On memories, neural ensembles and mental flexibility

Dimitris A. Pinotsis[1,2], Scott L. Brincat[1] and Earl K. Miller[1]

[1] The Picower Institute for Learning & Memory and Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]The Wellcome Trust Centre for Neuroimaging, University College London, WC1N 3BG, UK

Correspondence: Dimitris A. Pinotsis

The Picower Institute for Learning & Memory

and Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Cambridge, MA 02139, USA

Tel (+1) 617-252-1790

Fax (+1) 617-452-2588

pinotsis@mit.edu

KEYWORDS: memory engrams; neural ensembles; working memory; neural field theory; auto-encoders; effective connectivity; characteristic path length; betweenness centrality

Acknowledgements. This work is supported by NIMH R37MH087027. DAP gratefully acknowledges travel support by USAF grant no FA8655/13/1/3019. He is also grateful to Professor Karl Friston for invaluable discussions and support and Professor Viktor Jirsa for elucidating discussions on neural fields.

Abstract

Memories are assumed to be represented by groups of co-activated neurons, called neural ensembles. Describing ensembles is a challenge: complexity of the underlying micro-circuitry is immense. Current approaches use a piecemeal fashion, focusing on single neurons and employing local measures like pairwise correlations. We introduce an alternative approach that identifies ensembles and describes the effective connectivity between them in a holistic fashion. It also links the oscillatory frequencies observed in ensembles with the spatial scales at which activity is expressed. Using unsupervised learning, biophysical modeling and graph theory, we analyze multi-electrode LFPs from frontal cortex during a spatial delayed response task. We find distinct ensembles for different cues and more parsimonious connectivity for cues on the horizontal axis, which may explain the oblique effect in psychophysics. Our approach paves the way for biophysical models with learned parameters that can guide future Brain Computer Interface development.

Introduction

Memories are assumed to be represented by groups of co-activated neurons, called neural ensembles. How to identify and describe neural ensembles has long been a central issue in neuroscience (Hebb, 1949). It is not an easy task: one has to deal with an immensely complex system where billions of neurons are linked to each other through trillions of connections. A further complication is that neurons can have multiple functions, especially in higher level cortex (Fusi et al., 2016; Rigotti et al., 2013). Thus, the same neurons may participate in many different ensembles and, conversely, different ensembles might share some of the same neurons. Clearly, the structure of these ensembles cannot be described in terms of anatomical connectivity only: if anatomical connectivity was all there was to ensembles, then activating one would activate others leading to a jumble of ensembles. Further, anatomy alone seems to preclude a hallmark of higher cognition: flexibility. Ensembles should be able to break apart and re-form from moment to moment without changing the underlying anatomy. Finally, ensembles are functional units and thus defining them based on anatomy alone is not possible in a behavioral context.

Previous work has attempted to identify neural ensembles using electrophysiological measures (Brown et al., 1998; Diba and Buzsáki, 2007; Johnson and Redish, 2007) and more recently optogenetics and immediate early gene (IEG) labelling (Ryan et al., 2015). However, due to slow dynamics, these approaches can only provide limited insights into fast activity and neural oscillations that are thought to play a key role in memory function and ensemble formation (Buschman et al., 2012; Fries et al., 2007; Fusi et al., 2016; Haegens et al., 2011; Miller and Buschman, 2013). Thus far, electrophysiology studies have only considered neural ensembles in a piecemeal fashion, that is, using pairwise correlations. They have focused on single neurons and/or functional connectivity between pairs of neurons and/or recording sites; the existence of an ensemble is thus inferred indirectly (Buschman et al., 2012; Gray, 1999; Modi et al., 2014).

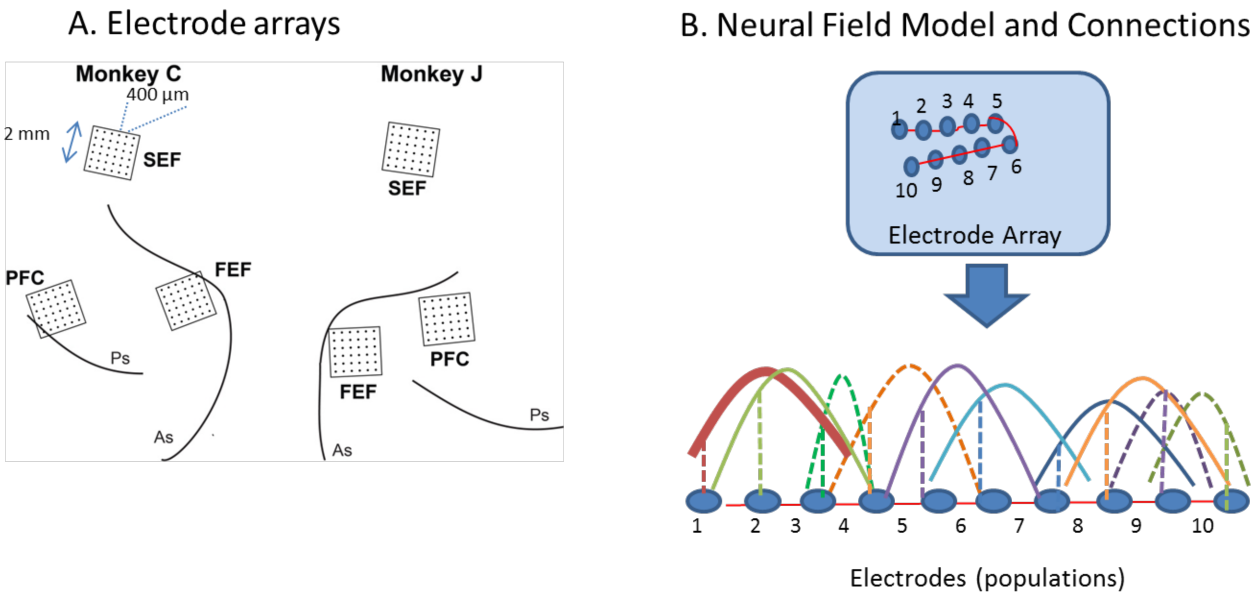
Here, we suggest an alternative approach to identifying ensembles based on effective connectivity. We describe ensemble properties using neurophysiological data combined with ideas from biophysical modelling, unsupervised learning and complex systems theory. We analysed multiple-electrode recordings obtained during a classic test of working memory: spatial delayed response (Funahashi et al., 1990; Fuster et al., 1985). We examined LFPs between simultaneously recorded electrodes in dorsolateral prefrontal cortex (PFC), supplementary eye field (SEF), and frontal eye fields (FEF). Our goal was to identify neural ensembles carrying spatial information in a *holistic*, not piecemeal, fashion and describe the connections that form them.

We used brain decoding algorithms, graph theory and spectral analysis to understand the structure of neural ensembles that give rise to observed patterns of LFP responses. This allowed us to treat neural ensembles as complex networks and to describe properties of the underlying connectivity. We obtained the estimates of effective connectivity underlying the ensembles by training a biophysical neural field model as a particular type of deep neural network called an auto-encoder. We found that we could describe ensemble properties and use them to decode the spatial location held in working memory using only a few parameters, which makes this approach computationally tractable. Further, it also revealed ensemble properties that cannot be observed using pairwise correlations. For example, using topological measures, we found that network connectivity in the spatial delayed response task was different for different cued locations. Cues on the horizontal axis had shorter characteristic path lengths (the least number of steps between different network nodes) than others. This could explain the oblique effect (psychophysics performance is better for stimuli on than off the horizontal/vertical axes). We also found connectivity and corresponding oscillatory dynamics across different spatial scales and different frequencies within cortical areas, which gives a new dimension to cortical network interactions.

Materials and Methods

Experimental Data and Recording Setup

Two adult male monkeys (monkey C, Macaca fascicularis, 9kg; monkey J, Macaca mulatta, 11kg) were handled in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care. They were trained to perform an oculomotor spatial delayed response task. This task required the monkeys to hold the location of one of six randomly chosen visual targets (at angles of 0, 60, 120, 180, 240 and 300 degrees, 12.5-degree eccentricity) in memory over a brief (750 ms) delay period and then saccade to the remembered location. If a saccade was made to the cued location, the target was presented with a green highlight and a water reward was delivered otherwise the target was presented with a red highlight and reward was withheld. Three 32-electrode chronic arrays were implanted unilaterally in PFC, SEF and FEF in each monkey (Figure 1A). Each array consisted of a 2 x 2 mm square grid, where the spacing between electrodes was 400 um. The implant channels were determined prior to surgery using structural magnetic resonance imaging and anatomical atlases. From each electrode, we acquired both threshold-crossing spike waveforms and local field potentials (extracted with a fourth order Butterworth low-pass filter with a cut-off frequency of 500Hz, and recorded at 1 kHz) using a multichannel data acquisition system (Cerebus, Blackrock Microsystems). We analyzed local field potentials (LFPs) during the delay period when monkeys held the cued locations in memory. We assumed that each electrode sampled LFP activity from a neural population in its proximity and modelled each brain area as a cortical area sampled at $N_S = 32$ locations. LFP activity was modelled by a mathematical model of wave dynamics known as a neural field. Electrodes were numbered in a monotonic fashion; neighbouring electrodes had adjacent numbers.


A. Electrode arrays

B. Neural Field Model and Connections

Figure 1. **A.** Recording setup. Three 32-electrode chronic microelectrode arrays were implanted in dorsolateral prefrontal cortex (PFC), supplementary eye field (SEF), and frontal eye field (FEF) in each monkey. Each array consisted of a 2 x 2 mm square grid, where the spacing between electrodes was 400 um. Ps: principal sulcus; As: arcuate sulcus. One monkey received implants in the left hemisphere and the other in the right hemisphere. These were located near Ps and As in both monkeys. **B.** Neural field model and connections. Neural fields provided a quantitative way to describe each ensemble's network interactions and patterns of activity across simultaneously recorded sites. The same model can describe different ensembles. Each electrode occupies a position on a cortical manifold (line) W parameterized by the variable υ and is connected to all other electrodes with connections whose strength follows a Gaussian profile (coloured solid and dashed lines), see also Equation (4).

Neural ensembles for memory maintenance

Using a neural field model allowed us to use patterns of LFP activity across recording sites to infer the underlying effective connectivity for each of the cued locations. Neural fields provided a quantitative way to describe each ensemble's network interactions and make predictions about patterns of activity that correspond to different attractor states, see Figure 1B. Each attractor state can be considered to reflect an ensemble or engram (Liu et al., 2012). This is also related to chimera states and metastability (Martens et al., 2016). Our goal was to obtain learned connectivity parameters that can describe the structure of neural ensembles activated while remembering different stimuli. The spacing between electrodes was larger (400um) than what is thought to be the origin of the LFP signal (250um, see Katzner et al., 2009). Using brain decoding algorithms and graph theoretic measures (centrality) we quantified the separability of neural ensembles in SEF and FEF (see below). However, volume conduction could in principle introduce confounds. These can be accommodated by using the effective connectivity parameters obtained here as priors to fit a more complicated biophysical model that accounts for volume conduction effects (Pinotsis et al., 2014). As it is common in computational neuroscience and modern machine learning approaches, we describe neural activity using a one dimensional model. In this model, space is defined along the line traced out by the electrodes. This is similar to ring models (Ben-Yishai et al., 1997; Somers et al., 1995), recurrent neural networks (Botvinick and Plaut, 2006; Sak et al., 2014; Shriki and Yellin, 2016) and deep convolutional neural networks (Krizhevsky et al., 2012). Also, one dimensional neural field models have proven very useful for explaining brain dynamics (Deco et al., 2008; Pinotsis et al., 2013; Robinson et al., 2014).

The neural field model describes transient fluctuations of neural activity around baseline. These fluctuations are similar to spontaneous fluctuations in resting state networks that are often described by neural fields (Pinotsis et al., 2013), but at the micro-scale. Similar patterns at a much larger spatial scale have been obtained with fMRI, e.g. (Sporns, 2013). In the classical neural field approach, current fluxes are considered as continuous processes on the cortical manifold described via partial differential or integro-differential equations depending on space and time (Jirsa and Haken, 1996; Coombes and Owen, 2004; Breakspear et al., 2006; Breakspear and Jirsa, 2007; Bressloff, 2010; Jirsa et al., 2010; Grindrod and Pinotsis, 2011; Pinotsis et al., 2012a). We here combine neural fields with unsupervised learning to obtain the cortical connectivity underlying interactions at the microscale (within a neural ensemble). This is different from earlier work in the biophysical modeling literature where the connectivity weights are usually chosen ad hoc so that model predictions resemble observed activity, e.g. large scale resting state activity measured with fMRI.

Deep neural fields for memory representations

We assumed that each of the three areas we recorded from (SEF, FEF and PFC) comprised a large number of neural populations (indexed by $i=1,..,N$ *,N=32),* that was equal to the number of electrodes we sampled from. These populations were sensitive to different cued locations (values of the visual angle θ). Each of these populations can be thought of as centered around a point υ. They also interact with other populations located at points ψ, via a connectivity function $K\left(\upsilon,\psi,t,t'\right)$. Our neural field model describes these interactions and its state represents cue maintenance.

