrix of the network, and are retrieved under the form of a neural state of persistent activity. These basic mechanisms are in principle implementable in cortical circuits, as cortical synapses have been shown to be plastic in an activity dependent manner (Markram et al., 1997; Sjöström et al., 2001) and are thus susceptible to sustain long-term storage of patterns of activity. Furthermore, the phenomenon of persistent activity has been observed in many cortical areas during working memory tasks such as delay-match to sample tasks (see e.g. Fuster (1995)).

A fully-connected module in our model can be considered as approximating a cortical patch as the one observed in visual and pre-frontal cortices that we described in more details in the introduction (DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Bosking et al., 1997; Pucak et al., 1996) - full connectivity is only an approximation in the sense that cortical patches have a size of the order of $1mm^2$, and it is known that for such large networks, the probability that two neurons touch each other depends on the distance between them (Holmgren et al., 2003; Perin et al., 2011). In our model, modules are connected to each other via diluted long-range connections, whose amount is controlled by a parameter γ. γ is the ratio between the number of contacts whose pre-synaptic neurons are outside of the module to which the post-synaptic neurons belongs to and the number of contacts within

this module. Anatomical studies indicate that this ratio is of order one: Braitenberg and Schütz (1991) estimated, in rodents, that short and long-range connections come in approximately similar numbers, while Stepanyants et al. (2009) estimated $\gamma \simeq 3$. Besides being as numerous as short-range connections, long-range connections are patchy in our model, in the sense that the long-range connectivity originating from one module target only a subset of all the other modules. This is in agreement with the above mentioned studies of cortical patches (DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Bosking et al., 1997; Pucak et al., 1996).

Another assumption regarding connectivity in our model is the choice of binary synapses. Whether real synapses are best described by binary variables, discrete variables with a large number of states or continuous variables, is still unresolved (Petersen et al., 1998; Montgomery and Madison, 2004; O'Connor et al., 2005; Enoki et al., 2009; Loewenstein et al., 2011).

In our model, the structure of the patterns of activity strictly corresponds to the anatomical structure of the networks, i.e. in a given pattern each module is either totally silent or fN neurons are active. fMRI experiments allow to study how neural representations of objects are distributed. Huth et al. (2012) found that visual presentations of objects elicit, in human observers, a pattern of activity spanning the entire cortex. In prefrontal cortex, these patterns consist of changes in activity of small pieces of cortex of a size of the order of a millimeter squared. In our model we assume that patches selective to a given object correspond to networks of connected cortical patches as the ones described by Pucak et al. (1996). This is not such a bold assumption as it has been shown that neurons belonging to connected patches in visual cortex tend to have similar selectivity (Bosking et al., 1997; Buzas et al., 2001).

Moreover, stored memories are split in categories, such that patterns of activity coding for memories belonging to the same category involve the same set of active modules. This is a first step to account for the observation that objects that are semantically close to each other are represented similarly on the cortical surface, while the neural representations of semantically far apart objects are dissimilar (Huth et al., 2012). In our model, from one category to another, the set of active modules are drawn independently. Experimental data rather suggests a smooth transition, in terms of neural representations, between categories that are semantically close to each others (Huth et al., 2012). This could be modeled by taking patterns with multiple levels of hierarchy, and it would be interesting to investigate how such patterns can be stored in networks with realistic connectivity constraints.

Our results show that the performance of a modular network as an auto-associative memory device is strongly determined by the parameters c, γ and η. In principle if these parameters could be estimated for a given cortical network, then it should be possible to determine whether it is well suited to store objects it represents via attractor dynamics. Although this seems a difficult task for a randomly taken set of cortical patches, this could be done for specific sets of patches, like the well identified network of face patches (Tsao et al., 2003).

5 Appendix

Estimating the storage capacity requires to compute the distributions of the inputs on foreground and background neurons. Because neural activity and synapses are binary, these inputs are sums of binary random variables. We first present general results about such sums, that will be used later to compute the distributions of inputs to neurons.

