NeuroImage 253 (2022) 119058

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Beyond dimension reduction: Stable electric fields emerge from and allow representational drift

Dimitris A. Pinotsis[a,b,*], Earl K. Miller[b]

[a] Centre for Mathematical Neuroscience and Psychology and Department of Psychology, City-University of London, London EC1V 0HB, United Kingdom
[b] The Picower Institute for Learning and Memory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Keywords:
Memory engrams
Neural ensembles
Working memory
Predictive coding
Auto-encoders
Effective connectivity

ABSTRACT

It is known that the exact neurons maintaining a given memory (the neural ensemble) change from trial to trial. This raises the question of how the brain achieves stability in the face of this representational drift. Here, we demonstrate that this stability emerges at the level of the electric fields that arise from neural activity. We show that electric fields carry information about working memory content. The electric fields, in turn, can act as “guard rails” that funnel higher dimensional variable neural activity along stable lower dimensional routes. We obtained the latent space associated with each memory. We then confirmed the stability of the electric field by mapping the latent space to different cortical patches (that comprise a neural ensemble) and reconstructing information flow between patches. Stable electric fields can allow latent states to be transferred between brain areas, in accord with modern engram theory.

1. Introduction

In the era of large scale electrophysiology (Steinmetz et al., 2018), neural recordings of high dimensionality are abundant. Yet this has revealed that brain areas seem to exchange information in low dimensions, using few task-related variables (latent variables) (Katlowitz et al., 2018). Indeed, brain dynamics evolve in low, not high, dimensional spaces (Gallego et al., 2018; Mastrogiuseppe and Ostojic, 2018). These spaces are found by dimensionality reduction, (Jazayeri and Ostojic, 2021; Urai et al., 2021). Low dimensionality underlies a variety of cognitive and motor tasks (Cunningham and Byron, 2014; Pang et al., 2016).

A key point is that low-dimension latent variables track information and task demands and are stable, highly correlated across trials (Pandarinath et al., 2018). This stands in contrast to higher-dimensional neural dynamics; while there is some overlap (Churchland et al., 2010), the specific neurons and synapses activated are variable across trials (Mongillo et al., 2017; Attardo et al., 2015; Ziv and Brenner, 2018). This appears paradoxical: which specific neurons are activated continuously changes, synapses rewire etc., yet at the functional/behavioral, stability comes from low dimensional, latent variables (Clopath et al., 2017; Lu and Zuo, 2021; Kozachkov et al., 2020). This low dimensional stability is important for normal cognition and behavior. Downstream neurons and networks need some consistency from upstream networks even though those upstream networks are under continuous reconfiguration.

The continuous reconfiguration is known as representational drift (Driscoll et al., 2017). It occurs at a time scale of days, minutes or seconds (Deitch et al., 2020). It helps ensure the robustness of brain circuits. If some neurons fail, others can do the same task (Marder et al., 2015). Plus, neurons, especially in higher cortical areas, have mixed selectivity which adds computational horsepower and cognitive flexibility (Fusi et al., 2016; Rigotti et al., 2013). Representational drift may also be important for the brain computations needed for Predictive Coding (Rule et al., 2019) and Reinforcement Learning (Kappel et al., 2018). But the biophysical mechanism that allows low-dimensional brain dynamics to emerge despite the representational drift, is still a mystery.

Here, we suggest that this low dimensional stability is an emergent property of the electric fields generated by neural activity. Consider the following: First, that ensembles are functionally integrated within larger brain networks (Park and Friston, 2013; Shine et al., 2016; Westphal et al., 2017). Networks must somehow represent the same memory at different times even though larger networks in which they embedded are in different states at different times. Given this fluctuating network activity, it is difficult to imagine how that memory could be represented by a specific set of neurons and connections, even if one assumes redundancy. Second, different combinations of electric sources can generate the same field (Jackson, 1999). Taken together, the above

[*] Corresponding author at: Centre for Mathematical Neuroscience and Psychology and Department of Psychology, City-University of London, London EC1V 0HB, United Kingdom.
E-mail address: pinotsis@mit.edu (D.A. Pinotsis).

https://doi.org/10.1016/j.neuroimage.2022.119058.
Received 5 January 2022; Received in revised form 3 March 2022; Accepted 3 March 2022
Available online 8 March 2022.
1053-8119/

two facts suggest that a changing input from the rest of the brain leads to a reconfiguration of the ensemble so that a stable electric field is maintained. Thus, a stable electric field level emerges from a high-dimensional representational drift of specific neurons.

It may help to consider the following analogy: Brain anatomy is like the road-and-highway system. It is where traffic could go. Current thoughts, memories etc. are the patterns of traffic at that moment. An exact network of specific neurons is one particular route through the road-and-highway system. But, importantly, the same destination can be reached by taking different routes at different times (i.e., representational drift). What really matters are the general patterns of the traffic, *not* the exact roads it takes. There are multiple ways to travel from location A to location B.

This motivates the following hypothesis: That ensemble representation at the electric field level is more robust and less variable than representation at the level of specific neurons and circuits. If true, this could explain how low-dimensional stable computations arise despite representational drift.

Here, we tested whether this hypothesis is supported by data from a spatial delayed saccade task. We characterized the stability of both the electric field and of the neural activity that generates this field. The same data were earlier used to build brain computer interfaces (Jia et al., 2017) and provide a neurobiological explanation of the oblique effect (Pinotsis et al., 2017). We here used them to train a biophysical neural network model as an autoencoder that learned to maintain spatial locations. This gave us the latent space similarly to other dimensionality reduction approaches (Pang et al., 2016; Gao and Ganguli, 2015). Then, we went one step further. We obtained single trial estimates of effective connectivity between different neurons. These describe how information propagates over a cortical patch occupied by the neural ensemble; and how neurons communicate via electric signals sent from one part of the patch to the other. This is a difference between our approach and other approaches. Our approach maps the latent space to a cortical patch. It goes beyond dimensionality reduction and reconstructs information flow.

Following (Pinotsis et al., 2017), we reconstructed the effective connectivity between neurons on the patch from the latent space obtained earlier. These connectivity estimates describe the exchange of electric signals within the ensemble. We found that they correlated with clusters found using pairwise correlations (Humphries, 2011). This extra step also allowed us to reconstruct the electric field produced by the ensemble. Having a detailed description of electric signals and neural activity within the patch, we computed the electric field near it, using a classic dipole model from electromagnetism (Schwartz et al., 2016). To sum up, we predicted neural activity and the electric field generated each time (trial) the same location had to be remembered. Then, we tested if they were the same across trials. We found that the electric field was different for different remembered locations and highly consistent across trials. It also contained stable information about the remembered locations, while specific neurons activated were variable across trials (representational drift).

2. Methods

2.1. Experimental data and recording setup

We reanalyzed data from (Jia et al., 2017). The same data were used in our earlier paper (Pinotsis et al., 2017). 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 (Supplementary Fig. 1B). 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 angle, 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 (Supplementary Fig. 1C). Each array consisted of a 2 × 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 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 angles in memory.

2.2. From the Wilson Cowan equations to deep neural fields

Below we derive the evolution equations for a biophysical neural network model whose connectivity parameters have been obtained after training it as an autoencoder. This describes the activity of a neural ensemble. Its connectivity is such that the mutual information between the remembered cue and the ensemble activity is maximized. The corresponding weights are optimal in an information-theoretic sense.

Consider a neural ensemble that consists of neurons occupying a cortical patch (two dimensional Euclidean manifold) M_A. Let u_a, v_a be two spatial variables parameterizing a M_A, $(u_a, v_a) \in M_A$, see e.g. (Ermentrout and Cowan, 1979; Pinotsis et al., 2012; Wilson and Cowan, 1973). Let $x^a_E(u_a, v_a, t)$ and $x^b_I(u_b, v_b, t)$ be the membrane potential of excitatory neurons and inhibitory neurons at locations (u_a, v_a) and (u_b, v_b) on the cortical surface and time t. The time evolution of $x^a_E(u_a, v_a, t)$ and $x^b_I(u_b, v_b, t)$ is given by the following neural network equations, known as the Wilson-Cowan Equations (Wilson and Cowan, 1973, Grossberg, 1967)

$$\begin{aligned}\dot{x}^a_E(u_a, v_a, t) &= -\tau_E x^a_E(u_a, v_a, t) + \sum_c K_{EE}(u_a, v_a, u_c, v_c) f\left[x^c_E(u_c, v_c, t)\right] \\ &\quad + \sum_d K_{EI}(u_a, v_a, u_d, v_d) f\left[x^d_I(u_d, v_d, t)\right] + S \circ U_E, \\ \dot{x}^b_I(u_b, v_b, t) &= -\tau_I x^b_I(u_b, v_b, t) + \sum_c K_{II}(u_b, v_b, u_c, v_c) f\left[x^c_I(u_c, v_c, t)\right] \\ &\quad + \sum_d K_{IE}(u_b, v_b, u_d, v_d) f\left[x^d_E(u_d, v_d, t)\right] + S \circ U_I, \end{aligned} \tag{1}$$

where $S : (\mathbb{R})^n \to (\mathbb{R})^n$ maps exogenous inputs to depolarization and f is vector-valued transfer function that describes the mapping from membrane potentials to current (spikes per second; Lipschitz continuous to guarantee local existence) of the population around point $(u_a, v_a) \in M_A$.