Mathematically, we obtained the attractor state by perturbing around baseline. Attractor states are solutions of neural fields together with spatially and temporally periodic patterns beyond Turing instabilities; for example, localised regions of activity such as bumps and travelling waves, e.g. (Pinto and Ermentrout, 2001;Laing and Troy, 2003). Attractor states have been shown to describe patterns of neural activity observed during the memory delay (Hansel and Sompolinsky, 1998; Tsodyks and Sejnowski, 1995). We performed a linear stability analysis and expanded in a Taylor series with respect to the space variable. Then a simple reformulation of the neural field as a Gaussian model resulted in a model of the sort considered in deep neural networks like variational auto-encoders:

$$Y = \sum_k G_k z_k + R \tag{1}$$

where the constant R will be defined below (see Equation 9) and G_k are the derivatives of the vector-valued function of perturbations $\hat{X}$ of depolarizations $X(\upsilon,t) = \begin{bmatrix} x_1(\upsilon,t) \\ \vdots \\ x_n(\upsilon,t) \end{bmatrix}$ around baseline activity $X(\upsilon,t) = X_0(\upsilon) + \hat{X}(\upsilon,t)$, with respect to the variable $\upsilon \in W$ that parameterizes the cortical manifold W occupied by the neural field,

$$G_k = \frac{\partial \hat{X}(\upsilon,t)}{\partial \upsilon^{(k)}} \tag{2}$$

The functions G_k are the *principal axes*. z_k are the latent variables, which we call the *connectivity components.* These functions are defined by the following equations

$$\begin{aligned} z_0(\upsilon) &= df_0 B^{-1} \int_W K(\psi,\upsilon) d\psi \\ z_k(\upsilon) &= \tfrac{df_0}{k!} B^{-1} \int_W K(\psi,\upsilon)(\psi-\upsilon)^k d\psi \end{aligned} \tag{3}$$

Neural fields predict average firing rate or depolarization. To obtain Equation (1), we start from an equation describing the general mathematical form of neural fields

$$\dot{X} = -BX + K * f \circ X + S \circ U \tag{4}$$

where * denotes the integral

$$K * Q = \iint K\left(\upsilon,\psi,t,t'\right) \cdot Q(\psi,t') d\psi dt' \quad (5)$$

ψ is the location where afferent input originates from and B is a matrix encoding the rate-constants of postsynaptic filtering.

Equation (4) says that the rate of change of depolarization $\dot{X}$ of a neural population occupying a location υ comprises three terms; the first is a simple decay, the second is due to presynaptic inputs from other parts of the cortical manifold and the final part is due to external inputs U, where $S:\mathbb{R}^n \to \mathbb{R}^n$ maps exogenous inputs to depolarization. Also, $f:\mathbb{R}^n \to \mathbb{R}^n,\ f(Q) = \dfrac{1}{1+\exp(\gamma(\eta - Q))}$ is a vector-valued transfer function that describes the mapping from postsynaptic potentials to average firing rate (Lipschitz continuous to guarantee local existence) of the population around point $\upsilon \in W$. Here, γ is synaptic gain and η is the postsynaptic potential at which the half of the maximum firing rate is achieved, see e.g. (Pinotsis et al., 2012b) for more details.

In the literature, the connectivity matrix $K\left(\upsilon,\psi,t,t'\right)$ often exhibits translational invariance, that is $K\left(\upsilon,\psi,t,t'\right) = K\left(\left|\upsilon-\psi\right|,t,t'\right)$. Here, we do not assume translational invariance. In other studies where there is no invariance constraint, entries of the connectivity matrix are chosen by hand. The novelty of our approach is that we obtained cue-specific entries for this matrix using a Bayesian algorithm. This can be thought of as a probabilistic analogue of the classical principal component analysis (PCA): obtaining the connectivity components amounts to obtaining principal components. This also maximizes the mutual information between the cue and its representation.

Obtaining efficient cue representations using neural fields

A common mathematical simplification in fixed point networks is the assumption of stationarity. This eschews the need for explicit numerical integration. It also means that we can drop the explicit time dependence from the equations and functions, e.g. write $K\left(\upsilon,\psi\right)$ instead of $K\left(\upsilon,\psi,t,t'\right)$. In our analyses below, we consider a single activity variable $X(\upsilon,t)=x_1(\upsilon,t)$ following a usual assumption sometimes known as a *stable attractor* (Brunel and Wang, 2001). This means that the variable $X(\upsilon,t)$ comprises both excitation and inhibition. This assumption does not exclude complex dynamics resulting from the interaction of excitatory and inhibitory populations: this was demonstrated by Brunel and Wang who considered mean fields of populations of spiking neurons based on the integrate-and-fire neuronal model. It is the mean field of the population activity that demonstrates the attractor state, which is the level of description chosen here, see also (Pinotsis et al., 2013). Brunel and Wang showed that noise induced in a population of spiking neurons will change the spiking dynamics, but the mean field will have a particular constant value as long as all other characteristics stay the same and this noise is small. These attractor states can be thought of as mediating the flexible formation of neural assemblies associated with different inputs or stimuli. This is not a new idea; attractor states are well known to mediate persistent activity in working memory and similar tasks, see e.g. (Wei et al., 2012). Under the above assumptions, it is straightforward to rewrite Equation (4) in the form of a Taylor expansion

$$\hat{X} \approx z_0G_0 + z_1G_1 + z_2G_2 + \ldots + z_nG_n \qquad (6)$$

where we linearized $f \circ X(\upsilon,t) \approx df_0 X(\upsilon,t)$. This result was obtained by expanding with respect to the space variable υ. In the following, we adopted a probabilistic framework. We assumed that cortical activity $\hat{X}(\upsilon,t) \in \tilde{X}$ was sampled from a random process $\tilde{X}$. This describes activity generated by the cue maintenance. It also means that different trials $\hat{X}^l, l=1,\ldots S$ correspond to different realizations of the process $\tilde{X}$, $\hat{X}^l \in \tilde{X}$. Adding Gaussian noise ε with precision $r^2{}_s = 1/s^2{}_s$, we obtained the following probabilistic model

$$\hat{X} = z_0 G_0 + z_1 G_1 + z_2 G_2 + \ldots + z_n G_n + \varepsilon$$
$$\varepsilon \sim (0, {s_s}^2 I) \qquad (7)$$

We then assumed that connectivity components were sampled from a normal distribution $z \sim N(0, I_Q)$. This ensured that they were uncorrelated. We also considered mean-centered activity

$$Y = \hat{X} - m$$
$$m = \frac{1}{S}\sum_l X^l \qquad (8)$$

We thus obtained Equation (1) where

$$R = m + \varepsilon \qquad (9)$$

We then used a Restricted Maximum-Likelihood (ReML) algorithm to obtain Bayesian optimal values for the connectivity components z_k (Harville, 1977). This algorithm optimizes the objective function *F*

$$F = \left(-\frac{1}{2}\right)\left[(Y - GZ)^T r_s^2 (Y - GZ) + \ln\left|s_s^2\right| + \ln\left|s^2{}_s \Delta^{-1}\right| + Z^T Z + \mathrm{co}\,nst\right] \qquad (10)$$

where *F* is called Free Energy (Neal and Hinton, 1998) and

$$\begin{aligned}\Delta &= s^2{}_s I + G^T G \\ Z &= \Delta^{-1} G^T Y\end{aligned} \tag{11}$$

The above expression of the Free Energy obtains from the sum

$$F = E_{z\sim Q}\left[\log p(Y|z)\right] - D_{KL}\left[Q(z|Y)\|N(0, I_Q)\right] \tag{12}$$

under a Laplace assumption, where Q is the approximate posterior and after substituting $z \sim \mathrm{N}(0, I_Q), q \sim Q(z|Z, s_s^2\Delta^{-1})$ and $Y \sim \mathrm{N}(0, GG^T + s^2{}_s I_P)$, see also (Friston, 2008). Equation (12) is the objective function used in variational auto-encoders where the data is generated by a directed graphical model $p(Y|z)$ and the encoder is learning an approximation q to the posterior $p \sim \mathrm{N}(z|Y)$.

In summary, we expressed a neural field model as a variational auto-encoder. This allowed us to project high dimensional data (electrophysiological time series) to low dimensional connectivity components z. We call a neural field model with parameters obtained by optimizing a Free Energy cost function, a deep neural field model.

Effective connectivity within a cortical area can be used for cue classification

We used Naïve Bayes and diagonal Linear Discriminant Analysis (LDA) to classify feature vectors comprising connectivity components (Witten et al., 2016). To classify a feature vector, diagonal LDA and Naïve Bayes use a maximum likelihood decision rule

$$\tilde{h} = \arg\max_{\theta} \left[\log p(y=\theta) + \log p(z|y=\theta) \right] \tag{13}$$

where *h* denotes the class (location or angle) and the first and second terms in the brackets are the class prior and the class conditional density respectively. Assuming a Gaussian form for the conditional density,

$$p(z|y=\theta) \sim N(\mu_\theta, \Sigma_\theta) \tag{14}$$

results in a decision rule of the form

$$\tilde{h} = \arg\min_{\theta} (z-\mu_\theta)\Sigma_\theta(z-\mu_\theta)^T + \log|\Sigma_\theta| \tag{15}$$

Taking $\Sigma_\theta = diag(ss_{\theta 1}^2, \ldots, ss_{\theta N}^2)$ yields Naïve Bayes. Then, the class densities have diagonal covariance matrices and the decision rule is the sum of components from each sample vector

$$\tilde{h} = \arg\min_{\theta} \sum_{l}^{M} \left(z_l - \mu_{\theta l}\right)^2 / ss_{\theta l}^2 + \log ss_{\theta l}^2 \tag{16}$$

If $\Sigma_\theta = \Sigma$, the covariance matrix is "tied" across classes and the decision rule estimates a pooled covariance matrix. For independent features, the decision rule takes a rather simple form

$$\tilde{h} = \arg\min_{\theta} \sum_{l}^{M} \left(z_l - \mu_l \right)^2 / ss_l^2 \tag{17}$$

This is known as diagonal LDA.

Obtaining the connectivity from its components

We assumed a Gaussian connectivity profile for the connectivity $K(\psi, \upsilon)$

$$K(\psi,\upsilon) = \left(C\sqrt{2\pi} \right)^{-1} \exp\left\{ -(\psi - \upsilon - \bar{\psi})^2 / 2C^2 \right\} \tag{18}$$

where $\bar{\psi}$ and C are the mean and standard deviation. These parameters can be readily obtained using the first three connectivity components A_0, A_1 and A_2 [1]:

$$\begin{aligned} \bar{\psi} &= \Re \, {}^{A_1}\!/_{A_0} \\ C &= \Re \, {}^{A_1}\!/_{A_0} \frac{\sqrt{A_0 A_2 - A_0^{\,2}}}{A_0} \end{aligned} \tag{19}$$

This follows standard results in the theory of linear inverse problems (Bertero et al., 1988)

[1] $\Re$ denotes that we take the real part of the expression that follows it.

that occur in various branches of physics (Heinz, 2013; Mersmann, 1995). Below, we used a small number of connectivity components to obtain the weights, that is A_0, A_1 and A_2 [2].

Topological features of neural ensembles

We used the connectivity matrices and graph theoretic measures to describe information flow within neural ensembles, cf. (Sporns, 2013). We first computed the *characteristic path length*

$$L_a = 1/(N-1)\sum_{a \neq b} L_{a,b} \tag{20}$$

We will see below that using this measure allowed us to quantify information transfer efficiency within neural ensembles.

We also used *betweenness centrality*

$$X_c = 1/(N-1)(N-2)\sum_{\substack{a,b \in E \\ a \neq b, a \neq c}} \frac{K_{ab\backslash c}}{K_{ab}} \tag{21}$$

where $K_{ab\backslash c}$ is the sum of shortest paths connecting electrodes *a* and *b* that do not cross electrode *c* and K_{ab} is the sum of all shortest paths connecting electrodes *a* and *b.* This allowed us to characterize the degree of overlap between different neural ensembles (see below).

[2] Our approach can be generalised to include different parameterizations for connectivity matrix (e.g. exponential or gamma functions) or numerically estimate each of the 1024 weights independently.

Results

Identifying neural ensembles

We first used the neural field model to characterise how different cued locations are represented in different cortical areas in a concise and holistic manner, in terms of a few parameters that describe microscale connectivity between recording sites within each area. The parameters that appear in the neural field model describe the dispersion and topography of neuronal connections that are likely to underlie observed patterns of LFPs across recording sites. To obtain them, we trained a neural field model using raw LFP time series and applied a PCA-type algorithm. This process yielded the principal components and axes. Mathematically, the latent variables in a linear model are the principal components. Here, the principal components describe the spatial information in the LFP data. That is, they describe the patterns of activity across recording sites. We focus on the principal components first. In the next section, we will consider another output of the PCA decomposition: the principal axes, which contain the temporal information.