We consider a random variable h

$$h = \sum_{k=1}^{K} X_k \tag{47}$$

where the X_k are independent binary random variables described by a parameter q:

$$X_k = \begin{cases} 1 & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases} \tag{48}$$

The sum h is then distributed according to a binomial distribution

$$P(h=S) = \binom{K}{S} q^S (1-q)^{K-S} \tag{49}$$

Note that to get this binomial distribution, we have to assume the X_k's are independent. In our case, this means that two synapses on the same neuron $W_{ij_1}^{m,n}$ and $W_{ij_2}^{m,n}$ are treated as independent variables. This is a reasonable assumption to make as we have

$$\mathbb{P}(W_{ij_1}^{m,m}=1) = 1-(1-f^2)^{pc} \tag{50}$$

and

$$\mathbb{P}(W_{ij_1}^{m,m}=1 | W_{ij_2}^{m,m}=1) = 1-(1-f^2)^{pc-1}(1-f) \underset{f\to 0, pc\to\infty}{\simeq} \mathbb{P}(W_{ij_1}^{m,m}=1) \tag{51}$$

and similarly for long-range connections $\mathbb{P}(W_{ij_1}^{m,n}=1|W_{ij_2}^{m,n}=1) \underset{f\to 0, pc\to\infty}{\simeq} \mathbb{P}(W_{ij_1}^{m,n}=1)$.

We will consider cases in which K and S are large. We can then use Stirling formula to express the binomial coefficients and write

$$P(h=S) = \exp\left(-K\Phi\left(\frac{S}{K}, q\right) + o(K)\right) \tag{52}$$

with

$$\Phi = \Phi^{fc}\left(\frac{S}{K}, q\right) = \frac{S}{K}\ln\left(\frac{S/K}{q}\right) + \left(1-\frac{S}{K}\right)\ln\left(\frac{1-S/K}{1-q}\right) \tag{53}$$

We use the superscript fc as this expression will be mainly used to describe fully connected sub-networks. For diluted enough networks, we will have $q, \frac{S}{K} \ll 1$, it is then useful to introduce

$$\Phi = \Phi^{dc}\left(\frac{S}{K}, q\right) = \frac{S}{K}\ln\left(\frac{S/K}{q}\right) - \frac{S}{K} + q \tag{54}$$

In our networks, when testing the stability of a given pattern $\vec{\Xi}^{\mu_0}$ it is useful to separate the total input into a local part and an external part. The local part is described by a couple (K, q), where $K = fN$ is the number of neurons active in a given local network, and $q = 1$ or g depending on whether we are considering the input onto a foreground or a background neuron. The external part can also be described by a couple (K, q) with $K = F(M-1)fN$ or $K = F(M-1)(1-f)N$ and $q = \frac{D}{N}$ or $\frac{D}{N}G$ for foreground or background neurons.

The distribution of the total input on a neuron can be written

$$\mathbb{P}(h_{i,m} = S) = \sum_{S_l, S_e / S_l + S_e = S} \mathbb{P}_l(S_l)\mathbb{P}_e(S_e) \tag{55}$$

To compute it, we first need to express the distribution of the inputs generated by the local module and the distribution of the inputs generated by the other modules. In the asymptotic limits we consider, this sum will be dominated by the most probable term of the sum, we will thus need to find the couple (S_l, S_e) that maximizes $\mathbb{P}_l(S_l)\mathbb{P}_e(S_e)$.

5.1 Distribution of inputs and probability of no-error in multi-modular network

5.1.1 Case $FM \to +\infty$

We apply the method sketched above, first to compute the distribution of inputs on foreground neurons, and then on background neurons.

Foreground neurons - The distribution of local inputs is a delta function $\mathbb{P}_l(S_l) = \delta(S_l - fN)$ as exactly fN neurons are active in each module in each pattern, and because of the fact that each module is a fully-connected network. The external component is the sum of the activity in each of the other $FM - 1 \simeq FM$ active modules when their states coincides with the pattern $\vec{\Xi}^{\mu_0}$ we are trying to retrieve. Given the above results on sum of binary variables, it writes

$$\begin{aligned}
\mathbb{P}_e(S_e) &= \binom{FMfN}{S_e}\left(\frac{D}{N}\right)^{S_e}\left(1 - \frac{D}{N}\right)^{FMfN - S_e} \\
&= \exp\left[-fN\Phi^{dc}\left(s_e, \frac{\gamma}{c}\right)\right]
\end{aligned} \tag{56}$$

with $s_e = \frac{S_e}{fN}$.