We then take the continuum limit of Equations (1). This is a common transformation of biophysical evolution equations (van Hemmen, 2004) and allows one to replace sums with integrals. It follows a standard process in mathematical physics that provides the continuous version of a discrete system (opposite of discretization). We then partition M_A into $N \times L$ cortical patches of neural densities ζ^a_{ij} with dimensions $(\Delta v, \Delta v)$ $i \in 1, ..., N$ and $j \in 1, ..., L$. Thus the subgroup of neurons in the square $T^a_{ij} = \{[i\Delta u, (i+1)\Delta u), [j\Delta v, (j+1)\Delta v)\}$ of M_A is given by $\rho^a_{ij} = \zeta^a_{ij}\Delta v \Delta v$. For mathematical convenience, consider a copy M_B of manifold M_A. The interaction between neurons in cortical patches $T^a_{ij} \in M_A$ and $T^b_{kl} \in M_B$ only depends on the duplets (i, j) and (k, l). A neuron at location (u_a, v_a) inside square T^a_{ij} receives input from all neurons in square T^b_{kl} with strength $K_{PP'}(i, j, k, l) = \tilde{K}_{PP'}(i\Delta u, j\Delta v, k\Delta u', l\Delta v')$, $P, P' = \{E, I\}$, where we use " ′ " to denote locations on manifold M_B. Also, $\tilde{K}$ is the continuous version of function K under the assumption that connectivity is constant within the square with sides of length Δv and Δv. For simplicity of notation, in the following we write K in place of $\tilde{K}$. Then, we can define the local spatially averaged activity variable X_P by $X_P(i\Delta u, j\Delta v, t) = (\rho^a_{ij})^{-1} \sum_{(i,j)\in T^a_{ij}} x_P(i, j, t)$ and consider the continuum limit $\Delta u, \Delta v, \Delta u', \Delta v' \to 0$: all nodes within the patches T^a_{ij} and T^a_{ij} occupy the same location in manifolds M_A and M_B. After replacing $u = i\Delta u$,

$v = j\Delta v$ and $u' = k\Delta u'$, $v' = l\Delta v'$, Equations (1) can be written as a system

$$\begin{aligned}\dot{X}_E(u,v,t) &= -\tau_E X_E(u,v,t) + \iint_{M_B} K_{EE}(u,v,u',v') f\left[X_E(u',v',t)\right] du'dv' \\ &+ \iint_{M_B} K_{EI}(u,v,u',v') f\left[X_I(u',v',t)\right] du'dv' + S\circ U_E \\ \dot{X}_I(u,v,t) &= -\tau_I X_I(u,v,t) + \iint_{M_B} K_{II}(u,v,u',v') f\left[X_I(u',v',t)\right] du'dv' \\ &+ \iint_{M_B} K_{IE}(u,v,u',v') f\left[X_E(u',v',t)\right] du'dv' + S\circ U_I \qquad (2)\end{aligned}$$

Similarly to (Pinotsis et al., 2017), we then consider perturbations $\hat{X}_P$ of membrane potentials around baseline: $X_P(u,v,t) = X_{0P}(u,v) + \hat{X}_P(u,v,t)\ \ P = \{E, I\}$. This yields an expression of the perturbations$\hat{X}_P(u,v,t)$ in terms of: (1) the functions G_k, which we previously called *principal axes* (Pinotsis et al., 2017)*;* and (2) the latent variables $z_{kl}^{PP'}$, which we called *connectivity components*—to resemble standard PCA terminology. Both are defined below. In that earlier work (Pinotsis et al., 2017), we found that the principal axes contained temporal information, while the connectivity components contained spatial information. The connectivity components $z_{kl}^{PP'}$ were defined by the following equations

$$\begin{aligned} z_0^{PP'}(u,v) &= df_0 {\tau_P}^{-1} \iint K_{PP'}(u,v,u',v') du'dv', P, P' = \{E, I\} \\ z_{kl}^{PP'}(u,v) &= \tfrac{df_0}{k!} {\tau_P}^{-1} \iint K_{PP'}(u,v,u',v')(u-u')^k (v-v')^l du'dv' \qquad (3)\end{aligned}$$

while the principal axes were given by

$$G_{kl}^P = \frac{\partial \hat{X}_P(u,v,t)}{\partial u^{(k)} \partial v^{(l)}} \qquad (4)$$

Using Eqs. (3) and (4), Eq. (2) yields the following expressions for the perturbations $\hat{X}_P(u,v,t)$:

$$\begin{aligned}\hat{X}_E(u,v,t) &= z_0^{EE}G_0^E + z_0^{EI}G_0^I + z_{10}^{EE}G_{10}^E + z_{10}^{EI}G_{10}^I + z_{01}^{EE}G_{01}^E + z_{01}^{EI}G_{01}^I \\ &+ z_{11}^{EE}G_{11}^E + z_{11}^{EI}G_{11}^I + z_{20}^{EE}G_{20}^E + z_{20}^{EI}G_{20}^I + z_{02}^{EE}G_{02}^E + z_{02}^{EI}G_{02}^I \\ &+ z_{21}^{EE}G_{21}^E + z_{21}^{EI}G_{21}^I + z_{12}^{EE}G_{12}^E + z_{12}^{EI}G_{12}^I + O(u^3, v^3) \\ \hat{X}_I(u,v,t) &= z_0^{II}G_0^I + z_0^{IE}G_0^E + z_{10}^{II}G_{10}^I + z_{10}^{IE}G_{10}^E + z_{01}^{II}G_{01}^I + z_{01}^{IE}G_{01}^E \\ &+ z_{11}^{II}G_{11}^I + z_{11}^{IE}G_{11}^E + z_{20}^{II}G_{20}^I + z_{20}^{IE}G_{20}^I + z_{02}^{II}G_{02}^I + z_{02}^{IE}G_{02}^E \\ &+ z_{21}^{II}G_{21}^I + z_{21}^{IE}G_{21}^I + z_{12}^{II}G_{12}^I + z_{12}^{IE}G_{12}^E + O(u^3, v^3) \qquad (5)\end{aligned}$$

Note that the above equation is obtained using linear stability analysis and includes a Taylor expansion over spatial coordinates. If we had separate data (depolarization or spike rates) for the excitatory and inhibitory populations, we could use Eq. (5) and this data to find $\hat{X}_E(u,v)$ and $\hat{X}_I(u,v)$ separately. We could estimate the connectivity components $z_{kl}^{PP'}$ for the excitatory and inhibitory populations separately. We will pursue this in future work using data from excitatory and inhibitory neurons. Here, our data included aggregate activity (LFPs) from both populations.

LFP recordings contain aggregate activity of excitatory and inhibitory populations together. Mathematically, this is expressed as a two factor sum of membrane depolarization of all populations for each location on the cortical surface, $\hat{X}(u,v) = \hat{X}_E(u,v) + r\hat{X}_I(u,v)$, where r is the ratio of excitatory to inhibitory activity, which we take *r=0.25.* This value for r was chosen according to Dale's principle that neurons can be either excitatory or inhibitory and there are four times more excitatory than inhibitory neurons (Eccles, Fatt and Koketsu, 1954, Song, Yang and Wang, 2016). For mathematical convenience and without loss of generality we also consider a (differentiable) change of coordinates $(u,v) \to (\tilde{u}, \tilde{v})$ where $\tilde{u}$ parameterizes the location of the excitatory populations and $\tilde{v}$ parameterizes the location of the inhibitory populations. We also assume that the Jacobian of this transformation $J(\tilde{u}, \tilde{v}) \neq 0$. In the Results section, we validated this assumption numerically. The rigorous mathematical justification of this assumption will be considered elsewhere. Following this, the principal axes G_{lk}^P and components can be simplified:

$$\text{(i) } \left\{G_{lk}^P, z_{kl}^{PP'}, z_0^{PP'}\right\} = \begin{cases} \{G_{lk}^P(\tilde{u}), z_{kl}^{PP'}(\tilde{u}), z_0^{PP'}(\tilde{u})\}, P = E, \\ \{G_{lk}^P(\tilde{v}), z_{kl}^{PP'}(\tilde{v}), z_0^{PP'}(\tilde{v})\}, P = I, \end{cases} P' = \{E, I\}$$

$$\text{(ii) } G_{lk}^P = \begin{cases} 0, P = E, \text{any} l, k \neq 0 \\ 0, P = I, \text{any} k, l \neq 0 \end{cases}$$

Thus: (i) Principal axes G_{lk}^P and components$z_{kl}^{PP'}$ describing excitatory populations depend on $\tilde{u}$ only and terms describing inhibitory populations depend on $\tilde{v}$ (this was the assumption above); (ii) Axes G_{lk}^E involving excitatory activity involving non zero sub-indices k can be removed from Eq. (5), because these axes contain mixed derivatives. Similarly for inhibitory activity and its axes G_{lk}^I that contain mixed derivatives with non zero sub-indices l. Because $\tilde{u}$ and $\tilde{v}$ are distinct (the locations of excitatory and inhibitory populations are different), we can consider the union of the spatial domains for $\tilde{u}$ and $\tilde{v}$ as a single, new spatial domain and join the spatial variables $\tilde{u}$ for the location of excitatory and $\tilde{v}$ for the location of the inhibitory populations into a single variable. Then, adding Eqs. (5a) and (5b), we obtain

$$\hat{X} \approx \tilde{A}_0^E G_0^E + \tilde{A}_0^I G_0^I + \tilde{A}_{10}^E G_{10}^E + \tilde{A}_{01}^I G_{01}^I + \tilde{A}_{20}^E G_{20}^E + \tilde{A}_{02}^I G_{02}^I + \varepsilon \qquad (6)$$

where the aggregate connectivity components $\tilde{A}^P$ are two factor sums of $z^{PP'}$ defined by Eq. (3):

$$\begin{aligned}\tilde{A}_0^P &= q^{P'}(z_0^{PP} + q^P z_0^{P'P}), q^P = \begin{cases} r, P = E \\ 1/r, P = I \end{cases}, P \neq P' \\ \tilde{A}_{kl}^P &= q^{P'}(z_{kl}^{PP} + q^P z_{kl}^{P'P})\end{aligned} \qquad (7)$$