The spatial principal components are matrix-valued functions with dimensionality equal to $N_T x N_S$, where $N_T = 600$ is the number of trials. We call each entry in these matrix functions the *component strength*. This expresses the *sum* of all *connectivity weights* that target the neurons that contribute to the LFPs observed from each electrode. The *connectivity weights* are parameters in the neural field model that describe the strength of the effective connections between the recording sites within each cortical area. They describe how the signal is amplified or attenuated when it propagates between recording sites. Large positive weights of connections targeting a certain electrode implies that large LFP responses would be expected from that recording site. These weights are the entries of the connectivity matrix K that multiplies input signal from other electrodes (second term in Equation 4). These recording sites also show high values of functional connectivity (see Supplemental Figure 1). Because the principal components describe the effective connectivity between sites we call them *connectivity components.*

In each panel in Figures 2A-C, we plotted the across-trial averages of the first four connectivity components (blue, orange, yellow and magenta lines, respectively) for all three areas. We will see in the next section that these correspond to the connectivity patterns that are expressed at four different spatial scales. The different panels show connectivity components for the different cued locations. In all these panels, the horizontal axis represents the electrode number and the vertical axis is the strength of the connectivity components.

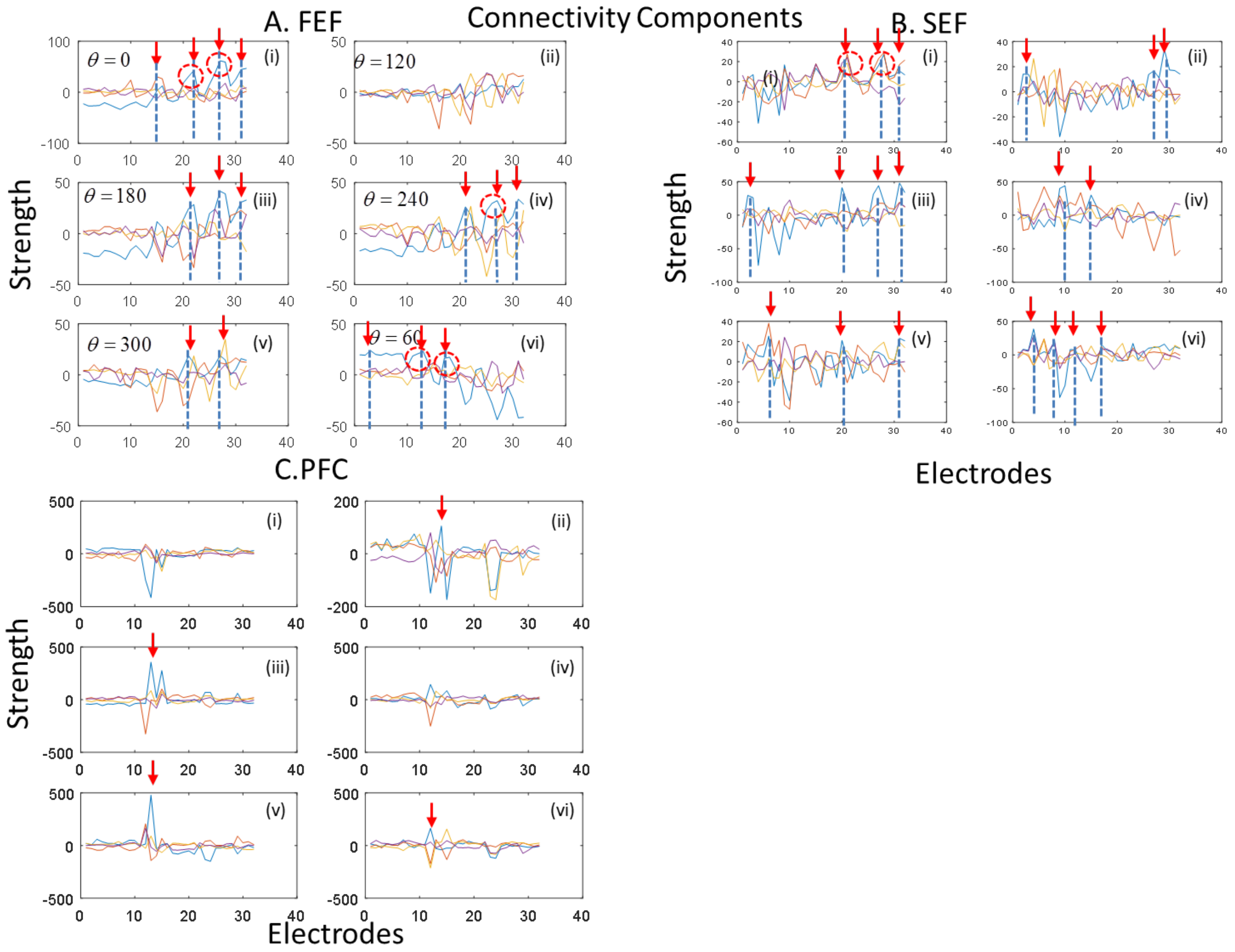


Figure 2. A. Trial average estimates of FEF connectivity components. Connectivity components are matrix-valued functions with dimensionality equal to $N_T x N_S$. These describe patterns of activity across recording sites and the connectivity underlying neural ensembles maintaining different cued locations. Red arrows and blue dashed lines indicate peaks of these components. These correspond to most activated electrodes and define neural ensembles. These ensembles are not due to between-trial variability; they reflect components that are consistently activated for a specific cued location. The value of the cue (angle) is shown in insets on the top left corner of each panel. **B.-C.** Connectivity components for SEF and PFC respectively. Order of cued locations follows that of Figure 2A.

Red arrows and blue dashed lines indicate peaks of these components that correspond to the recording sites that are most strongly activated: we call these *"peak"* sites. We suggest that these peak sites describe part of an ensemble for each cued location. Because we took trial-averages of connectivity components, these ensembles are not due to between-trial variability; instead they reflect components that are consistently activated for that cued location.

We then focused on the spatial structure of ensembles and asked whether they showed any traits of topographic clustering. In other words, that cues adjacent in the visual field induce responses that are also adjacent on the cortical surface (Vaina et al., 2014). Our focus was not on optimizing clustering but on preserving the spatial order of recording sites in the connectivity components. This is crucial for comparing and contrasting results from different brain areas[3]. Traits of topographic clustering can be observed by revisiting the results of Figure 2. We saw above that large component strengths imply high levels of activation for the corresponding recording sites (peak sites). To test for topographic clustering, we need to consider the connectivity component strengths that correspond not only to peak sites but also their neighbouring sites. To this end, the overall shape of the connectivity components can be quite informative: the FEF and SEF connectivity components have a regular shape that resembles waves (Figure 2A-B), while PFC

[3] See section *Between-area differences in topography and topology of ensemble connectivity* below.This also justifies our choice not to use modularity maximization or similar algorithms for optimizing clustering.

components have only local (isolated) peaks (they lack the waves, Figure 2C). Considering this together with the fact that large component strengths imply high levels of activation, we concluded that recording sites that are nearby to peak sites will show similar levels of activation as peak sites. This can be seen in the panels of Figure 2 as follows: connectivity component strengths that correspond to sites *adjacent* to peak sites, are depicted by coloured line segments that lie inside the *red circles* in panels 2A(i),(iv) and (vi) and 2B(i). Because these segments are smooth and (approximately) horizontal, sites adjacent to peak sites correspond to the same values of the component strength (depicted on the vertical axis) and therefore exhibit the same levels of activation as the peak sites. In other words, a continuous shape of the connectivity component strength implies a continuous activation profile as we move along electrodes placed on the *x*-axis in the panels of Figure 2. All in all, the regular, wave-like form of the FEF (and to a lesser extent, SEF) connectivity components taken together with the fact that sites that are nearby on the cortex are also adjacent on the horizontal axes of the panels in Figure 2 imply that FEF (and to a lesser extent, SEF) responses will show traits of topographic clustering. This could be explained by the stronger retinotopy in FEF (Funahashi et al., 1989). This can also be seen in the corresponding functional connectivity matrices that show traits of lattice or block structure (Supplemental Figure 1).

To sum so far, we have looked at the organisation of neural activity induced by different cued locations within cortical areas. We used connectivity components to identify neural ensembles associated with these locations. This also allowed us to characterize clustering within the cortical area.

Neural ensembles oscillate at theta, alpha and beta frequencies

In the previous section, we focused on the principal components obtained after training the neural field model on LFP data using a PCA-type algorithm. Next, we turn to principal axes. These contain temporal information (they are the only time-dependent terms in the linear Gaussian model). They predict the oscillatory profile of fluctuations around baseline. These fluctuations can be thought of as transient non Turing patterns (patterns that decay back to baseline) whose Lyapunov exponents determine the frequencies observed in sampled LFP activity.

To understand how the principal axes describe these fluctuations consider an analogy with the physical quantities of velocity and acceleration: velocity describes the distance over which a vehicle has moved in unit time. Then acceleration captures changes in velocity in unit time. This means that, for a vehicle moving at an arbitrary velocity, acceleration will vary faster than velocity over time because it is sensitive to changes in velocity. This means that it will operate at a finer *temporal* scale. Now, one could replace the word "temporal" with "*spatial"* and think of the second and third principal axes as velocity and acceleration over *space* instead of time: the first principal axis is a linear approximation to fluctuations around the mean and captures activity changes at the largest spatial scale. Then, the second is sensitive to changes of the first axis at a smaller scale, the third is sensitive to changes of the second at an even smaller scale and so on. Mathematically, they are the partial derivatives of neural activity with respect to the spatial parameter in the neural field model (this parameter describes the location of an electrode). Thus, each principal axis operates at a distinct spatial scale. This also means that the corresponding connectivity component is associated with a certain spatial scale. This follows simply from the usual correspondence between principal axes and principal components in common PCA. Principal axes and connectivity components are thus associated with a hierarchy of spatial scales: each scale being *finer* than the previous one.

Principal axes revealed characteristic frequencies in the LFP activity induced by a particular cue (similar to the modes of a spring that is perturbed from its original position and then exhibits transient oscillations). Principal axes are matrix-valued functions of dimensionality $N_T xT$, where *T*=720ms is the length of the raw LFP time series. We call each entry in these matrices the *axis strength*. This corresponds to an instantaneous scale factor with which the corresponding component strength must be multiplied to reconstruct the observed LFP. Taking across-trial averages of the first four principal axes for different cues and areas, we obtain the upper left panels of Figures 3-5.

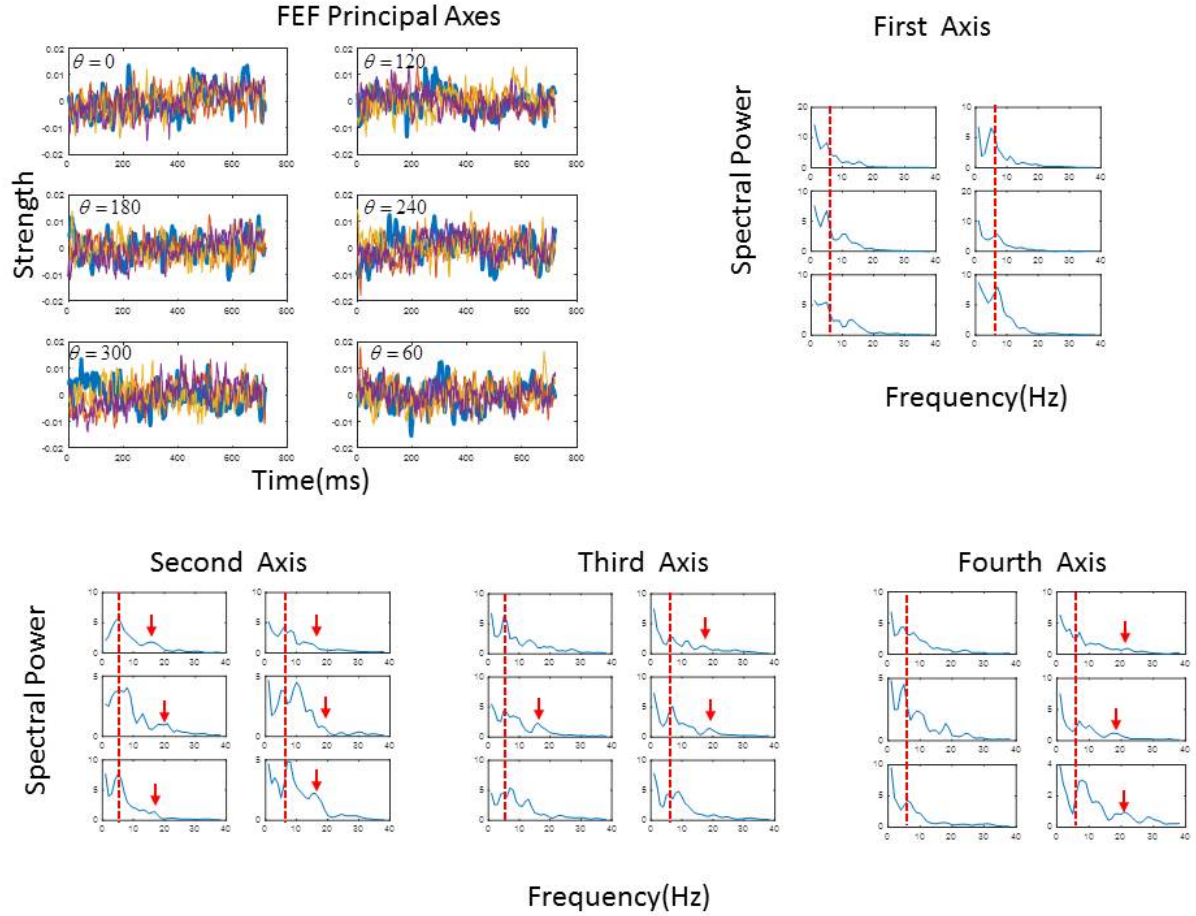


Figure 3. Trial averages of FEF principal axes and corresponding power spectra. Upper left panels: The first axis is depicted with a thick blue line. Orange, yellow and magenta lines depict the second, third and fourth principal axes. Each of the six upper left panels corresponds to a different cued location. Upper right panels: Spectral power associated with the first axis in FEF for all six cued locations seems to be expressed in the theta range: the abscissa of the *red* dashed line in these panels corresponds to 6Hz (upper right panel). This is the power we expect to dominate at large spatial scales because the spectra associated with the first axis operate at the largest spatial scale. Each of the six upper right panels corresponds to a different cued location. Bottom panels: Power spectra associated with the second, third and fourth principal axes and for all cued locations. Since each higher order axis operates at a finer spatial scale than the previous one, these correspond to more localised responses. The bottom panels reveal local beta band activity (depicted with *red arrows*). In all panels in this figure order of cued locations follows that of Figure 2A.