The total input on foreground neurons is then

$$\mathbb{P}(h_{i,m} = S = fN + S_e | \Xi_{i,m}^{\mu_0} = 1) = \exp\left[-fN\Phi^{dc}\left(\frac{S_e}{fN}, \frac{\gamma}{c}\right) + o(fN)\right] \tag{57}$$

Background neurons in active modules - The local input now fluctuates because inputs are mediated by synapses that have been potentiated during the presentation of randomly drawn patterns $\vec{\Xi}^{\mu\neq\mu_0}$. It is distributed according to

$$\begin{aligned} P_l(S_l) &= \binom{fN}{S_l} g^{S_l}(1-g)^{fN-S_l} \\ &= \exp\left[-fN\Phi^{fc}\left(\frac{S_l}{fN}, g\right) + o(fN)\right] \end{aligned} \tag{58}$$

where g is defined in eq. (25). Similarly, the external part of the input is distributed according to

$$\begin{aligned} P_e(S_e) &= \binom{FMfN}{S_e} \left(\frac{D}{N}G\right)^{S_e} \left(1-\frac{D}{N}G\right)^{FMfN-S_e} \\ &= \exp\left[-fN\Phi^{dc}\left(\frac{S_e}{fN}, \frac{\gamma}{c}G\right) + o(fN)\right] \end{aligned} \tag{59}$$

The distribution of the total input is now written

$$\begin{aligned} &\mathbb{P}(h_{i,m} = S|\Xi^{\mu_0}_{i,m} = 0, \Xi^{\mu_0}_m = 1) = \\ &\sum_{S_l, S_e/S_l+S_e=S} \exp\left[-fN\left(\Phi^{fc}\left(\frac{S_l}{fN}, g\right) + \Phi^{dc}\left(\frac{S_e}{fN}, \frac{\gamma}{c}G\right)\right) + o(fN)\right] \\ = &\exp\left[-fN\left(\Phi^{fc}\left(s_l^*, g\right) + \Phi^{dc}\left(s_e^*, \frac{\gamma}{c}G\right)\right) + o(fN)\right] \end{aligned} \tag{60}$$

where $s_l^* = \frac{S_l^*}{fN}$ and $s_e^* = \frac{S_e^*}{fN} = s - s_l^*$ (where $s = \frac{S}{fN}$) are given by the condition

$$\frac{\partial\left(\Phi^{fc}(s1, g) + \Phi^{dc}(s - s_l, \gamma FG)\right)}{\partial s_l}(s_l^*) = 0 \tag{61}$$

Solving this equations yields

$$\begin{aligned} s_l^* &= \frac{1}{2}\left(1 + s + \frac{(1-g)\frac{\gamma}{c}G}{g}\right) - \frac{1}{2}\sqrt{\left(1 + s + \frac{(1-g)\frac{\gamma}{c}G}{g}\right)^2 - 4s} \\ s_e^* &= -\frac{1}{2}\left(1 - s + \frac{(1-g)\frac{\gamma}{c}G}{g}\right) + \frac{1}{2}\sqrt{\left(1 - s + \frac{(1-g)\frac{\gamma}{c}G}{g}\right)^2 + 4\frac{(1-g)\frac{\gamma}{c}G}{g}s} \end{aligned} \tag{62}$$

Background neurons in silent modules - There is only long-range inputs in this case and the fraction of activated long-range synapses $G' \simeq \alpha c F$ is given by (28),

$$\begin{aligned} \mathbb{P}(h_{i,m} = S|\Xi^{\mu_0}_{i,m} = 0, \Xi^{\mu_0}_m = 0) &= \binom{FMfN}{S} \left(\frac{D}{N}G'\right)^{S_e} \left(1-\frac{D}{N}G'\right)^{FMfN-S} \\ &= \exp\left[-fN\Phi^{dc}\left(\frac{S}{fN}, \frac{\gamma}{c}G'\right) + o(fN)\right] \end{aligned} \tag{63}$$