Letting $H = [G_{10}^E, G_{01}^I, G_{20}^E, G_{02}^I, ..., G_{i\cdots 0}^E, G_{0\cdots i}^I]^T$ Eq. (6). is a *deep neural field*, and can be rewritten in the general form of a Gaussian Linear Model (GLM; cf Eq. (1). in (Pinotsis et al., 2017)),

$$\begin{aligned} Y &= \textstyle\sum_j H_j w_j + m + R \\ w &= \left[A_0, A_1, A_2, ..., A_{i-1}, A_i\right]^T \\ m &= N^{-1} \textstyle\sum_N X^l \end{aligned} \qquad (8)$$

where for simplicity of notation we have relabelled, $A_k = [\tilde{A}_0^P, \tilde{A}_{01}^P, \tilde{A}_{10}^P, ..., \tilde{A}_{i0}^P, \tilde{A}_{0i}^P]$ and have dropped the superscript P, because we do not distinguish between neural populations in what follows. This simply relabels components with two sub-indices as components with a single sub-index. Note that A_k are 1D, while $z_{kl}^{PP'}$ are 2D. Since there is only one spatial variable in A_k, only one sub-index was needed. We have also assumed that cortical activity $\hat{X}(\tilde{u}, \tilde{v}, t) \in \tilde{X}$ was sampled from a random process$\hat{X}$ and $Y = \hat{X} - m$.

The above 1D reduction was obtained under certain mathematical assumptions. To validate them, we compared our effective connectivity estimates against two established approaches (see Methods section below and Results section). We found that our results correlated significantly with results obtained with these methods. The rigorous mathematical justification of these assumptions will be pursued elsewhere.

In brief, starting from a neural network model for coupled excitatory and inhibitory populations (Eq. 1), we have shown how it can be reformulated as a deep neural field model (Eq. 6) – and then a GLM (Eq. 8). This is useful because it allows us to obtain the effective connectivity that characterises information flow within the neural ensemble. This is described in the next section.

2.3. Connectivity components and kernels

The connectivity components A_k are the latent states of the autoencoder trained by optimizing the cost function, known as the Free Energy, *F*,

$$\begin{aligned} F &= \left(-\frac{1}{2}\right)\left[(Y - Hw)^T r_s^2 (Y - Hw) + \ln\left|s_s^2\right| + \ln\left|{s^2}_s \Delta^{-1}\right| + Z^T Z + const\right] \\ \Delta &= {s^2}_s I + H^T H \\ Z &= \Delta^{-1} H^T Y \end{aligned} \qquad (9)$$

using a Restricted Maximum-Likelihood (ReML) algorithm (Harville, 1977). This assumed a directed graphical model $p(Y|w)$ used in autoencoders that yields an approximation q to the posterior $p \sim \mathrm{N}(w|Y)$, see (Pinotsis et al., 2017) for more details. Note that the cost function defined by Eq. (9), is the same cost function like the one used in Predictive Coding.

To summarize, Equations (3) define the connectivity components $z^{PP'}$ of the neural network (1). Similarly, Equations (7) define the 1D connectivity components A_k of the deep neural field (Eq. (6)) as two factor sums of $z^{PP'}$. Training the GLM to optimize the cost function (9) we obtain single trial estimates of effective connectivity components A_k. Their averages across trials are shown in Fig. 2 of (Pinotsis et al., 2017). If we had separate recordings of excitatory and inhibitory neurons we could get the effective connectivity components $z^{PP'}$ of the neural network (1) in a similar way. This will be pursued elsewhere. Here, we used single trial A_k estimates to identify neural ensembles that maintained location during each trial. We also compared them to similar measures obtained using other approaches for ensemble identification (see Methods below and Results).

We now turn to connection weights of the neural network (1). We call these connectivity kernels $K_{PP'}$. In Equations (3), the connectivity components are integrals of the connectivity kernel $K(u_a, v_a, u_b, v_b, t, t')$. Here we have dropped the sub-indices P, P' because the kernel is not spatially discrete; instead, it depends on continuous variables (u, v).

In (Pinotsis et al., 2017), after obtaining A_k, we assumed that cortical connectivity has a Gaussian profile and computed $K(u_a, u_b) = (C\sqrt{2\pi})^{-1}\exp\{-(u_a - u_b - \bar{u})^2/2C^2\}$. This is shown schematically in Fig. 1A as Gaussian (bell shaped) curves connecting any two electrodes sampling from the patch. We also obtained trial average estimates of $K(u_a, u_b)$, where $\bar{u}$ and C are the mean and standard deviation of axonal dispersion. Here, we first considered the same profile and focused on the corresponding single trial estimates of $K(u_a, u_b)$. We also considered a more general expression for the connectivity profile involving a weighted Gaussian (see *Mapping the latent space to a cortical patch* section below). Examples of connectivity kernels are shown in Fig. 2B.

2.4. Comparison of our approach to established approaches in the literature

To validate our approach, we compared our estimates of connectivity components and kernels to methods that are established in the literature. First, we considered a correlation-based method, see (Humphries, 2011). This yields neuronal ensembles, where neurons in the same ensemble have dense connections with each other and weak connections to other neurons. This is achieved by maximizing a graph theoretic measure known as modularity and is similar to finding communities in social networks (Newman and Girvan, 2004). It computes similarity measures including cosine similarity and the correlation coefficient that we used here. The method was initially developed to analyse spike train data, but we here adapted it to deal with LFPs. It provides a spectral decomposition of the modularity matrix using a stochastic algorithm (Newman and Girvan, 2004). It employs a consensus algorithm to ensure that the same clustering is obtained for different initialisations (Lancichinetti and Fortunato, 2012). This method has been applied to neural activity in visual cat areas (Humphries, 2011) and the Aplysia pedal ganglion (Bruno et al., 2015).

Second, we used a higher dimensional SVD method known as canonical decomposition (CD (Carroll and Chang, 1970; Kolda and Bader, 2009)). This provides a generalization of the usual SVD which factorizes a tensor in terms of R arrays. It allows one to obtain an approximation of the data represented by a third order tensor $Y \in (\mathbb{R})^{N_T \times N_S \times T}$ given by

$$\tilde{Y}_{ijk} \approx f_{i1}b_{j1}c_{k1} + f_{i2}b_{j2}c_{k2} + \dots + f_{iR}b_{jR}c_{kR} \tag{10}$$

where $f_{im} \in (\mathbb{R})^{N_T \times m}, b_{jm} \in (\mathbb{R})^{N_S \times m}$ and $c_{km} \in (\mathbb{R})^{T \times m}$ are three matrices known as "modes" in the mathematical literature (Kolda and Bader, 2009). Their first dimensions are either number of trials (N_T), or electrodes (N_S) or time (T). R is known as the rank of Y, with *m=1,…R*. Eq. (10). includes a sum of combinations of elements $f_{i1}, b_{j1}, c_{k1}, f_{i2}, b_{j2}, c_{k2}, \dots$. Taking together (i.e. for all *m=1,…R*) all elements with the *same* first dimension, e.g. the dimension denoted by index "*i*", that is,$f_{i1}, .f_{i2}, .., f_{iR}$ we obtain a matrix $F = [f_{im}]$ and similarly for $B = [b_{jm}]$ and $C = [c_{km}]$. *F, B* and *C* are known as *modes* Each mode is a matrix where the first dimension (denoted by *i, j* or *k*) is equal to one of the above three dimensions of the LFP array, that is, a number of trials (*i=1,..,*N_T), electrodes (*j=1,…,*N_S) or time points (*k=1,…,T*). Thus, each term in the sum $\tilde{Y}_{ijk}$ is a product of elements from the three modes $f_{im}b_{jm}c_{km}$. This product is called a *factor*. The second dimension (denoted by *m*) is the same for all three modes that belong to the same factor and is different for each term in the sum (i.e. each factor). It ranges between 1 and some arbitrary number *R*, $m = 1, \dots, R$. Thus, *R* is equal to the number of factors in the CD approximation $\tilde{Y}_{ijk}$. In the Results section, we will see that *R* can be estimated based on some measures from statistics.

The approximation $\tilde{Y}$ is obtained using an alternating least squares algorithm (ALS) that minimizes the reconstruction error$\min_{\tilde{Y}} \|Y - \tilde{Y}\|_F$, where $\|Y\|_F$ is the Frobenius norm of Y. The ALS approach fixes $B = [b_{jm}]$ and $C = [c_{km}]$ to find $F = [f_{im}]$. The conditional least square estimate of A is then

$$F = Y(B \otimes C)^T\left[(B \otimes C)(B \otimes C)^T\right]^{-1} \tag{11}$$

where ⊗ is known as the Kronecker product. ALS continues by then fixing F and C to find B and finally F and C to find *B*. The CD approximation is unique up to permutation and scaling of the modes. Thus, ALS is used iteratively. For more details see (ten Berge, 1993). CD has recently been applied to analyse spiking data and identify neural ensembles in (Williams et al., 2018). Here, we adapted this work to identify neural ensembles using LFPs. The CD approach by (Williams et al., 2018) does *not* provide single trial estimates of connectivity components and kernels, like those we considered here. However, we compared our results to CD estimates after averaging across all trials that corresponded to the same stimulus.