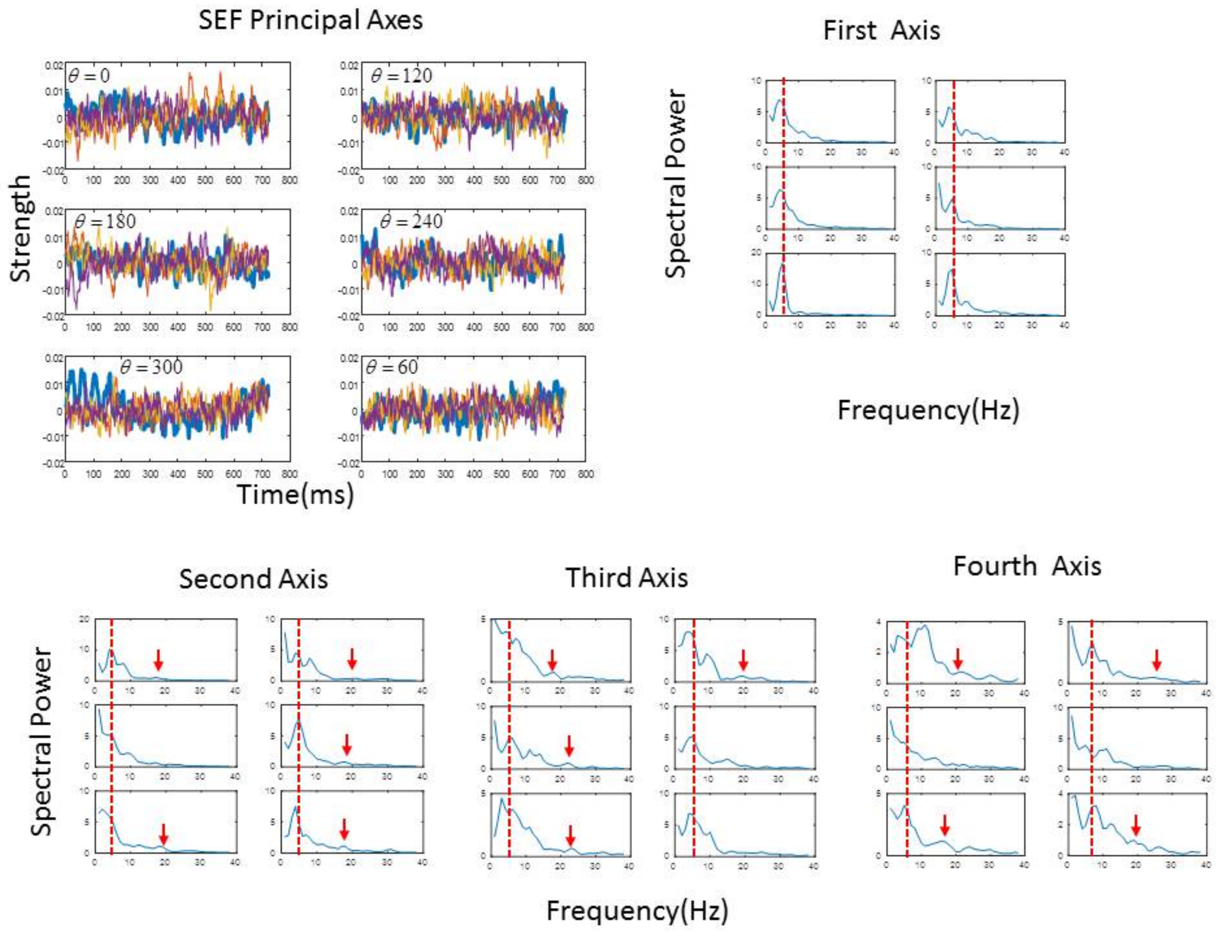


Figure 4. Trial averages of SEF principal axes and power spectra following the format of Figure 3. Similarly to power spectra in FEF, SEF spectral power associated with the first mode is expressed in the theta range while spectra of higher order principal axes also include beta activity. Order of cued locations for each panel follows that of Figure 2A.

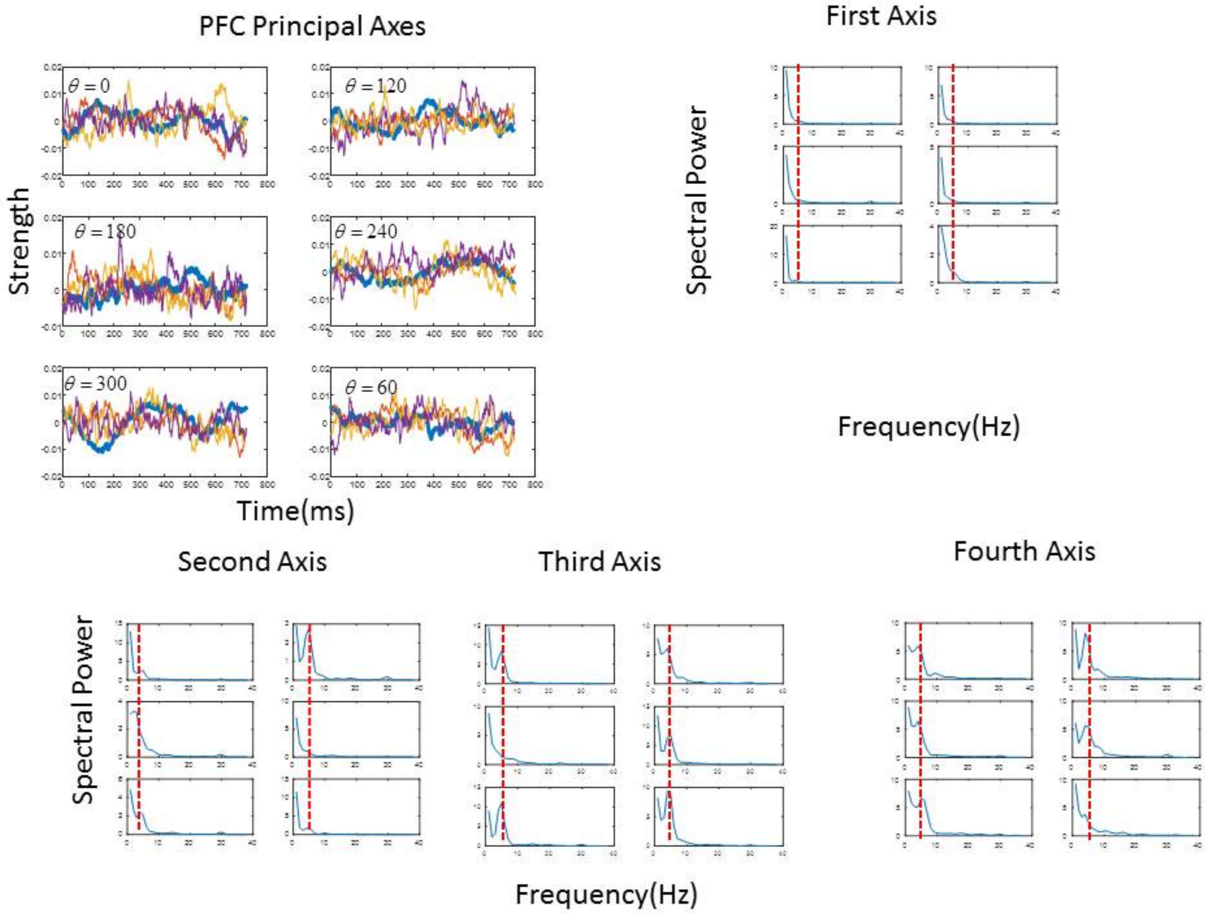


Figure 5. Trial averages of PFC principal axes and power spectra following the format of Figure 3. In contrast to the power spectra of figures 3 and 4 that exhibited global theta activity, theta activity in PFC appears to be localised: theta peaks appear only in the spectra of higher order axes and are absent in the spectra of the first axis. Order of cued locations for each panel follows that of Figure 2A.

In the upper left panels, the first mode is depicted with a thick blue line. Orange, yellow and magenta lines depict the second, third and fourth modes. The power spectra associated with different cued locations and each of these modes are shown in the upper right and three bottom panels of Figures 3-5. The upper right panels of Figures 3 and 4 reveal that spectral power associated with the first mode in FEF and SEF seems to be expressed primarily in the theta range: the abscissa of the *red* dashed line in these panels corresponds to 6Hz. This line corresponds to the peak frequency in the first mode power spectra. This is the power we expect to dominate at large spatial scales because the first mode spectra operate at the largest spatial scale. Also, alpha activity operates at the same scale. In the bottom panels, we plotted the power spectra associated with the second, third and fourth modes for different cue locations. In FEF and SEF, we notice beta rhythms (bottom panels of Figures 3 and 4). Since each higher order mode operates at a finer spatial scale than the previous one, beta rhythms correspond to more localised responses. In PFC spectra, theta peaks appear only in the higher order modes and are absent in the first mode (compare the upper right panel with the bottom panels of Figure 5). This means that theta responses in PFC are more localised than in the other two areas.

To sum up, our approach allowed us to look at oscillatory dynamics across different spatial scales and different frequencies. Our approach also gives a new dimension to cortical interactions by linking functional connectivity in different frequency bands to the spatial scales that generate it. This link can be established through the use of connectivity components. Each connectivity component expresses the functional connectivity between sites in one or more frequency bands. These are the bands in which characteristic frequencies appear in the corresponding principal axis. This follows from the usual correspondence between principal axes and connectivity components in PCA-type algorithms. Since the first and second FEF and SEF principal axes seem to be primarily expressed in the theta band (top right and bottom left panels of Figure 3), it follows that the first and second FEF and SEF connectivity components describe the connectivity in the theta band. In other words, FEF and SEF functional connectivity in the theta band results from interactions at the largest and second largest spatial scales. FEF functional connectivity matrices in the theta band for different cued locations are shown in the left panels of Figure 6. Here, different matrices depict functional connectivity of different neural ensembles. We also looked at functional connectivity in the beta band. This is described by the second and higher order connectivity components (beta band activity seems prominent in second and higher order principal axes, bottom panels of Figure 3). This is shown in the right panels of Figure 6. We noticed that functional connectivity strengths in the beta band are much weaker than in the theta band. Most electrode pairs have strengths around zero with a few electrode pairs with strength 0.05 (orange and yellow) or -0.05(dark blue). There is little synchronised beta activity between electrodes. This result confirmed our earlier conclusion that beta activity corresponds to more localised responses.