Probability of no errors - We have derived the expressions for the distribution of inputs to both foreground and background neurons. In order to compute the probability that there is no error $\mathbb{P}_{ne}$ in the retrieval of pattern $\vec{\Xi}^{\mu_0}$, we have to estimate the probability that the inputs are above or below threshold, as written in the main text in equations (14). To do so we first note that

$$\mathbb{P}(h_{i,m} \geq \theta f N | \Xi_{i,m}^{\mu_0}) = \mathbb{P}(h_{i,m} = \theta f N | \Xi_{i,m}^{\mu_0}) \sum_{s \geq \theta} \frac{\mathbb{P}(h_{i,m} = s f N | \Xi_{i,m}^{\mu_0})}{\mathbb{P}(h_{i,m} = \theta f N | \Xi_{i,m}^{\mu_0})} \tag{64}$$

where the '$\sum$' term will not contribute to the final expression of $\mathbb{P}_{ne}$ in the large N limit, as has been shown in Dubreuil et al. (2014). In practice we thus replace the probability to be above threshold by the probability to be at threshold. We can apply the same reasoning for the probability to be above the activation threshold for background neurons. We now have all the elements to express Φ^f, Φ^b and $\Phi^{b'}$ in formulas (14):

$$\Phi^f = \Phi^{dc}\left(\theta - 1, \frac{\gamma}{c}\right) \tag{65}$$

$$\Phi^b = \Phi^{fc}\left(s_l^*(\theta), g\right) + \Phi^{dc}\left(s_e^*(\theta), \frac{\gamma}{c} G\right) \tag{66}$$

and

$$\begin{aligned} \Phi^{b'} &= \Phi^{dc}\left(\theta, \frac{\gamma}{c} G'\right) \\ &\underset{F \to 0}{\simeq} \theta \log(1/F) \end{aligned} \tag{67}$$

5.1.2 Case $FM = O(1)$

In this case the microscopic dilution term $\frac{D}{N}$ is finite and we have to use Φ^{fc} instead of Φ^{dc} to describe the external inputs. Following the same reasoning as above, the rate functions are

$$\Phi^f = (\theta - 1) \ln\left(\frac{\theta - 1}{FM\frac{D}{N}}\right) + (FM - (\theta - 1)) \ln\left(\frac{1 - (\theta - 1)/FM}{1 - D/N}\right) \tag{68}$$

and

$$\Phi^b = \Phi^{fc}\left(s_l^*(\theta), g\right) + s_e^*(\theta) \ln\left(\frac{s_e^*(\theta)}{FM\frac{D}{N}G}\right) + (FM - s_e^*(\theta)) \ln\left(\frac{1 - s_e^*(\theta)/FM}{1 - \frac{D}{N}G}\right) \tag{69}$$

with s_l^* and s_e^* given by

$$
\begin{aligned}
s_l^* &= \frac{\lambda(FM-s)+1+s}{2(1-\lambda)} - \frac{1}{2}\sqrt{\left(\frac{\lambda(FM-s)+1+s}{1-\lambda}\right)^2 - 4\frac{s}{1-\lambda}} \\
s_e^* &= \frac{-\lambda(FM+s)+s-1}{2(1-\lambda)} + \frac{1}{2}\sqrt{\left(\frac{-\lambda(FM+s)+s-1}{(1-\lambda)}\right)^2 + 4\frac{\lambda FMs}{1-\lambda}}
\end{aligned}
\tag{70}
$$

with

$$
\lambda = \frac{\gamma}{cFM}\frac{G(1-g)}{(1-G\frac{\gamma}{cFM})g} \tag{71}
$$

The expression for $\Phi^{b'}$ remains unchanged as connectivity between two randomly taken modules is effectively highly diluted.

5.2 Capacity calculation for a single module with diluted connectivity

Here we focus on a model made of a single module with N neurons whose dynamics obey equation (1), where we store P patterns $\vec{\xi}^{\mu}$ with coding level f ($\sum_i \xi_i^\mu = fN$). As for modular networks we focus on the limits $N \to +\infty$, $f = \beta\frac{\ln N}{N}$ and $P = \frac{\alpha}{f^2}$ with $\alpha, \beta = O(1)$. Patterns are stored on a diluted connectivity matrix, such that at the end of the learning phase the synaptic matrix is given by $W_{ij} = w_{ij}d_{ij}$. With $w_{ij} = 1$ if there exists a pattern such that neurons i and j are co-activated, $w_{ij} = 0$ otherwise ; and d_{ij} is drawn randomly in $\{0, 1\}$ being 1 with probability $d \ll 1$.
To compute the capacity we follow the procedure described in the 'Methods' section. We first set the network in a state corresponding to one of the patterns $\vec{\xi^{\mu_0}}$, and ask whether this a fixed point of equation (1). This is done by computing $\mathbb{P}_{ne}$, the probability that all the fields are on the right side of the activation threshold. Similarly to equation (13),