In the expansion (10) above, the number of components is arbitrary. To find the rank *R,* we used two criteria: consistency and congruence. Consistency was introduced as an alternative way to obtain the rank in CD approximations (Bro and Kiers, 2003). It uses certain elements of CD theory, known as CD factors, to compute an alternative approximation of the data matrix, known as Tucker3 approximation (Tucker, 1966). The Tucker3 approximation also contains (mixtures of) CD factors. For a given *R,* consistency quantifies the difference between data fits using the CD and Tucker3 approximations. *R* should be such that this difference is minimal. According to (Bro and Kiers, 2003), this corresponds to consistency values between 50-100%. In brief, consistency quantifies the degree that the LFP data contain a trilinear variation, see (Bro and Kiers, 2003) for more details. It is optimal for that particular value of *R,* that renders the core of the corresponding Tucker3 approximation (the Tucker3 approximation with the same CD factors) superdiagonal.

Congruence, on the other hand, is simply based on uncorrected correlation coefficients (CC) between any two sets of factor matrices {F_1, B_1, C_1} and {F_2, B_2, C_2} . These are averaged over a different implementations of the ALS algorithm starting from different initial conditions and then the maximum value is subtracted from 1, i.e. congruence (CG) is given by $CG = 1 - \max(1/N \sum_t \rho_t))$ where ρ_t is the CC computed in the *t*-th initialisation and we have assumed *N* initialisations. Congruence was initially used to remedy instabilities and slow convergence that are knowns to affect ALS, due to its iterative nature. A low value of congruence implies that the CD approximation was *not* stuck in local minima and CD factors are stable (Kiers, 1998). In Results, we chose a rank *R* with high consistency and low congruence. This results in a stable CD approximation that includes a trilinear variation in the data.

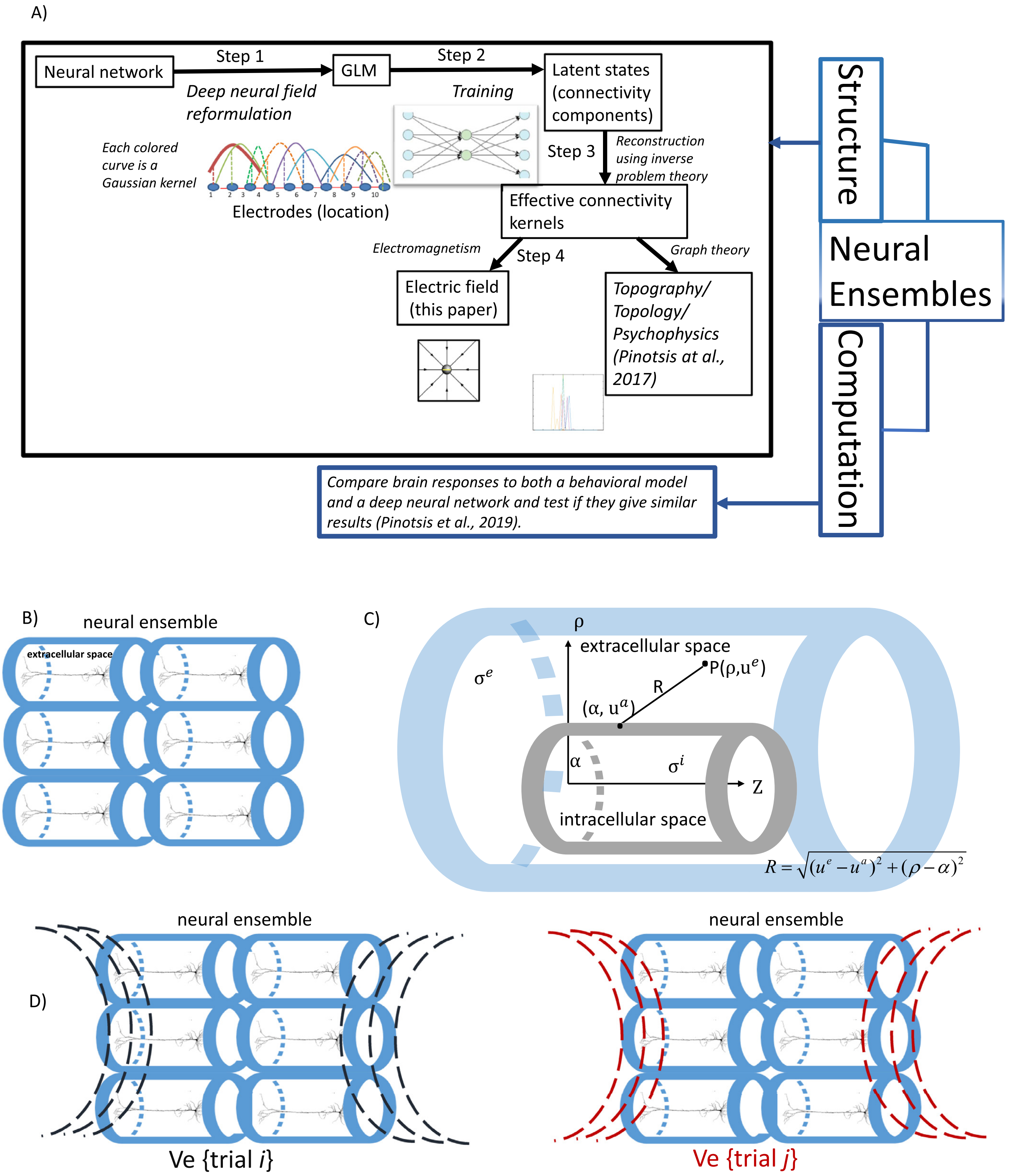
A)
Neural network
Step 1
Deep neural field reformulation
GLM
Step 2
Training
Latent states (connectivity components)
Step 3
Reconstruction using inverse problem theory
Each colored curve is a Gaussian kernel
1 2 3 4 5 6 7 8 9 10
Electrodes (location)
Effective connectivity kernels
Electromagnetism
Step 4
Graph theory
Electric field (this paper)
Topography/ Topology/ Psychophysics (Pinotsis at al., 2017)
Compare brain responses to both a behavioral model and a deep neural network and test if they give similar results (Pinotsis et al., 2019).
Structure
Neural Ensembles
Computation
B)
neural ensemble
extracellular space
C)
ρ
extracellular space
P(ρ,u^e)
σ^e
R
(α, u^a)
α
σ^i
Z
intracellular space
$R=\sqrt{(u^e-u^a)^2+(\rho-\alpha)^2}$
D)
neural ensemble
Ve {trial i}
neural ensemble
Ve {trial j}

Fig. 1. A. Outline of our approach. We first reformulated a neural network (described by Wilson Cowan equations) as a neural field model and then a Gaussian Linear Model (GLM; step 1. We trained this model as an autoencoder and obtained the latent states (connectivity components; step 2). Then, using inverse problem theory, we obtained the corresponding connectivity kernels (step 3). In (Pinotsis et al., 2017), we used the kernels and graph theory to characterize the topography and topology of neural ensembles. Here, we use the kernels and electromagnetism (dipole theory) to study the stability of the electric field generated by an ensemble (step 4). This paper and (Pinotsis et al., 2017) focus on the structure and biophysics of neural ensembles. In related work (Pinotsis et al., 2019), we also studied the computations performed by ensembles using deep neural networks and behavioural models. B. Extracellular space around each neuron within the ensemble (blue cylindrical fibers). C. Bidomain model for the electric field generated by a cylindrical fiber in a conductor. The extracellular and intracellular space are depicted by blue and grey cylindrical fibers (see Methods for the meaning of various symbols) D. Extracellular electric potential V^e corresponding to two random trials: *i* (black dashed lines, left panel) and *j* (red dashed lines, right panel).