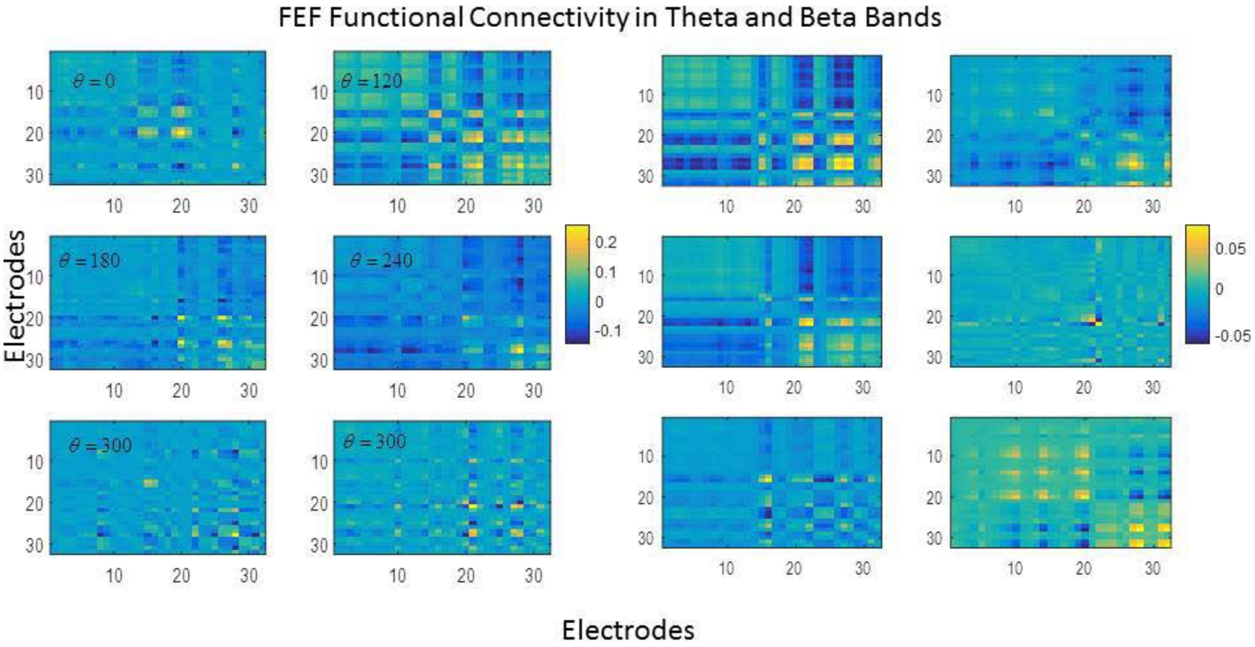


Figure 6. FEF Functional Connectivity in the theta (left) and beta (right) bands for different cued locations. Notice the block structure that is reminiscent of retinotopy. Also, functional connectivity values for the beta band are relatively weak. This is because beta oscillations are localised.

Neural ensembles for different cued locations do not overlap

Earlier, we described spatial aspects of neural ensembles in terms of connectivity components. In other words, we showed that cue maintenance is sub-served by different neural ensembles that are linked together via sets of connectivity patterns described by connectivity components. However, we have not yet tested the degree to which *differences* among neural ensembles are expressed by these components. Can we distinguish between memorized cues based on connectivity components? Earlier, we noted that many neurons may participate in multiple ensembles. A large degree of overlap between ensembles might mean that the connectivity components between different recording sites could overlap for different ensembles to the extent that we could not detect any difference in connectivity for the different cues. If, on the other hand, we can, it would help establish this approach as a means for detecting and describing ensembles. Below, we show that we can. This process also allows us to determine whether only a few connectivity components might be sufficient for decoding ensemble activity. From a computational perspective this is important, as having to use too many components would be computationally intractable.

To formally test which and how many connectivity components can sufficiently characterise differences in connectivity for different remembered cues, we used their entries (the 32 connectivity strengths) as *classification features* of different trials by cued location[4]. Recall that each strength expresses the sum of all connection weights that target the neurons that contribute to the LFPs from each of the 32 electrodes.

For our classification analyses, we varied the total number of classification features used by including a different number of connectivity components: the total number of classification features was equal to $32xN_c$, where $N_c = 1,\ldots,4$ is the number of connectivity components considered. Letting $N_c = 1$ meant that we only used one connectivity component, while $N_c = 4$ meant that we used all four shown in Figure 2 and so on. The specific goals of this analysis were: 1. To determine whether we can predict the location of the cue on each trial; 2. To appraise the predictive power (relative accuracy) of each of the connectivity

[4] A difference in our approach in comparison to similar work is that we used model parameters (the connectivity components shown in Figure 2) for classification as opposed to raw time series or some data features like oscillatory power or average firing rate, as is often used in the literature (De Martino et al., 2008; Formisano et al., 2008; Misaki et al., 2010;Jia et al., 2017).

components (the curves of different colours shown in the panels of Figure 2) and their combinations. In other words, we wanted to test whether some connectivity components might be more informative than others and whether accuracy would increase when we considered different or more connectivity components for classification (varied N_c). We used two different methods for classification, namely Naïve Bayes and diagonal LDA. We used data from 600 trials (450 trials for training and 150 trials for testing. Naïve Bayes and LDA are among the most commonly used classification algorithms.

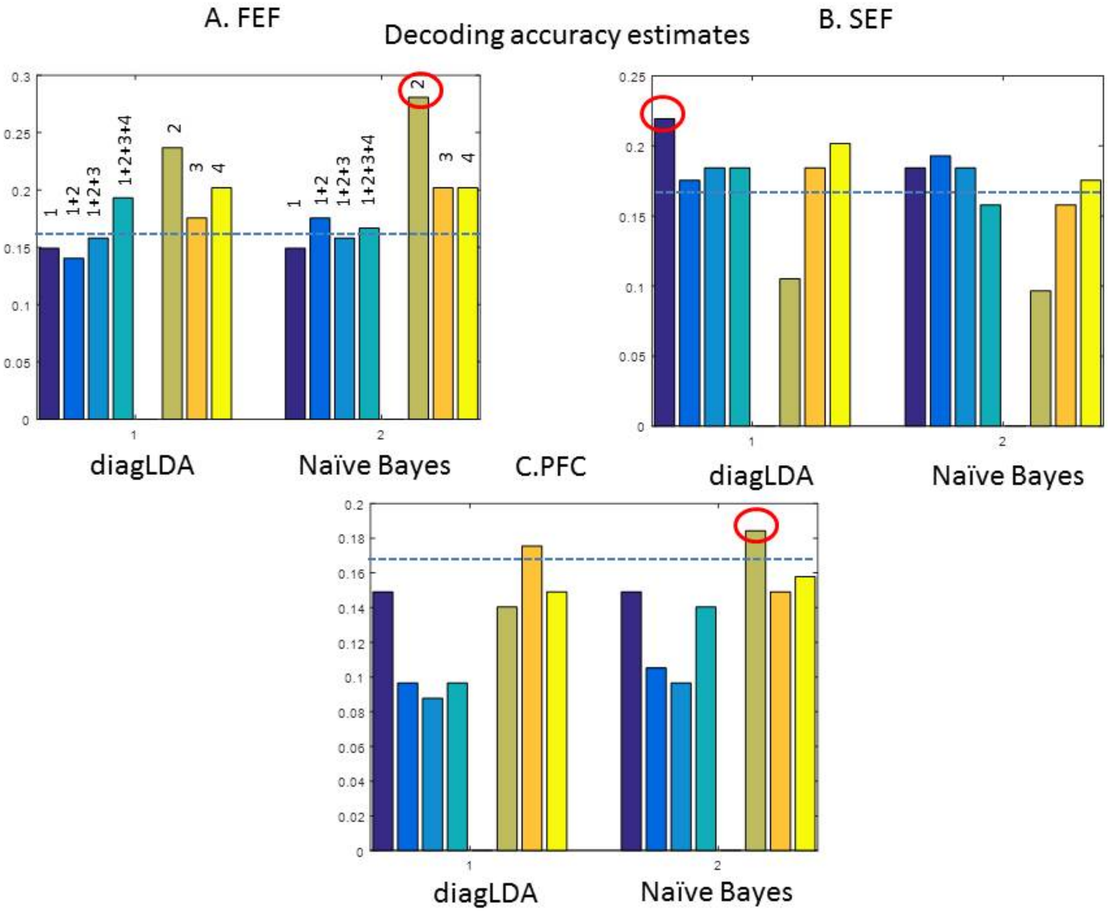


Figure 7. **A. - C.** Classification of different trials by cued location for each of the three brain areas. We varied the total number of classification features used by including a different number of connectivity components

and used Naïve Bayes and diagonal LDA as classification algorithms. Either a single component or multiple components were used (see the numbers above each bar). We obtained significant results for both FEF and SEF. Maximum accuracy in FEF was obtained with the second connectivity component (A), while the first connectivity component yielded maximum accuracy in SEF (B). PFC accuracy estimated appear to be less satisfactory, however the second component resulted in above chance level accuracy (C).

The results of our analysis are shown in Figure 7. The blue dashed line depicts chance level (0.1667). Results obtained using diagonal LDA are shown on the left while accuracy obtained after applying Naïve Bayes is shown on the right of each panel. Using both methods, we tested whether seven different combinations of connectivity components can be used for classification of the trials by cue location. Either a single component or multiple components were used (see the numbers above each bar, e.g. “1+2” means that we used the first and second components). Overall, both methods gave qualitatively similar results in terms of which features are informative for classification.

We obtained significant results for both FEF and SEF. This revealed which spatial scales of LFP activity are most informative for distinguishing between ensembles corresponding to different cued locations. As we saw in the previous section, each connectivity component corresponds to a spatial scale: the component of highest accuracy corresponds to the most informative spatial scale. In FEF, the second connectivity component (*orange* curve in the panels of Figure 2A) appears to be most informative (Figure 7A, , cf. the third green-brown bar from the right in the top left panel with the number “2” on top) for distinguishing between various cues maintained in FEF. This means that using a few components is sufficient for describing unique ensemble activity for different cues. It also means that we expect differences in FEF functional connectivity between neural ensembles to be expressed in a fine spatial scale.

In SEF, the first connectivity component (*blue* curve in the panels of Figure 2B) yields accuracy well above chance, cf. the first (blue) bar from the left in the top right panel. Thus, in SEF, the most informative spatial scale is larger than in FEF. This result fits nicely with the greater degree of topographic organisation of FEF responses observed earlier. A higher degree of topographic organisation of FEF in comparison to SEF responses would imply a more organised spatial structure that should be evident at a *finer* spatial scale.

The differences in the most informative components in FEF and SEF described above imply differences in the way LFP electrodes capture information in different frequency bands in the two areas. Functional connectivity was not equally informative in both areas. We found that SEF functional connectivity was informative in more frequency bands than in FEF. It was more sensitive to differences between neural ensembles. We concluded this as follows: We computed pairwise differences in the SEF functional connectivity between neural ensembles in two different frequency ranges: 1. In a narrow range that only included the theta band (cf. Figure 7 left panels) 2. In a broader range that included both theta and alpha bands. We summed together all pairwise differences and obtained two matrices of functional connectivity differences between ensembles: one for the narrow and one for the broad frequency range. We then subtracted these two matrices from each other and thresholded the resulting matrix to focus on the top 50% of the entries[5]. The result is shown in the right panel of Figure 8. We also computed the corresponding result for FEF (Figure 8, left panel). We noticed that including the alpha on top of the theta band when computing the functional connectivity in SEF (considering a broader frequency range) enhances the differences between ensembles: the right hand side matrix in Figure 8 includes a larger number of non-zero entries in comparison to the left hand side matrix. This means that a larger number of SEF electrodes were sensitive to differences in functional connectivity between ensembles when we used both theta and alpha bands (to compute the functional connectivity) in comparison to FEF. Thus, a broader range of frequencies is informative for distinguishing between SEF ensembles (in comparison to FEF). This is because SEF activity in the theta and alpha frequency bands is most informative in the largest spatial scale.

[5] Thresholding was done for visualisation purposes and we reached the same conclusion by just considering matrix norms without thresholding.

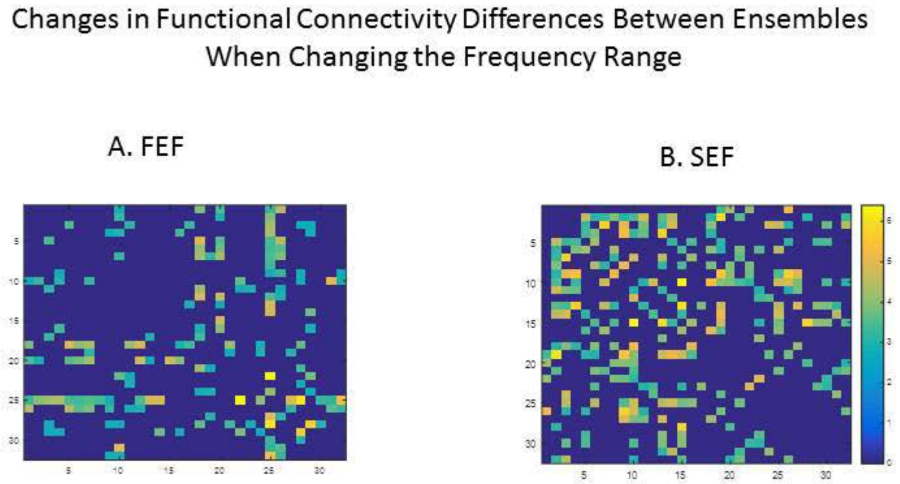


Figure 8. Changes in functional connectivity differences between ensembles when changing the frequency range. We computed such differences for a narrow (theta band) range and a broader range that included both theta and alpha bands. We found that a larger number of electrodes were sensitive to changes in the differences in functional connectivity between ensembles when we included the alpha band in SEF in comparison to FEF. Thus, a broader range of frequencies is informative for distinguishing between SEF ensembles.

Although most data features we used yielded results well above chance in both FEF and SEF, the maximum decoding accuracy obtained using FEF data features was higher than when using SEF features: the accuracy achieved for the second FEF connectivity component (0.28) was higher than the accuracy achieved for the first SEF component (0.22), cf. the leftmost bar in the top right panel *encircled* (Figure7B). The bottom panel in Figure 7 shows the results for PFC. The effects are weaker. However, the second component (*orange* curve in the panels of Figure 2C) resulted in accuracy above chance level when used for classification with Naïve Bayes; cf. the third bar from the right.