$$
\mathbb{P}_{ne} = (1 - \mathbb{P}(h_i \leq fN\theta \,|\, \xi_i^{\mu_0} = 1))^{fN} (1 - \mathbb{P}(h_i \geq fN\theta \,|\, \xi_i^{\mu_0} = 0))^{(1-f)N} \tag{72}
$$

Using the equations (52), (54) that describe the distributions of inputs and the fact that $\mathbb{P}(h_i \leq fN\theta \,|\, \xi_i^{\mu_0} = 1) \simeq \mathbb{P}(h_i = fN\theta \,|\, \xi_i^{\mu_0} = 1)$ (see Dubreuil et al. (2014)) we can write

$$
\mathbb{P}(h_i \leq fN\theta \,|\, \xi_i^{\mu_0} = 1) = \exp(-fN\Phi^{dc}(\theta, d) + o(fN)) \tag{73}
$$

and

$$
\mathbb{P}(h_i \geq fN\theta \,|\, \xi_i^{\mu_0} = 1) = \exp(-fN\Phi^{dc}(\theta, dq) + o(fN)) \tag{74}
$$

where q is the fraction of synapses on i, w_{ij}, that are 1 after the learning phase, which can be expressed as

$$
q = 1 - \exp(-\alpha) \tag{75}
$$

Table 1: Notations

M	number of modules
N	number of neurons in each module
F	fraction of active modules in a memory
f	fraction of active neurons in an active module
β	$\beta = f\frac{N}{\ln N} = O(1)$
P	number of stored memories
$\mathscr{P}$	number of categories
p	number of memories in each category
c	number of categories in which a module is activated
α	storage load $\alpha = pf^2 = O(1)$
γ	ratio between the numbers of synaptic contacts from long and short-range origins
θ	activation threshold
g	fraction of local pairs of neurons co-activated at least once during learning
G	fraction of non-local pairs of neurons co-activated at least once during learning
E	amount of error to correct
η	amount of local inhibition

For $\mathbb{P}_{ne}$ to go 1 in the large N limit, equations (17),(18) have to be satisfied. Saturating these inequalities leads to a choice of activation threshold $\theta \to d$ and a coding level with

$$\beta = \frac{1}{d(-\ln q + 1 - q)} \tag{76}$$

In the specific case we are studying the general expression for the storage capacity (20) can be written

$$I = \frac{1}{\ln 2}\frac{\alpha}{\beta d} \tag{77}$$

Using the two equations (75) (76), it can be expressed more simply

$$I = \frac{\ln(1-q)(\ln q + 1 - q)}{\ln 2} \tag{78}$$

A maximal storage capacity $I = 0.26$ is reached at $q = 0.24$.
Note that for modular networks we can not get such a closed form for I since the $\mathbb{P}(h_i = fN\theta)$ needs to be estimated numerically.

References

Amit, D. J. (1989). *Modeling brain function*. Cambridge University Press.

Amit, D. J. and Fusi, S. (1994). Dynamic learning in neural networks with material synapses. *Neural Computation*, 6:957–982.

Bosking, W., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J Neurosci.*, 17:2112–2127.

Braitenberg, V. and Schütz, A. (1991). *Anatomy of the cortex*. Springer-Verlag.

Brunel, N. (2003). Network models of memory. In *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School*, pages 407–476.

Brunel, N. (2005). Network models of memory. In Chow, C., Gutkin, B., Hansel, D., Meunier, C., and Dalibard, J., editors, *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School 2003*. Elsevier.

Buzas, P., Eysel, U. T., Adorjan, P., and Kisvarday, Z. F. (2001). Axonal topography of cortical basket cells in relation to orientation, direction, and ocular dominance maps. *J Comp Neurol*, 437:259–285.