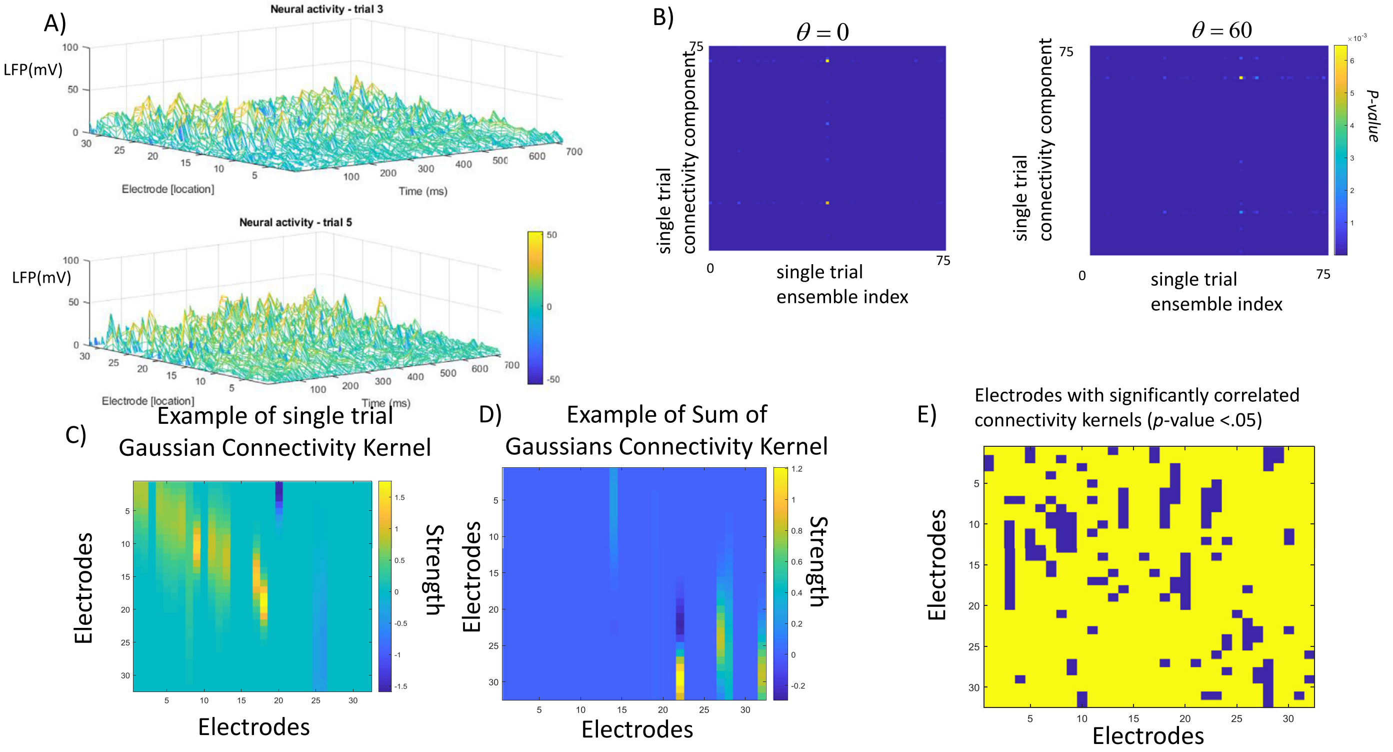

Fig. 2. A. Examples of neural activity for two different individual trials corresponding to the same task condition. Local field potentials (LFPs, in *mV*) are shown on the vertical axis. The electrodes (location on the cortex) and time (in *ms*) are shown on the two horizontal axes. B. Significance (*p*-value) of Pearson correlations between the single trial connectivity components and ensemble indices obtained by the approach of (Humphries, 2011). Trials where a horizontal location was maintained (θ=0 degrees) are shown in the left panel. Similarly, trials for cued angle at θ=60 degrees are shown in the right panel. Estimates for all trials correlated perfectly ($p<10^{-2}$). C. Example of effective connectivity kernel with a Gaussian profile. This describes the weights which scale neural activity propagating between any pair of populations located near one of the electrodes. This kernel characterizes information flow at the single trial level. D. Example of an alternative expression of the effective connectivity kernel obtained as a weighted Gaussian using a series expansion. E. Correlations between the connectivity kernels in panels C. and D. R=87% of connectivity weights were significantly correlated at the p<.05 level. These are shown in yellow. Blue denotes weights that were not significantly correlated.

2.5. The electric potential and electric field generated by a neural ensemble

To model the *electric potential* (EP) generated by synaptic activity (EPSPs and IPSPs) in a neural ensemble we use the bidomain model of the neural tissue (Schwartz et al., 2016; Goldwyn et al., 2017; Mc Laughlin et al., 2010). This assumes that the neural ensemble can be represented by a cylindrical fiber of radius a (grey cylinder in Fig. 1C). Pyramidal neurons are assumed to be aligned and the EP (and the electric field) varies primarily along the dendritic axis. Also, they receive synchronous synaptic input. Under these assumptions, the extracellular space of each neuron can be described by a cylindrical fiber (small blue cylinders in Fig. 1B). Then, the electromagnetic principle of superposition allows us to replace the individual cylindrical fibers of Fig. 1B (for each neuron) with a larger cylindrical fiber corresponding to the extracellular space surrounding the neural ensemble (blue cylinder in Fig. 1C). The extracellular current near each neuron can be added to obtain an aggregate current that flows in the extracellular space around the ensemble. Similarly, current flowing along the dendrites of each neuron (in the intracellular space) can be added to obtain an aggregate current that flows within the ensemble (grey cylinder in Fig. 1C). The EP and the electric field have rotational symmetry. The potential is a function of two coordinates (ρ, Z), where $Z = u_a$ is the coordinate along the fiber axis and ρ is coordinate vertical to it, see (Plonsey, 1974) and Fig. 1C. Deviations from symmetry (spatial inhomogeneity) and asynchronous input will change the current flowing and the electric field outside the ensemble (Goldwyn et al., 2017) (e.g. axes of individual cylindrical fibers of Fig. 1B might cross). Because of the principle of superposition in electromagnetics, this will reduce the overall extracellular electric field. However, it will not affect qualitive results, like the stability of the electric field we will discuss later. This is because the laws of electromagnetism do not change and can still be applied by splitting the extracellular and intracellular spaces into smaller parts (cylindrical fibers) where symmetry and synchrony still apply. Below, we use the bidomain model, to derive the extracellular potential V^e and the extracellular electric field generated by the neural ensemble, E^e. The extracellular potential V^e (corresponding to two different trials, see next section) is shown in Fig. 1D using black and red dashed lines. Details of this derivation can be found in the references above. Here, we included a summary for the convenience of the reader. The bidomain model describes the potential in the two sides of the neuron membrane, that is, the intracellular V^i and extracellular V^e potentials. Their difference $V^m = V_0^e - V_0^i$ is the transmembrane potential and results in a spatial discontinuity also for the electric field $E^a = -\nabla V_o^a, a = \{e, i\}$. V_0^e and V_0^i are the values of the extracellular and intracellular EPs on the two sides of the membrane. Note that ∇ denotes the gradient operator. According to the theory of electromagnetism, this discontinuity gives rise to dipole sources with moments (Jackson, 1999)

$$p_a = \nabla^2 V^m / r \tag{12}$$

Here r is the brain resistivity with r = 2.2Ohm (Rush and Driscoll, 1969) and we have assumed that the number of neurons is large and that each cell is very small compared to the distance at which the LFP electrode is placed. Also, the current density $I^a(u_a, v_a)$ that results from EPSPs and IPSPs is given by

$$I^a(u_a, v_a) = p_a / \Omega \tag{13}$$

where Ω is the total volume of the ensemble. Neglecting ephaptic interactions $V^m \approx V^i$, and the extracellular electric potential generated by the current density $I^a(u_a, v_a)$ is given by

$$V^e(u^e, v^e, w^e) = (4\pi\sigma^e)^{-1} \int I(u^a, v^a)\nabla(1/R)d\Omega \tag{14}$$

where $\sigma^e$is the conductivity of the extracellular space, and R is the distance between the current source at the point (α, u^a) of the neural ensemble and the point (ρ, u^e) in the extracellular space where we measure V^e, i.e. the location of the LFP electrode,

$R = \sqrt{(u^e - u^a)^2 + (\rho - \alpha)^2}$, see Fig. 1C. Then, according to the bidomain model, Eq. (14) can be written as (Henriquez, 1993; Roth, 1997)

$$V^e(u^e, v^e, w^e) = -(4\pi\sigma^e/\sigma^i)FT^{-1}\left[\hat{V}^m(k)W(k)\right] \tag{15}$$

where $\hat{V}^m(k)$ is the Fourier Transform of the transmembrane potential V^m and FT^{-1} is its inverse Fourier Transform, that is,

$$\begin{aligned} \hat{V}^m(k) &= \int_{-\infty}^{\infty} V^m(\rho)e^{ik\rho}d\rho \\ FT^{-1}\left[\hat{V}^m(k)\right] &= V^m(\rho) = \int_{-\infty}^{\infty} \hat{V}^m(k)e^{-ik\rho}dk \end{aligned} \tag{16}$$

The function $W(k)$ is given in terms of the modified Bessel functions of the first $I_0(\rho), I_1(\rho)$ and second $K_0(\rho), K_1(\rho)$ kind (Abramowitz et al., 1988),

$$W(k) = \frac{I_1(|k|d)K_0(|k|\rho)}{I_0(|k|d)K_1(|k|d) + \sigma^i/\sigma^e I_1(|k|d)K_0(|k|d)} \tag{17}$$

Then, the *extracellular electric field* (EF) generated by the neural ensemble, E^e, is just the gradient of V^e, $E^e = -\nabla V^e$.