Note that we did not consider the connectivity strengths of more than four components as classification features. Doing so, might have led to higher accuracies but we were not after the best accuracy; rather we wanted to test whether using only the first few connectivity components might result in above chance classification accuracy. If so, ensemble activity can be described by a biophysical, neural field model that is computationally tractable. Our results suggest that this is indeed the case: classification accuracy did not dramatically change when we varied the number of components between 1-4: the mean accuracy and its standard deviation after using all possible classification features based on these components and both methods for each area (the mean and standard deviation when we consider all accuracy estimates in the panels of Figure 7) was as follows: 0.162 ± 0.07 for FEF, 0.150 ± 0.06 for SEF and 0.117 ± 0.05 for PFC respectively.

Between-area differences in topography and topology of ensemble connectivity

As said, entries of the connectivity components shown in Figure 2 are sums of connectivity weights. To describe the connectivity patterns underlying ensemble activity we used a neural field. Each recording site (electrode) is connected to all other recording sites with connections whose strength follows a Gaussian profile. However, neural fields (and similar biophysical models) are formulated in terms of connectivity weights, not their sums. Assuming a Gaussian connectivity profile, it is straightforward to obtain the connectivity weights from the components. The Gaussianity assumption is not restrictive; it is one of many possibilities, adopted here for the following three reasons: (i) Connectivity within the neural ensemble should aim to minimise metabolic cost. That is, it is metabolically advantageous to create connections between spatially adjacent network nodes (in the present case, electrodes) as opposed to connecting distant nodes together. This is the same fundamental constraint that has been shown to be crucial for explaining large scale connectivity between brain areas (Laughlin and Sejnowski, 2003; Sporns et al., 2004); (ii) It makes the model computationally tractable. The Gaussianity assumption reduces the problem of finding 32x32=1024 independent connectivity parameters (we are measuring activity from 32 electrodes) to finding only 64 parameters. (iii) It provides a simple, intuitive explanation of the parameters of the connectivity matrix. Using these matrices we fitted our neural field model to the data and obtained the corresponding coefficients of determination η^2 for all regions averaged over all stimuli and trials. This served as a validation of our model and the Gaussian assumption for the connectivity profile. The results are shown in Figure 9 (mean and 90% confidence intervals). Variance explained for both FEF and SEF were around 41%, while PFC showed a lower coefficient η^2, around 14%. This fits comfortably with the traits of topographic clustering in FEF and SEF observed earlier. Had we used a larger number of connectivity components (instead of only three) and determined each of the 1024 strengths independently, we might have got better model performance. However, our current focus was on obtaining a computationally tractable model that involves a small number of parameters. In future work, we will use these parameters as priors to fit Dynamic Causal Models (DCMs), see (Pinotsis et al., 2014, 2015). These models could explain more variance in the data due to the increased number of parameters and description of neurobiological properties of different brain areas.

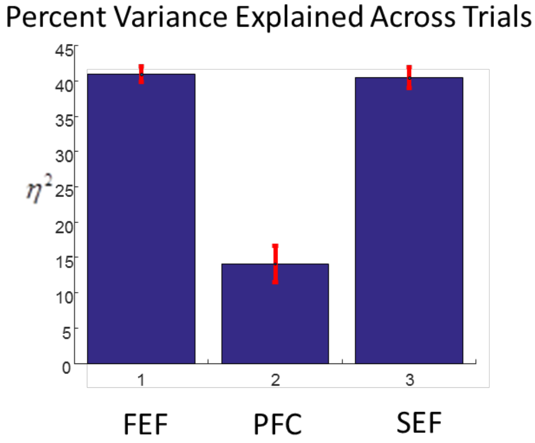


Figure 9. Percent variance explained after fitting our deep neural field model to data and averaging across all stimuli and trials.

Obtaining the connectivity matrices allowed us to address questions regarding cortical interactions at the microscale. The connectivity matrices can provide holistic measures for describing the large scale *topography* and *topology* of the observed patterns of LFP activity. Topography pertains to the spatial properties of LFP activity, for example, in which area does it occur, how extended it might be and how it is placed with respect to activity induced by different stimuli or in different tasks in the same or different brain areas. Topology on the other hand, is a notion borrowed from graph theory: Recording sites are thought to occupy the nodes of a graph and to be related to each other through edges that correspond to some pairwise relation measure – usually functional connectivity. Then the topology of cortical activity pertains to non–spatial aspects of relations between responses at different recording sites in terms of graph theoretic measures, like *characteristic path length* and *betweenness centrality* (Freeman, 1977). Each site can be connected to any other site in the ensemble in multiple ways that are called paths. One of them is the shortest path. Paths are defined for any pair of sites. Thus each site participates in many shortest paths. Characteristic path length refers to the *mean* shortest path length between all pairs of sites. Sites that participate in a larger number of shortest paths are more important for the robustness of the neural ensemble. These important sites are also called “central” sites. The term central is reminiscent of graph theory. Betweenness centrality is the ratio of shortest paths in which any given site participates over all shortest path lengths in the ensemble. First, we will discuss the topography of brain activity.

A Gaussian connectivity profile implies that the connection strength between two recording sites is described by two parameters: mean and dispersion. Dispersion is a measure of how extended the corresponding ensemble activity is. The Gaussian connectivity matrices for the three cortical areas and all cued locations are depicted in Supplemental Figure 2. We performed a two-way ANOVA using brain area and cued location (six possible cues) as factor on the mean and dispersion of the connectivity matrices for all areas and possible cues. For each ANOVA we looked at pairwise differences for each possible combination of two brain areas.

For SEF vs PFC, we found a significant main effect of area on connectivity dispersion, *$F(1, 888) = 100.27, p <0 .0001$*. Recall that the connectivity dispersion characterizes the extent of connections linking together neural ensembles across the cortical surface. The fact that we obtained a significant result for this main effect signifies a larger spread of *effective connections* in PFC in comparison to SEF. We did not find a significant effect of cue on dispersion, *$F(5, 888) = 0.6, p =0.7$* nor an interaction between cue and area factors *$F(5, 888)$*

= 1.67, p =0.15. Connectivity dispersion in PFC was larger (*C = 0.08 ± 0.004)* than SEF *(C = 0.14 ± 0.004)*, see Figure 10. A larger dispersion of PFC connectivity in comparison to SEF connectivity also means that PFC activity is more extended on the cortical surface in comparison to SEF activity. This fits nicely with an earlier result in section *Neural ensembles oscillate at theta, alpha and beta frequencies* above. There we found that power in PFC is concentrated in lower frequencies than in SEF. Had we not known that earlier result, we would have been able to predict it using the result we just found. Given that activity in PFC is more extended than in SEF, and that more extended activity might correspond to more power at lower frequencies (Leopold et al., 2003), it follows immediately that responses in PFC will be more extended than in SEF. This is an illustration of how the neural field model introduced here can link together –seemingly disconnected- spatial and temporal aspects of LFP responses.

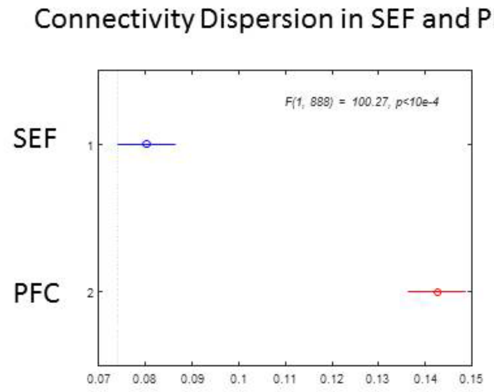


Figure 10. Differences in the average values of the mean and dispersion of Gaussian connectivity for SEF vs PFC using two-way ANOVA. Interactions between area and cue factors revealed a significant main effect of area on connectivity dispersion, *F(1, 888) = 100.27, p <0 .0001*.

We then turned to topological aspects of neuronal organization. We first calculated the characteristic path length for each of the three areas and all cues. The results are shown in Figure 11. In complex systems theory, paths in a network are considered as routes along which information propagates (Sporns, 2013). Characteristic path lengths are used to characterise how efficient information transfer within the network might be. Here, the characteristic path length of an ensemble describes how many processing steps, i.e. steps between recording sites, are required to maintain a cue in memory. Our use of path lengths is similar to their use in functional connectivity studies: instead of correlations, we here consider the effective connectivity weights obtained after training the neural field model as an auto-encoder. All results using graph theoretic measures below exploit the connectivity weight matrices found earlier after taking averages of connectivity components across all trials.

In FEF, we found a smaller characteristic path length for the neural ensembles that maintained cued locations on the *horizontal* axis. This means that a smaller number of processing steps is required for transmitting the information within those neural ensembles in comparison to cued locations on other directions. Also, information propagates faster when characteristic path lengths are smaller. This speaks to a general result in complex systems theory: a shorter characteristic path length implies that the network requires less energy and performs information processing faster (Laughlin and Sejnowski, 2003; Sporns et al., 2004). In short, our results suggest that information flow might be faster and metabolically more efficient for cued locations on the horizontal axis. The brain seems to respond preferentially to horizontal cues. This conclusion of our model speaks to a result in psychophysics known as *oblique effect* (Appelle, 1972) also observed in non-human primates (Bauer et al., 1979). This effect states that better performance is achieved for horizontal (and vertical) contours in visual perception tasks (the vertical axis was not tested in these data).

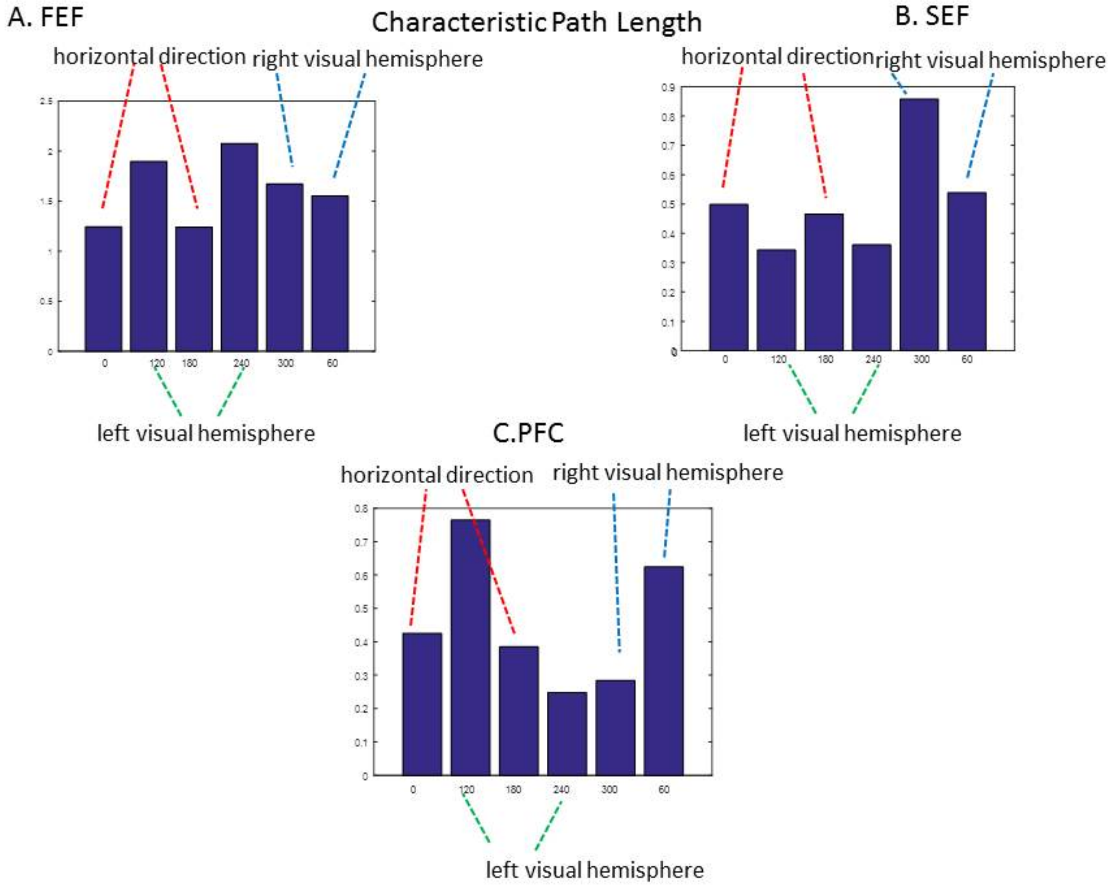


Figure 11. **A.** Characteristic path length estimates for all FEF ensembles and cued locations. These appear to partition the space of possible cued locations into locations on the horizontal (red dashed lines) and in the left (green dashed lines) and right (blue dashed lines) hemisphere. Also the smallest estimates are obtained for horizontal cued locations; this is reminiscent of the oblique effect. **B.-C.** Characteristic path length estimates for SEF and PFC. Estimates for both cued locations on the horizontal direction are similar.