DeFelipe, J., Conley, M., and Jones, E. G. (1986). Long-range focal collateralization of axons arising from corticocortical cells in monkey sensory-motor cortex. *J Neurosci.*, 6:3749–3766.

Dubreuil, A. M. (2014). *Memory and cortical connectivity, PhD thesis*. Université Paris Descartes.

Dubreuil, A. M., Amit, Y., and Brunel, N. (2014). Memory capacity of networks with stochastic binary synapses. *PLoS computational biology*, 10(8):e1003727.

Enoki, R., Hu, Y. L., Hamilton, D., and Fine, A. (2009). Expression of long-term plasticity at individual synapses in hippocampus is graded, bidirectional, and mainly presynaptic: optical quantal analysis. *Neuron*, 62(2):242–253.

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

Fuster, J. M. (1995). *Memory in the cerebral cortex*. MIT Press.

Fuster, J. M. and Alexander, G. (1971). Neuron activity related to short-term memory. *Science*, 173:652–654.

Gilbert, C. D. and Wiesel, T. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci.*, 9:2432–2442.

Hellwig, B. (2000). A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological cybernetics*, 82(2):111121.

Holmgren, C., Harkany, T., Svennenfors, B., and Zilberter, Y. (2003). Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J. Physiol.*, 551:139–153.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, 79:2554–2558.

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Johansson, C. and Lansner, A. (2007). Imposing biological constraints onto an abstract neocortical attractor network model. *Neural Comput.*, 19(7):1871–1896.

Kalisman, N., Silberberg, G., and Markram, H. (2005). The neocortical microcircuit as a tabula rasa. *Proc Natl Acad Sci U S A*, 102(3):880–885.

Knoblauch, A., Palm, G., and Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289341.

Kropff, E. and Treves, A. (2005). The storage capacity of Potts models for semantic memory retrieval. *J. Stat. Mech.*, 8:P08010.

Loewenstein, Y., Kuras, A., and Rumpel, S. (2011). Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo. *J. Neurosci.*, 31(26):9481–9488.

Mari, C. F. (2004). Extremely dilute modular neuronal networks: Neocortical memory retrieval dynamics. *Journal of Computational Neuroscience*, 17:57–79.

Mari, C. F. and Treves, A. (1998). Modeling neocortical areas with a modular neural network. *Biosystems*, 48(1):47–55.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215.

Meli, C. and Lansner, A. (2013). A modular attractor associative memory with patchy connectivity and weight pruning. *Network*, 24:129–150.

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

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335:817–820.

Montgomery, J. M. and Madison, D. V. (2004). Discrete synaptic states define a major mechanism of synapse plasticity. *Trends Neurosci.*, 27(12):744–750.

Nadal, J.-P. (1991). Associative memory: on the (puzzling) sparse coding limit. *J. Phys. A: Math. Gen.*, 24:1093–1101.

O'Connor, D. H., Wittenberg, G. M., and Wang, S. S.-H. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proc Natl Acad Sci U S A*, 102:9679–9684.

O'Kane, D. and Treves, A. (1992). Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25:5055.

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

Petersen, C. C., Malenka, R. C., Nicoll, R. A., and Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proc.Natl.Acad.Sci.USA*, 95:4732–4737.

Pucak, M. L., Levitt, J. B., Lund, J. S., and Lewis, D. A. (1996). Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *J. Comp. Neurol.*, 338:360–376.

Romo, R., Brody, C. D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399:470–474.

Roudi, Y. and Treves, A. (2004). An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010.

Roudi, Y. and Treves, A. (2006). Localized activity profiles and storage capacity of rate-based autoassociative networks. *Physical Review E*, 73(6):061904.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32:1149–1164.

Stepanyants, A., Martinez, L. M., Ferecsko, A. S., and Kisvarday, Z. F. (2009). The fractions of short-and long-range connections in the visual cortex. *Proceedings of the National Academy of Sciences*, 106(9):35553560.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995.

van Vreeswijk, C. and Sompolinsky, H. (2003). Irregular activity in large networks of neurons. In *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School*, pages 341–402.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.*, 24:455–463.

Willshaw, D., Buneman, O. P., and Longuet-Higgins, H. (1969). Non-holographic associative memory. *Nature*, 222:960–962.