2.6. Gauge functions and ensemble electric fields

Multiple extracellular EPs V^e can give rise to the same EF $E^e = -\nabla V^e$ in extracellular space. This is a well-known result in the theory of electromagnetism called *Gauge invariance.* It follows from the conservation of electrical charges Jackson, 1999). The same electric field can be expressed in terms of different potential functions, V^e (Darrigol, 2003). These describe different arrangements of electric sources and capture symmetries of the Maxwell equations that describe the evolution of the extracellular EF, E^e (Maxwell, 2021). In the case of LFP measured with multielectrode arrays, each trial gives rise to an LFP recording. This, in turn, results from a different EP generated by current flow within a neural ensemble in each trial. In Results, we test the *hypothesis* that the *EF is the same for all trials corresponding to the same remembered stimulus*, $E^e\{\text{trial}i\} = E^e\{\text{trial}j\}$. To test this hypothesis, we first needed an estimate of the EP at an arbitrary trial *j*, $V^e\{\text{trial}j\}$. This is shown by red dashed lines in the right panel of Fig. 1D (see also the discussion in the previous section). We obtained this using Eqs. (15) and ((17) above, in two ways. First, using simulations of our deep neural field model. Second, using recorded LFPs as proxies for the transmembrane potential at arbitrary trial *j*, $V^m\{\text{trial}j\}$. Then by taking the gradient of $V^e\{\text{trial}j\}$, we found the extracellular EF for trial *j*, $E^e\{\text{trial}j\}$. Having obtained EF estimates, we tested the hypothesis that the EF is stable in three ways: First, we looked whether EFs where correlated across trials. Second, we asked if EF estimates were consistently different for neural ensembles that maintain different cued angles. We tested if we could distinguish between memorized cues based on EFs. We used EFs as classification features in two commonly used classification algorithms, Naïve Bayes and diagonal LDA (Pinotsis et al., 2017). Third, we used Gauge functions that connect the recorded LFPs. If the EF was stable, the EPs are related by *Gauge functions* $\chi = \chi(\rho, \zeta, t)$ (Jackson, 1999)

$$V^e\{\text{trial}i\} = V^e\{\text{trial}j\} + \partial\chi/\partial t \tag{18}$$

A Gauge function describes the difference between two electric potential functions that result in the same electric field. In Fig. 1D, $V^e\{\text{trial}i\}$ is shown using black dashed lines (left panel). Subtracting $V^e\{\text{trial}j\}$ from $V^e\{\text{trial}i\}$, we obtain the temporal derivative (rate of change) of the Gauge function. In short, a third way to test the stability of EFs is to test if the Gauge functions can be used to distinguish between different cued angles (see Results). According to Eq. (18), the time derivative of the Gauge function $\partial\chi/\partial t$ is equal to the difference of EPs corresponding to any two trials. Eq. (18) should hold for any arbitrary pair of trials. Thus, we asked whether we could decode cued angles using Gauge function derivatives $\partial\chi/\partial t$ as classification features. These, in turn, were obtained after subtracting LFP recordings. An independent experimental validation could also be carried out using intracellular recordings: If Eq. (18) holds, then a similar Equation for the intracellular potential V^i also holds with the same Gauge function $\chi = \chi(\rho, \zeta, t)$. Thus, the Gauge function$\chi(\rho, \zeta, t)$, can be found experimentally by measuring V^i during any two trials, $V^i\{\text{trial}i\}$ and $V^i\{\text{trial}j\}$.

2.7. Mapping the latent space to a cortical patch

The extra step that allowed us to obtain the electric field above was the mapping of the latent space to a cortical patch (Pinotsis et al., 2017). Starting from the connectivity components, we obtained the weights that scaled incoming input to each population from all other populations in the ensemble, called the connectivity kernel. This describes information exchange and electrical activity on the patch. Having this, we then reconstructed the EF. Consider Eq. (1). The connectivity kernels $K_{PP'}(u_X, v_X, u_c, v_c)$, $X = \{a, b\}$ include the weights that scale input from a population at location (u_c, v_c) to an excitatory population at (u_a, v_a) or an inhibitory population at (u_b, v_b). Above, we considered the continuum limit of Equations (1), that is, Equations (2) and similarly the continuum limit of the connectivity kernels$K_{PP'}(u, v, u', v')$. These have the same meaning as$K_{PP'}(u_X, v_X, u_c, v_c)$. Only a difference in notation: the subindices denoting location have been replaced by continuous spatial variables that lie on a patch $u, v \in M_A$. Then, given the connectivity components A_0, A_{kl}, we can find $K_{PP'}$. In mathematical terms, the kernels are probability distribution functions and can be estimated using a variety of methods from inverse problems theory, including splines (Gehringer and Redner, 1992), series expansions (Amindavar and Ritcey, 1994) and other methods (Heinz, 2013; Mersmann, 1995).

We here considered a Gaussian connectivity profile used in (Pinotsis et al., 2017) and an alternative expression for the connectivity kernel that includes sums of Gaussian profiles weighted by polynomial factors known as Hermite polynomials, H_n (Abramowitz et al, 1988). These sums are known as Gram-Charlier series. In brief, the connectivity kernel of the neural ensemble can be approximated by

$$\begin{aligned} K(u, u') &= \sum_0^k d_n H_n(u) g_C(u, u') \\ g_C(u, u') &= \exp\left\{-(u - u' - \bar{u})^2/2C^2\right\} \\ \bar{u} &= {}^{A_1}/_{A_0} \\ C &= \frac{\sqrt{A_0A_2 - {A_1}^2}}{A_0} \end{aligned}$$

$$\begin{aligned} K(u, u') &= d_0H_0(u)g_C(u, u') + d_1H_1(u)g_C(u, u') + \ldots. + d_kH_k(u)g_C(u, u') \\ g_C(u, u') &= \exp\left\{-(u - u' - \bar{u})^2/2C^2\right\} \\ \bar{u} &= {}^{A_1}/_{A_0} \\ C &= \frac{\sqrt{A_0A_2 - {A_1}^2}}{A_0} \end{aligned} \tag{19}$$

where the Hermite polynomials, H_n, *n=0,…k,* are *known* and the coefficients d_n can be found by substituting (19) and the definition of H_n into

$$d_n = C^{2n}/n! \int K_{PP'}(u, u')H_n(u)du \tag{20}$$

Interestingly, Eq. (20) using the binomial theorem and the definition of connectivity components gives

$$A_k(u,u') = \gamma \int \textstyle\sum_1^k d_n H_n(u) g_C(u,u') \sum_0^k (-1)^n \binom{k}{n} u^k u'^{k-n} du' \\ \gamma = \frac{df_0}{k!} {\tau_P}^{-1} \tag{21}$$

The above expression seems complicated. However, one can use the properties of the Hermite polynomials to find the coefficients $d_n, n = 0, 1, ..., k$.

$$\begin{aligned} &d_0 = A_0 \\ &d_1 = A_1 \\ &d_2 = 1/2\left[A_2 + 2(u' - \bar{u})A_1 + (u'^2 + \bar{u}^2 - C^2 - 2C\bar{u})A_0\right] \\ &\dots \end{aligned} \tag{22}$$

Substituting the above expressions and the expressions for Hermite polynomials into Eq. (19), we obtain an alternative expression for the connectivity kernel $K(u, u')$ that involves a Gaussian function weighted by terms involving connectivity components (keeping the first three terms in the series expansion given by Eq. (19)):

$$K(u,u') \approx g_C(u,u') \cdot \\ \cdot \begin{pmatrix} A_0 + A_1(u - \bar{u})/C^2 \\ +1/2\left[A_2 + 2(u' - \bar{u})A_1 + (u'^2 + \bar{u}^2 - C^2 - 2C\bar{u})A_0\right](u - \bar{u})^2/(C^4 - 1/C^2) \end{pmatrix} \tag{23}$$

3. Results

3.1. Deep neural fields describe neural ensemble structure in a holistic fashion

This paper follows upon our recent work that focused on groups of neurons that represent memories known as neural ensembles (Fig. 1A). In (Pinotsis et al., 2019) we studied computations performed by neural ensembles during a flexible sensorimotor decision making task (Siegel et al., 2015). We showed that neural ensembles in the same brain area performed different computations based on the rule applied during each trial, although the stimulus processed was the same. This result was obtained by comparing brain responses to both a behavioral model and a deep neural network and testing if they give similar results. In a parallel line of work (Pinotsis et al., 2017), we also studied the structure of neural ensembles and obtained their effective connectivity. Our analyses below build upon that earlier work and used the same dataset. This includes a spatial working memory task, where the angle of a cue had to be remembered (delayed saccade task; Supplementary Figure 1A). We analysed LFP data recorded during the delay period.

We analysed neural activity (LFPs) recorded from a multielectrode array of $N_s = 32$ electrodes implanted in the FEF of two macaque monkeys. LFPs are thought to describe neural activity from a population in the proximity of each electrode (Buzsáki et al., 2012, Lindén et al., 2011). Analysing LFPs allowed us to identify neural ensembles and test if they overlap in different trials. Electrodes were numbered in a monotonic fashion; neighbouring electrodes had adjacent numbers (Supplementary Fig. 1B). Our approach and the main results of (Pinotsis et al., 2017) are summarized below and in Fig. 1A.

Our approach has the following steps (each step is depicted by an arrow in Fig. 1A, see below): (1) Start with a neural network model. Reformulate this as a biophysical Gaussian Linear Model (GLM), that we called deep neural field. (2) Use LFP data to obtain the latent states (connectivity components) of the deep neural field model. The term “deep” was used in our earlier work (Pinotsis et al., 2017) to distinguish this model (with learned connectivity parameters) from common neural field models where connectivity weights are chosen ad hoc, e.g. (Bojak and Liley, 2010; Atay and Hutt, 2006; Deco et al., 2008; Coombes, 2010). The learned parameters are obtained after training the neural field as an autoencoder (see also (Pinotsis et al., 2017)). Thus, the term “deep” refers to the hidden layer of the corresponding training network. (3) Use the connectivity components and inverse problem theory to obtain the effective connectivity (connectivity kernels). We will come back to components and kernels (and their differences) below. (4) Use electromagnetism and the connectivity kernels to obtain the electric field generated by the neural ensemble. This will be discussed later.

In (Pinotsis et al., 2017), we used average connectivity estimates and graph theory and showed that path length portioned the space of cued angles. The smallest values occurred for cues on the horizontal meridian, i.e. information propagates faster. This provided an explanation of the oblique effect in psychophysics (Appelle, 1972) [1]. Here, we use single trial connectivity estimates and electromagnetism to reconstruct the electric field produced by a neural ensemble. We assumed that this is predicted by the bidomain model of the neural tissue (Schwartz et al., 2016; Goldwyn et al., 2017; Mc Laughlin et al., 2010) (Fig. 1C), see also *Methods*. In this model, the extracellular spaces of individual neurons are described by cylindrical fibers (blue cylinders in Fig. 1B). The bidomain model predicts that extracellular currents measured in different trials (e.g. trials *i* and *j*), generate different electric potentials $V^e\{\text{trial}i\}$ (black dashed curves, left panel in Fig. 1D) and $V^e\{\text{trial}j\}$ (red dashed curves, right panel in Fig. 1D). This will be discussed in detail later.