We also found that characteristic path length values corresponding to other cued locations (not only the ones on the horizontal direction) also have behavioral relevance. Characteristic path length values appeared to partition the space of possible cued locations into three subsets: locations on the horizontal axis ($\theta_h = 0,180$ degrees) and in the left ($\theta_l = 120,240$ degrees) and right ($\theta_r = 60,300$ degrees), visual hemispheres. They were shortest for the horizontal axis, next shortest for locations in the right visual hemisphere, and longest for locations in the left visual hemisphere. The values of characteristic path length corresponding to each of the two cued locations within each of the above three subsets are shown with dashed lines (*red* for θ_h, *green* for θ_l and *blue* for θ_r). Note the similar path length values for cued locations within each subset in FEF. Also note that the characteristic path length values for each pair of locations on the horizontal meridian in all three areas are similar in value.

We also evaluated the betweenness centrality for all areas and cues and found central sites. Betweenness centrality provides an alternative way to characterise overlap in ensembles: if central sites were the same different between different ensembles, this means that different ensembles would overlap. Earlier, we found that different ensembles for different cues do not completely overlap by decoding ensemble responses (Figure 7). This can also be found using betweenness centrality. The results are shown in Figure 12.

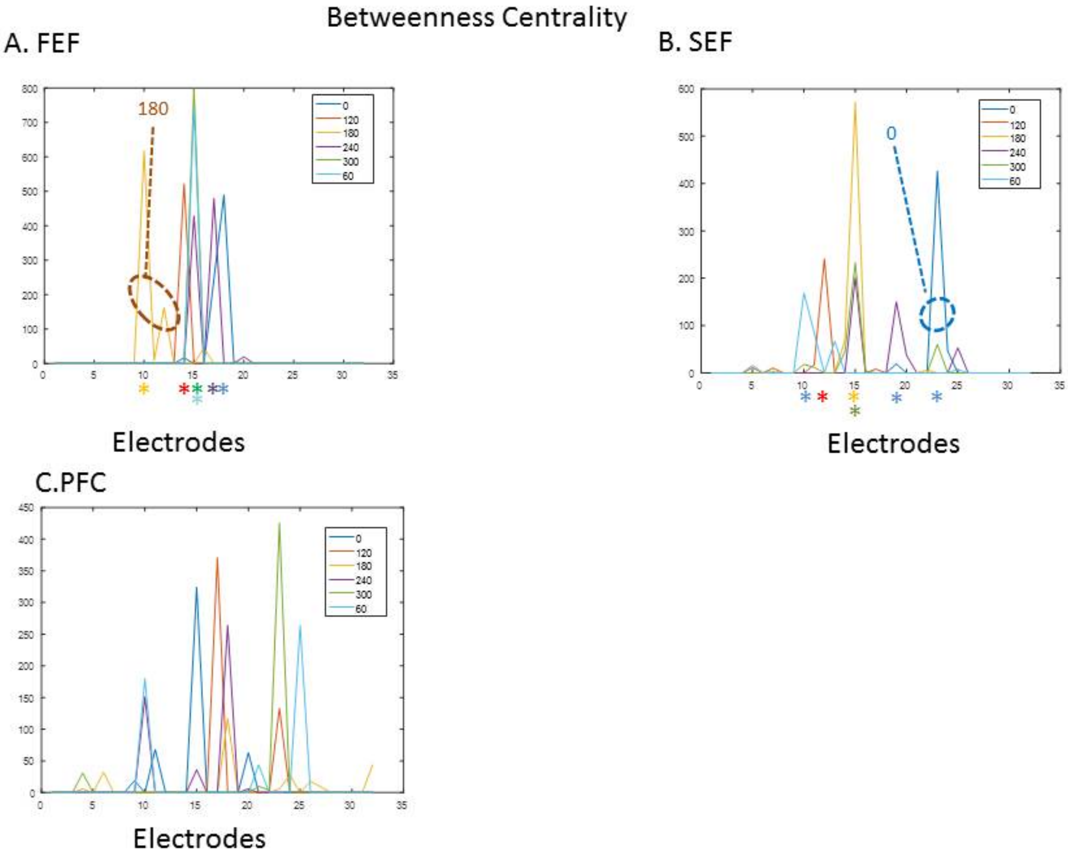


Figure 12. A. Betweenness centrality estimates for all FEF ensembles and cued locations. Neural ensembles do not overlap. Central sites are different for different cued locations. **B.** SEF betweenness centrality estimates. Similarly to FEF, SEF neural ensembles do not overlap. Notice that SEF central sites are similar to FEF central sites for the same cued location. **C.** PFC betweenness centrality estimates. Neural ensembles overlap. Central sites are the same for multiple cued locations.

Let us e.g. focus on FEF betweenness centrality values depicted in Figure 12A. Values corresponding to a cue at $\theta = 180$ degrees are shown by *brown* dashed lines and ellipses. These describe a neural ensemble comprising electrodes (9-13) on the left end of the figure. Electrode 10 is central in this ensemble. Different electrodes exhibited high betweenness centrality values for distinct cued locations: electrodes 14, 15, 17 and 18 showed peak betweenness centrality values for cued locations at $\theta = 120, 300$ or 60, 240 and 0 degrees respectively. Stars of different colours denote these central electrodes. So, with the exception of electrode 15 the rest of the central electrodes correspond to different cued locations in a unique manner. The exception of electrode 15 is not surprising: in the deep neural field model, this electrode is placed in the middle of the cortical manifold, therefore its role as a hub of information flow for the maintenance of more than one cues is expected.

In SEF, we found similar results to FEF: we observed a neural ensemble of electrodes (22-25) on the right end of the figure that exhibits high values of betweenness centrality for $\theta = 0$ degrees (*blue* dashed lines and ellipses in Figure 12B). Again, we found distinct central electrodes for different cued locations (shown by coloured stars): electrodes 10, 12, 15, 19 and 23 showed peak betweenness centrality values for cued locations at $\theta = 60, 120, 300$ or 180, 240 and 0 degrees respectively. Interestingly, there appears to be some homology between electrodes for which betweenness centrality values peaked for the same cued locations ($\theta = 120, 300, 240$ and 0 degrees) in both areas: electrode 15 for $\theta = 300$ and then electrodes 14,17 and 18 in FEF as opposed to 12,19 and 23 in SEF for $\theta = 120, 240$ and 0 degrees respectively. This means that there might be some structure in neural ensembles that is conserved across brain areas and might relate to topographic clustering we observed in the two areas. All in all, using graph theoretic measures, we found non-overlapping stimulus-specific neural ensembles in FEF and SEF that maintain behaviourally relevant information efficiently. In PFC, the situation is different from the other two areas: neural ensembles based on betweenness centrality appear to overlap more. This is in accord with earlier results of lowest decoding accuracy achieved in PFC (Figure 7C). Many electrodes exhibited high betweenness centrality values for more than one cued locations and different neural ensembles obtained using betweenness centrality appeared to overlap with each other (Figure 12C).

Discussion

We presented an approach for identifying and describing neural ensembles. Our approach applies ideas from statistical learning in neuroscience (Fiser et al., 2010; Hong et al., 2016, 2016; Jazayeri and Movshon, 2006; Tacchetti et al., 2016; Tsai and Cox, 2015) and is based on a neural field model with learned parameters (Olshausen, 1996; Simoncelli and Olshausen, 2001).

We focused on the connectivity patterns that underlie ensemble responses. The strengths of these connections (connectivity weights) were inferred by analysing LFPs recorded during a spatial working memory task. One motivation for focusing on connectivity weights is recent evidence that working memories are stored in connectivity weights as opposed to neural activity *per se* (Lundqvist et al., 2011, 2016; Stokes, 2015). The connectivity weights measured the *effective* connectivity between recording sites. Effective connectivity is a particular sort of functional connectivity where connectivity estimates appear as parameters in a model of observed brain activity. Here, these estimates appear in a neural field model. Traditionally, effective connectivity has been used to describe *mesoscale* interactions between brain areas using neuroimaging (David et al., 2006; Friston, 1994). To the best of our knowledge this is the first use of effective connectivity to describe microscale interactions underlying neural ensemble dynamics.

To obtain the connectivity weights, we trained a neural field model as a particular type of deep neural network, called an auto-encoder using raw LFP data. Thus, our approach provides links between unsupervised learning literature and biophysical modeling. It is the first illustration of how a neural field model can be used as an auto-encoder network that can learn and maintain its inputs. We call this model a *deep* neural field. As it is common, we use the term “deep” to describe a neural network architecture with multiple layers, like an auto-encoder. Deep neural fields maximise the mutual information between the remembered cue and the ensemble activity. Therefore, the connectivity weights we obtained are *optimal* in an information-theoretic sense.

Computing the connectivity weights allowed us to address questions regarding the cortical micro-circuitry and neuronal interactions. We found that the graph theoretic measures based on effective connectivity at the microscale captured behaviorally relevant information. Characteristic path length estimates in FEF appeared to partition the space of possible cues into those located in the left and right visual field. In addition, the characteristic path lengths were smaller for cues on the horizontal direction, that is, at an angle of 0 or 180 degrees. This means that propagation of information flow within the corresponding neural ensembles occurred faster and engaged shorter paths when storing memories of these cues. The brain seems more efficient on processing stimuli on the horizontal meridian (the *oblique effect* - Bauer et al., 1979). This in turn might be the result of more parsimonious micro-circuitry.

We also obtained betweenness centrality estimates. These revealed that neural ensembles corresponding to different cued locations do not completely overlap in FEF and SEF (that is, we could detect distinct ensembles or each). This was confirmed using brain decoding algorithms that used patterns of ensemble activity to predict the cued locations held in memory. We concluded this was possible and found the best performance for FEF ensembles. This may be related to the evidence we found for topographic clustering in the effective connectivity between FEF (and to a lesser extent, SEF) recording sites. Topographic clustering explained the structure of the ensemble connectivity and could also explain a homology between FEF and SEF central sites (recording sites that showed maximum betweenness centrality values) corresponding to the same cued locations. These central sites were near one another within FEF and SEF electrode arrays. They were indexed by adjacent electrode numbers, e.g. electrodes 17 and 19 were central for cued locations at 240 degrees in FEF and SEF respectively. This homology might be explained as follows. If topographic clustering is due to retinotopy, the way FEF and SEF responses are spatially distributed on the cortical surface would be similar. The way these responses are sampled by our multi-electrode arrays could also be similar: electrodes in FEF and SEF arrays are numbered in a similar (monotonic) fashion. Of course, the two arrays did not necessarily have the same orientation when placed on the cortical surface. However, such differences were not captured by the deep neural field model we used here and did not affect the betweenness centrality estimates of our model: the spatial variable appearing in the deep neural field is one-dimensional. All in all, the similarity in the patterns of activity of neural ensembles corresponding to the same stimulus and the way activity is sampled in the two regions could explain the homology between the corresponding central sites.

Finally, we considered oscillatory responses of neural ensembles recorded from different brain areas (Jones, 2016; Schroeder and Lakatos, 2009). Theta rhythms seemed to be prominent in FEF and SEF and were expressed at the largest spatial scale. This is in accord with other studies where theta activity has been shown to be associated with spatial coding both in PFC (Jones and Wilson, 2005) and hippocampus (O'Keefe and Burgess, 2005). We also found beta rhythms. Beta activity has been observed in PFC (Antzoulatos and Miller, 2016; Buschman et al., 2012; Kawasaki and Yamaguchi, 2012; Lundqvist et al., 2016; Stanley et al., 2016). Our approach further revealed connectivity across different spatial scales for theta vs beta frequencies. In FEF, beta was expressed in smaller spatial scales and had weaker functional connectivity than theta. In PFC, power was concentrated in lower frequencies and had larger connectivity dispersion, i.e. extent (spread) of connections compared to SEF. This is consistent with observations that power in lower frequencies is more extended on the cortical surface than power in higher frequencies (Leopold et al., 2003).

Although based on neural fields, our approach deviates significantly from existing work with such and similar biophysical models, see e.g. (Spencer, 2009). To date, neural fields have been used to 1. *simulate* activity e.g. during sleep, anaesthesia (Hutt, 2013; Steyn-Ross et al., 2013; Suder et al., 2001) or cognitive tasks (Beim Graben et al., 2008; Potthast and Graben, 2009;Lipinski et al., 2009, 2012); 2. *fit* real data (Freestone et al., 2011; Pinotsis et al., 2012a, 2014). A common assumption in all previous work is that connectivity matrices are *homogeneous.* This simply means that connectivity in these models depends only on the distance between neurons in the model. This is also the case in related approaches like ring models, recurrent and convolutional neural networks (Ben-Yishai et al., 1997; Shriki and Yellin, 2016; Somers et al., 1995). In this work, connectivity weights are often postulated *ad hoc*. Here, we deviated from these assumptions.

We considered *inhomogeneous* connectivity matrices and showed how these can be obtained after training a neural field as an auto-encoder. Our approach yields connection weights that contain information about the particular cue maintained in memory. From a statistical learning perspective, these are learned parameters. From a neurophysiological perspective, they can be thought of as synaptic weights in an animal that has learned to remember the cues. Thus, our approach provides a principled way to obtaining learned connectivity parameters instead of setting them by hand. This paves the way for detailed biophysical models with learned connectivity parameters. Similarly to the deep neural field model introduced here, these models will be able to predict the dynamics of neural ensembles representing different stimuli. They have the potential to explain differences in brain activity associated with stimuli of different values or levels (e.g. cued locations or colours). Importantly, these models could reveal the principles underlying the biophysics and information processing in neural ensembles. These advances could also guide the future development of Brain Computer Interfaces and neuroprosthetics (Nicolelis and Lebedev, 2009).