Examples of neural activity for two different individual trials corresponding to the same task condition are shown in Fig. 2A. LFP amplitudes (in *mV*) are shown on the vertical axis. The horizontal axis are electrode number (location) and time (in *ms*). We assumed that FEF comprised a large number of neural populations (indexed by *j*=1,…,*N*=32*),* that was equal to the number of electrodes we sampled from, see also (Pinotsis et al., 2017). Each of these populations can be thought of as centred around a point (u_a, v_a) on the 2D cortical surface. They also interact with other populations located at point (u_b, v_b), via an effective connectivity kernel $K(u_a, v_a, u_b, v_b, t, t')$.

We previously identified neural ensembles based on their effective connectivity kernel averaged across trials (Pinotsis et al., 2017). This connectivity was expressed in terms of two measures: (1) the latent variables of an autoencoder that we called *connectivity components* and (2) the *connectivity kernel* $K(u_a, v_a, u_b, v_b, t, t')$ of a biophysical rate model (neural field). This kernel was obtained from the connectivity components after assuming a Gaussian connectivity profile over space. Here, we followed a similar approach and focused on effective connectivity of a neural ensemble and its components at the single trial level (i.e., without averaging). We also considered a more general weighted Gaussian as a connectivity profile over space. Our starting point was different to [29]: we modelled each neural ensemble as a 2D neural network model of interacting excitatory and inhibitory populations (Wilson-Cowan Equations; see Methods). By changing the variable that parameterised the cortical surface from discrete to continuous, the neural network was reformulated as a mean field model, known as a neural field (Wilson and Cowan, 1973; Amari, 1977; Coombes, 2007). In (Pinotsis et al., 2017), our starting point was a usual neural field.

Since we are measuring aggregate activity (LFPs), we could not distinguish between locations of excitatory vs inhibitory populations. At the same time, these locations do not overlap. Intuitively, this means that we can join the 2 spatial variables in the neural network, describing locations of excitatory and inhibitory populations, into one. In conclusion, the original 2D neural network model was first transformed to a 2D neural field and then to an 1D deep neural field considered in (Pinotsis et al., 2017). The details of this reduction are included in Methods. Its mathematical implications will be considered elsewhere. Here, we assessed whether this reduction allowed us to correctly identify neu-

[1] The oblique effect is the relative deficiency in perceptual performance for oblique angles or contours compared to horizontal or vertical stimuli (Appelle, 1972). In (Pinotsis et al., 2017), we reconstructed the connectivity of ensembles maintaining oblique and horizontal angles and found that it was sparsest in the latter case.

ral ensembles, by comparing our results to those obtained using other methods.

3.2. Effective connectivity components of deep neural fields reveal clusters obtained using pairwise correlations

We compared the effective connectivity components and kernels obtained with the deep neural field model to estimates of ensemble connectivity obtained with other methods. Below, we show that our effective connectivity estimates correlated significantly with connectivity estimates obtained using pairwise correlations (Humphries, 2011) and a high dimensional SVD approach (Carroll and Chang, 1970).

We first discuss connectivity components (denoted by A_k in Methods). These are the latent variables of the low dimensional space obtained after training our deep neural field as an autoencoder. We will see below that they describe aggregate synaptic input to neural populations located at a certain point on the cortical patch. They cluster neurons into task related groups (Williams et al., 2018). Specifically, we obtained single trial component estimates in the following way. We trained a deep neural field model using a cost function considered in predictive coding and autoencoder networks. Component averages across trials were shown in Fig. 2 of (Pinotsis et al., 2017). In that paper, we also showed that connectivity components were matrix-valued functions with dimensionality equal to$N_T \times N_S$, where $N_T = 600$ is the number of trials. For each trial, we obtained a vector of dimension N_S whose entries were called *component strengths*. These were similar to loadings or principal components in PCA. Because we here used aggregate neural activity from both kinds of populations (LFPs), we obtained effective connectivity components for both populations together. This is similar to other dimensionality reduction approaches, e.g. (Pang et al., 2016; Williams et al., 2018).

Here, we validated the effective connectivity components obtained in (Pinotsis et al., 2017) (summarized also in (Pinotsis and Miller, 2017)) using two independent methods. First, using a correlation-based method, see (Humphries, 2011). This was originally used to identify similarities between spike trains. It was based on pairwise correlations. Similarities were then used to define neural ensembles –assuming that neurons with similar spiking patterns represented the same stimulus or sequence. Thus, one obtains neural ensembles. Each neuron is included in an ensemble (called a “cluster” in the original paper), indicated by an ensemble index. In other words, the approach by (Humphries, 2011) did not yield effective connectivity per se, but one can map the ensemble index to effective connectivity components and kernels that we obtained.

Below, we compare our effective connectivity estimates to the clusters obtained using the method presented in (Humphries, 2011). This yields an alternative way to obtain the same neural ensembles described by our deep neural field model in an unsupervised way, using a *k*-means algorithm. We adapted the original algorithm from (Humphries, 2011) to work with LFP data, instead of spike trains. For each trial, the method assigned each electrode to an ensemble using an ensemble index. Assuming that electrodes sample from populations in their proximity (Buzsáki et al., 2012; Lindén et al., 2011), this process also assigns populations to ensembles. We then computed the correlation between the ensemble indices and the first connectivity components for different cued angles. We asked whether the ensemble index correlated with the component strength for each electrode and trial. In (Pinotsis et al., 2017), we showed that the component strengths are aggregate sums of all the weights of all connections that target the electrode at hand. They describe changes of signal as it propagates between electrodes. Thus, different values of component strengths correspond to different levels of activity (drive) that each electrode receives.

Pearson correlations were obtained for ensembles obtained from trials with different cued angles. Recall that the monkey performed a spatial delayed saccade task (Supplementary Fig. 1A). The *p*-values across all 32 electrodes are shown in Fig. 2B for a remembered stimulus at angle $\theta = 0$ (left) and $\theta = 60$ (right) degrees. Correlations were also significant for trials that involved different angles (other stimuli, not shown). It should be noted that both the ensemble index and the component strength of the deep neural fields are single trial measures. The fact that they were significantly correlated implies that electrodes formed ensembles based on the drive that the neural population in the vicinity of each electrode received during each trial. This is similar to intercolumnar synchronization observed in perceptual grouping studies (Gray et al., 1989).

In conclusion, we found that the effective connectivity components of our deep neural field model describe the same clusters as those found using pairwise correlations obtained using LFPs and the method of (Humphries, 2011). This provides an independent validation of our effective connectivity estimates at the single trial level.

3.3. Connectivity kernels correlate with ensemble indices

Recall that, besides connectivity components, our approach also yields the connectivity kernel. Its entries, called *connectivity weights* quantify the strength of the effective connections between the recording sites within each cortical area, see Supplementary Fig. 1B. They multiply input signal from other electrodes that targets a certain electrode measuring activity from a part of the neural ensemble. In other words, 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.

The connectivity kernel enabled us to map the latent space (spanned by the components) to a cortical patch, that the neural ensemble occupies. Later, we will also use the connectivity kernel to predict the electric field generated by the ensemble. First, we assessed whether the connectivity kernel could also identify neural ensembles, similarly to the connectivity components above.

We assumed a Gaussian connectivity profile over space and obtained single trial connectivity kernel estimates $g_C(u_a, u_b)$ (Methods). We also considered a more general weighted Gaussian profile. This is a generalization of the widely used Gaussian kernel (Bojak and Liley, 2010; Atay and Hutt, 2006; Coombes, 2010) that follows from a series expansion (Amindavar and Ritcey, 1994). Here, a Gaussian kernel is weighted by known Hermite polynomials (Abramowitz et al., 1988).

Example connectivity kernels obtained using data from a random trial are shown in Fig. 2C and D. Fig. 2C shows a single Gaussian kernel, while Fig. 2D shows the more general weighted Gaussian. Note that because of the Gaussian profile, only elements around the main diagonal are non-zero Fig. 2.E shows correlations between the two expressions obtained. *R*=87% of connectivity weights of the kernels shown in Fig. 2C and D were correlated.

For simplicity, in the analyses below we used the expression involving a single Gaussian kernel. Similar analyses can be carried out using alternative expressions. First, we asked whether the connectivity kernel could be used to identify neural ensembles, similar to the analyses for connectivity components presented above. We computed correlations between single trial connectivity kernels and ensemble indices, obtained using the method of (Humphries, 2011).