References

Antzoulatos, E.G., and Miller, E.K. (2016). Synchronous beta rhythms of frontoparietal networks support only behaviorally relevant representations. ELife *5*, e17822.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. Psychological Bulletin *78*, 266.

Atay, F.M., and Hutt, A. (2006). Neural fields with distributed transmission speeds and long-range feedback delays. SIAM Journal on Applied Dynamical Systems *5*, 670–698.

Bauer, J.A., Owens, D.A., Thomas, J., and Held, R. (1979). Monkeys show an oblique effect. Perception *8*, 247–253.

Beim Graben, P., Pinotsis, D., Saddy, D., and Potthast, R. (2008). Language processing with dynamic fields. Cognitive Neurodynamics *2*, 79–88.

Ben-Yishai, R., Hansel, D., and Sompolinsky, H. (1997). Traveling waves and the processing of weakly tuned inputs in a cortical network module. Journal of Computational Neuroscience *4*, 57–77.

Bertero, M., De Mol, C., and Pike, E.R. (1988). Linear inverse problems with discrete data: 11. Stability and regularisation. Inverse Problems *4*, 573–594.

Botvinick, M.M., and Plaut, D.C. (2006). Short-term memory for serial order: a recurrent neural network model. Psychological Review *113*, 201.

Breakspear, M., and Jirsa, V. (2007). Neuronal dynamics and brain connectivity. Handbook of Brain Connectivity 3–64.

Breakspear, M., Roberts, J.A., Terry, J.R., Rodrigues, S., Mahant, N., and Robinson, P.A. (2006). A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. Cerebral Cortex *16*, 1296–1313.

Bressloff, P.C. (2010). Metastable states and quasicycles in a stochastic Wilson-Cowan model of neuronal population dynamics. Physical Review E *82*, 051903.

Brown, E.N., Frank, L.M., Tang, D., Quirk, M.C., and Wilson, M.A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. The Journal of Neuroscience *18*, 7411–7425.

Brunel, N., and Wang, X.-J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. Journal of Computational Neuroscience *11*, 63–85.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., and Miller, E.K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron *76*, 838–846.

Coombes, S., and Owen, M.R. (2004). Evans functions for integral neural field equations with Heaviside firing rate function. SIAM Journal on Applied Dynamical Systems *3*, 574–600.

David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., and Friston, K.J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage *30*, 1255–1272.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage *43*, 44–58.

Deco, G., Jirsa, V.K., Robinson, P.A., Breakspear, M., and Friston, K. (2008). The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields.

Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. Nature Neuroscience *10*, 1241–1242.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in Cognitive Sciences *14*, 119–130.

Formisano, E., De Martino, F., and Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magnetic Resonance Imaging *26*, 921–934.

Freeman, L.C. (1977). A set of measures of centrality based on betweenness. Sociometry 35–41.

Freestone, D.R., Aram, P., Dewar, M., Scerri, K., Grayden, D.B., and Kadirkamanathan, V. (2011). A data-driven framework for neural field modeling. NeuroImage *56*, 1043–1058.

Fries, P., Nikolić, D., and Singer, W. (2007). The gamma cycle. Trends in Neurosciences *30*, 309–316.

Friston, K. (2008). Hierarchical models in the brain. PLoS Comput Biol *4*, e1000211.

Friston, K.J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. Human Brain Mapping *2*, 56–78.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. Journal of Neurophysiology *63*, 814–831.

Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Current Opinion in Neurobiology *37*, 66–74.

Fuster, J.M., Bauer, R.H., and Jervey, J.P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. Brain Research *330*, 299–307.

Gray, C.M. (1999). The temporal correlation hypothesis of visual feature integration: still alive and well. Neuron *24*, 31–47.

Grindrod, P., and Pinotsis, D.A. (2011). On the spectra of certain integro-differential-delay problems with applications in neurodynamics. Physica D: Nonlinear Phenomena *240*, 13–20.

Haegens, S., Nácher, V., Hernández, A., Luna, R., Jensen, O., and Romo, R. (2011). Beta oscillations in the monkey sensorimotor network reflect somatosensory decision making. Proceedings of the National Academy of Sciences *108*, 10708–10713.

Hansel, D., and Sompolinsky, H. (1998). 13 Modeling Feature Selectivity in Local Cortical Circuits.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association *72*, 320–338.

Hebb, D.O. (1949). The organization of behavior: A neuropsychological approach (John Wiley & Sons).

Heinz, S. (2013). Statistical mechanics of turbulent flows (Springer Science & Business Media).

Hong, H., Yamins, D.L., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nature Neuroscience *19*, 613–

622.

Hutt, A. (2013). The anesthetic propofol shifts the frequency of maximum spectral power in EEG during general anesthesia: analytical insights from a linear model.

Jazayeri, M., and Movshon, J.A. (2006). Optimal representation of sensory information by neural populations. Nature Neuroscience *9*, 690–696.

Jia, N., Brincat, S., Salazar-Gómez, A., Panko, M., Guenther, F., and Miller, E. (2017). Decoding of intended saccade direction in an oculomotor brain-computer interface. Journal of Neural Engineering.

Jirsa, V.K., and Haken, H. (1996). Field theory of electromagnetic brain activity. Physical Review Letters *77*, 960.

Jirsa, V., Sporns, O., Breakspear, M., Deco, G., and McIntosh, A.R. (2010). Towards the virtual brain: network modeling of the intact and the damaged brain. Archives Italiennes de Biologie *148*, 189–205.

Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. The Journal of Neuroscience *27*, 12176–12189.

Jones, S.R. (2016). When brain rhythms aren't 'rhythmic': implication for their mechanisms and meaning. Current Opinion in Neurobiology *40*, 72–80.

Jones, M.W., and Wilson, M.A. (2005). Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task. PLoS Biol *3*, e402.

Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D.L., and Carandini, M. (2009). Local origin of field potentials in visual cortex. Neuron *61*, 35–41.

Kawasaki, M., and Yamaguchi, Y. (2012). Individual visual working memory capacities and related brain oscillatory activities are modulated by color preferences. Frontiers in Human Neuroscience *6*.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pp. 1097–1105.

Laing, C.R., and Chow, C.C. (2001). Stationary bumps in networks of spiking neurons. Neural Computation *13*, 1473–1494.

Laing, C.R., and Troy, W.C. (2003). PDE methods for nonlocal problems. SIAM Journal of Dynamical Systems *2*, 487–516.

Laughlin, S.B., and Sejnowski, T.J. (2003). Communication in neuronal networks. Science *301*, 1870–1874.

Leopold, D.A., Murayama, Y., and Logothetis, N.K. (2003). Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. Cerebral Cortex *13*, 422–433.

Lipinski, J., Sandamirskaya, Y., and Schöner, G. (2009). Swing it to the left, swing it to the right: enacting flexible spatial language using a neurodynamic framework. Cognitive Neurodynamics *3*, 373–400.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J.P., and Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. Journal of Experimental Psychology: Learning, Memory, and Cognition *38*, 1490.

Liu, X., Ramirez, S., Pang, P.T., Puryear, C.B., Govindarajan, A., Deisseroth, K., and Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. Nature *484*, 381–385.

Lundqvist, M., Herman, P., and Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. Journal of Cognitive Neuroscience *23*, 3008–3020.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and beta bursts underlie working memory. Neuron *90*, 152–164.

Martens, E.A., Panaggio, M.J., and Abrams, D.M. (2016). Basins of attraction for chimera states. New Journal of Physics *18*, 022002.

Mersmann, A. (1995). Crystallization technology handbook. Drying Technology *13*, 1037–1038.

Miller, E.K., and Buschman, T.J. (2013). Brain rhythms for cognition and consciousness. Neurosciences and the Human Person: New Perspectives on Human Activities 1–11.

Misaki, M., Kim, Y., Bandettini, P.A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage *53*, 103–118.

Modi, M.N., Dhawale, A.K., and Bhalla, U.S. (2014). CA1 cell activity sequences emerge after reorganization of network correlation structure during associative learning. Elife *3*, e01982.

Neal, R.M., and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in Graphical Models, (Springer), pp. 355–368.

Nicolelis, M.A., and Lebedev, M.A. (2009). Principles of neural ensemble physiology underlying the operation of brain–machine interfaces. Nature Reviews Neuroscience *10*, 530–540.

O'Keefe, J., and Burgess, N. (2005). Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. Hippocampus *15*, 853–866.

Pinotsis, D.A., Moran, R.J., and Friston, K.J. (2012a). Dynamic causal modeling with neural fields. Neuroimage *59*, 1261–1274.

Pinotsis, D.A., Moran, R.J., and Friston, K.J. (2012b). Dynamic causal modeling with neural fields. Neuroimage *59*, 1261–1274.

Pinotsis, D.A., Hansen, E., Friston, K.J., and Jirsa, V.K. (2013). Anatomical connectivity and the resting state activity of large cortical networks. Neuroimage *65*, 127–138.

Pinotsis, D.A., Brunet, N., Bastos, A., Bosman, C.A., Litvak, V., Fries, P., and Friston, K.J. (2014). Contrast gain control and horizontal interactions in V1: a DCM study. Neuroimage *92*, 143–155.

Pinotsis, D.A., Leite, M., and Friston, K.J. (2015). On conductance-based neural field models. Frontiers in Computational Neuroscience *7*.

Pinto, D.J., and Ermentrout, G.B. (2001). Spatially structured activity in synaptically coupled neuronal networks: I. Traveling fronts and pulses. SIAM Journal on Applied Mathematics *62*, 206–225.

Potthast, R., and Graben, P.B. (2009). Inverse problems in neural field theory. SIAM Journal on Applied Dynamical Systems *8*, 1405–1433.

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

Robinson, P.A., Sarkar, S., Pandejee, G.M., and Henderson, J.A. (2014). Determination of effective brain connectivity from functional connectivity with application to resting state connectivities. Physical Review E *90*, 012707.

Rubin, J.E., and Troy, W.C. (2004). Sustained spatial patterns of activity in neuronal populations without recurrent excitation. SIAM Journal on Applied Mathematics *64*, 1609–1635.

Ryan, T.J., Roy, D.S., Pignatelli, M., Arons, A., and Tonegawa, S. (2015). Engram cells retain memory under retrograde amnesia. Science *348*, 1007–1013.

Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth Annual Conference of the International Speech Communication Association, p.

Schroeder, C.E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences *32*, 9–18.

Shriki, O., and Yellin, D. (2016). Optimal Information Representation and Criticality in an Adaptive Sensory Recurrent Neuronal Network. PLoS Comput Biol *12*, e1004698.

Somers, D.C., Nelson, S.B., and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. The Journal of Neuroscience *15*, 5448–5465.

Spencer, K.M. (2009). The functional consequences of cortical circuit abnormalities on gamma oscillations in schizophrenia: insights from computational modeling. Frontiers in Human

Neuroscience *3*, 33.

Sporns, O. (2013). The human connectome: origins and challenges. Neuroimage *80*, 53–61.

Sporns, O., Chialvo, D.R., Kaiser, M., and Hilgetag, C.C. (2004). Organization, development and function of complex brain networks. Trends in Cognitive Sciences *8*, 418–425.

Stanley, D.A., Roy, J.E., Aoi, M.C., Kopell, N.J., and Miller, E.K. (2016). Low-Beta Oscillations Turn Up the Gain During Category Judgments. Cerebral Cortex.

Steyn-Ross, M.L., Steyn-Ross, D.A., and Sleigh, J.W. (2013). Interacting Turing-Hopf instabilities drive symmetry-breaking transitions in a mean-field model of the cortex: a mechanism for the slow oscillation. Physical Review X *3*, 021005.

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

Suder, K., Wörgötter, F., and Wennekers, T. (2001). Neural field model of receptive field restructuring in primary visual cortex. Neural Computation *13*, 139–159.

Tacchetti, A., Isik, L., and Poggio, T. (2016). Spatio-temporal convolutional neural networks explain human neural representations of action recognition. ArXiv Preprint ArXiv:1606.04698.

Tsai, C.-Y., and Cox, D.D. (2015). Measuring and understanding sensory representations within deep networks using a numerical optimization framework. ArXiv Preprint ArXiv:1502.04972.

Tsodyks, M.V., and Sejnowski, T. (1995). Rapid state switching in balanced cortical network models. Network: Computation in Neural Systems *6*, 111–124.

Vaina, L.M., Soloviev, S., Calabro, F.J., Buonanno, F., Passingham, R., and Cowey, A. (2014). Reorganization of retinotopic maps after occipital lobe infarction. Journal of Cognitive Neuroscience *26*, 1266–1282.

Valenti, A.P., Brady, M., Scheutz, M.J., Holcomb, P.J., and Pu, H. A Neural Field Model of Word Repetition Effects in Early Time-Course ERPs in Spoken Word Perception.

Wei, Z., Wang, X.-J., and Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. The Journal of Neuroscience *32*, 11228–11240.

Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J. (2016). Data Mining: Practical machine learning tools and techniques (Morgan Kaufmann).