Earlier, we found that ensemble indices were correlated with connectivity components. Correlations were significant for all trials. This implied that electrodes formed ensembles, where electrodes in the same ensemble had neural populations in their vicinity driven by the same input. Similarly, we found that ensemble indices also correlated with the connectivity *kernels* we obtained. Correlations between ensemble indices and connectivity kernels for cued angle at $\theta = 240$ degrees are shown in Fig. 3A . Correlation coefficients are shown in the vertical axis and trials in the horizontal axis. Trials with significant correlations at *p*<0.05 are denoted with asterisk. Overall, 25-40% of single trial kernel estimates correlated with ensemble indices for different angles. The percentage of significantly correlated trials for each cued angle is shown in Fig. 3B. Connectivity components were correlated with ensemble indices across

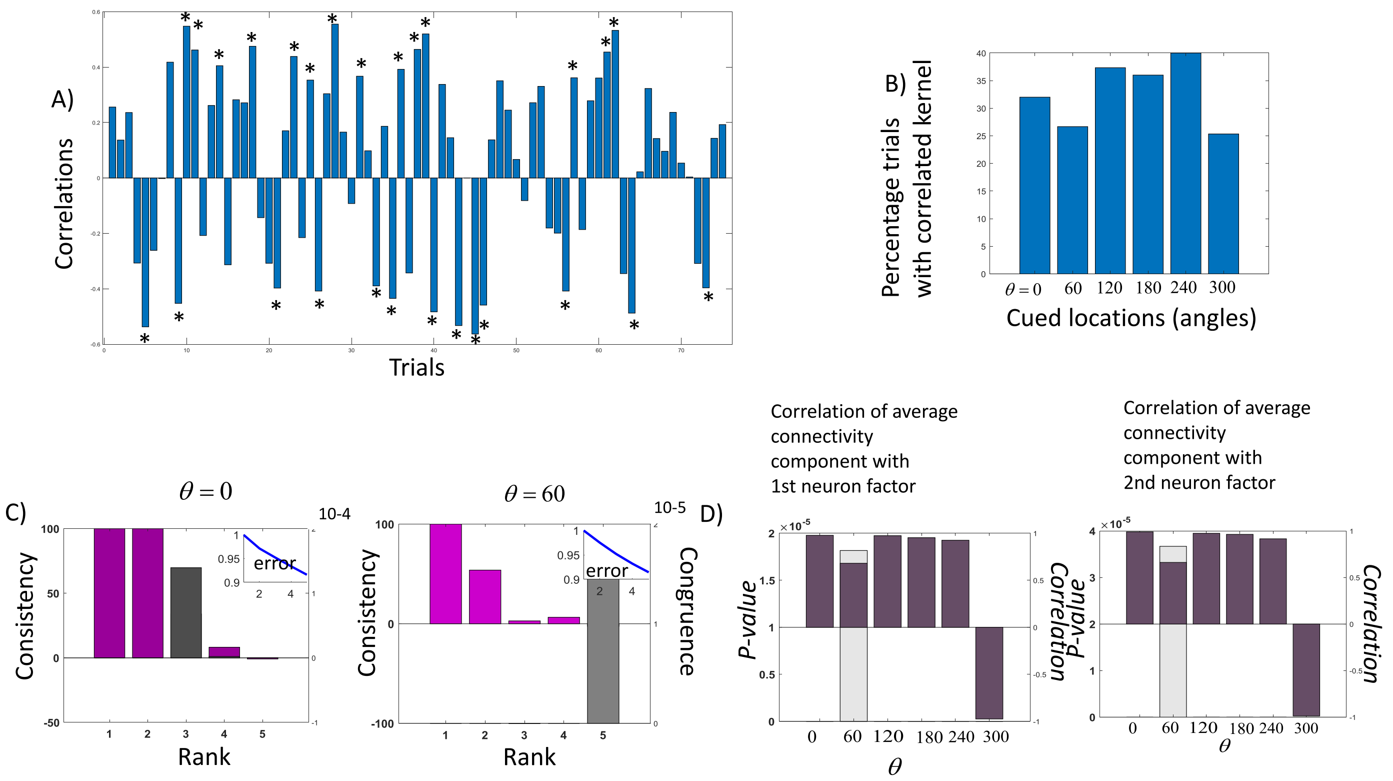

Fig. 3. A. Correlations between single trial ensemble indices and connectivity kernels for cued angle at $\theta = 240$ degrees. Correlation values are shown on the vertical axis. Individual trials are shown on the horizontal axes. Trials with significant correlations at $p<0.05$ are denoted with an asterisk. B. Percentage of significantly correlated trials for each cued angle. Locations are shown on the horizontal axis. Overall, 25-40% of single trial kernel estimates correlated with ensemble indices for different angles. C. Canonical Decomposition. Left and right panels show results for cued angles at θ=0 and θ=60 degrees. The number of factors (rank) is shown on the horizontal axis. Consistency is shown using magenta bars, while congruence is shown using grey bars. Different bars correspond to different ranks. Consistency values are shown on the left vertical axes, while congruence values are shown on the right vertical axes. ALS algorithm reconstruction error is shown in the insets. D. Correlations between connectivity components and first (left panel) and second (right panel) neuron factors obtained via Canonical Decomposition. Cued angles are shown on the horizontal axes. *P*-values (grey bars) are shown on the left vertical axes. Correlation coefficients are shown on the right vertical axes (burgundy bars).

all trials. Connectivity kernels across some of the trials where the same cued angle was maintained, not all. Note that the connectivity kernels were a priori constrained to have a Gaussian (parametric) form, while the components were unconstrained. This explains why the percentage of significant correlations is smaller in the case of kernels. Some ensemble indices show Gaussianity too – but there is nothing intrinsic in the method of (Humphries, 2011) that requires this assumption—which, on the other hand, was intrinsic to the Gaussian profile we assumed for kernels. If ensemble indices are not Gaussian, there are no significant correlations.

All in all, the above results suggest that the connectivity kernels identified the clusters obtained with the pairwise correlation method of (Humphries, 2011). The advantage that these kernels have over the previously considered components is that they describe actual connectivity on a cortical patch, not latent space.

3.4. Connectivity components of deep neural fields correlate with high dimensional SVD components

We also validated the effective connectivity components obtained using our deep neural field approach using a second method. This is based on some old extension of high dimensional SVD, known as Canonical Decomposition (CD), see (Carroll and Chang, 1970) and (Williams et al., 2018) for a recent application. Recall that, to obtain effective connectivity estimates, we trained the biophysical model as an autoencoder. This is similar to classical principal component analysis (PCA): Obtaining the connectivity components amounts to obtaining principal components. Thus, another validation of our components can be achieved by comparing them to components obtained using an SVD approach like CD. Note that CD components do not correspond to single trial ensemble connectivity like the components obtained using our method. CD provides an estimate of average (across trials) connectivity that we had found in (Pinotsis et al., 2017). The authors of (Williams et al., 2018) called this average the neuron mode and suggested that it describes the "*spatial structure that is common across all trials*". Below, we will see that this is similar to the average connectivity component across trials. We will also compare the CD neuron mode with the average connectivity component. We will show that the two correlated perfectly.

CD yields an approximation of our data: this is a three-dimensional LFP array $Y \in \mathbb{R}^{N_T \times N_S \times T}$ (trials x electrodes x time; Methods). The CD approximation includes three matrices (or "modes") that describe patterns over each of the three dimensions: a trial, electrode and time mode. These are dominant patterns in the data, similar to PCA components that describe dominant patterns in time or space. Examples of such PCA components include those obtained with the dimensionality reduction approach of (Wang et al., 2018) that outputs motor timing, i.e. trajectories in a low dimensional domain spanned by PCA components in the time domain; also in (Mante et al., 2013) PCA components included trajectories traced out by neurons representing motion and color.

We used the CD approximation to validate our effective connectivity estimates. Of particular interest for our current analysis is the neuron mode. This is an $N_S \times R$ matrix, where N_S is the number of electrodes and R is a constant, known as rank, that will be estimated below based on some measures from statistics. We assume that each electrode measures activity from a neural population in its proximity. The columns of this matrix are vectors of dimension equal to the number of electrodes.

Each entry is an approximate LFP measured at each electrode (averaged across time and trials). The paper (Williams et al., 2018) used spiking data and the CD approach to obtain the neuron mode. These authors suggested that one can think of the neuron mode as a prototypical firing rate across neurons.

According to (Williams et al., 2018), a neuron mode corresponds to "*the synaptic weights from each latent input to each neuron*". This is similar to the definition of the component strengths included in [29]: *"(component strengths) express the sum of all connectivity weights that target the neurons that contribute to the LFPs observed from each electrode"*. Thus, our connectivity components and CD neuron modes are generalisations of principal components in three dimensions, and they have similar definitions.

Note that here, we used the word "mode" instead of "factor", because the word "factor" is commonly used in the mathematical literature to denote terms in the CD approximation (Kolda and Bader, 2009). In other words, our "neuron mode" is the "neuron factor" of (Williams et al., 2018).

We asked whether our connectivity components and CD neuron modes found using our LFP data were correlated. We considered connectivity components averaged across trials. To find the CD neuron modes we used a standard iterative Alternating Least Squares (ALS) algorithm (ten Berge, 1993). Before obtaining the modes we needed to find the rank, *R* (Methods). This is part of the CD approach. We assumed different values for rank *R*=1,…,5. For each value of *R*, we calculated the sum of squares reconstruction error. This is plotted on the vertical axis appearing in the top right insets of the panels in Fig. 3C. On the horizontal axis, we plotted the number of factors (rank, *R*). The left panel shows results obtained for LFP responses when a cue stimulus was presented at angle $\theta = 0$ degrees. The right panel shows similar results for a cue to $\theta = 60$ degrees.

For both stimuli (and all other angles, Supplementary Fig. 2A), the error reduced with an increasing number of factors (blue line in the insets). This provides a sanity check that the ALS algorithm produces meaningful results as the rank increased. This is similar to PCA, where the more principal components are included, the higher the variance explained. To find the optimal value for rank *R,* we computed two statistical measures: consistency and congruence.

Consistency was introduced as way to obtain the rank, *R*, in CD approximations in the paper (Bro and Kiers, 2003). It uses certain elements of CD theory, known as CD factors, to compute an alternative approximation of the data matrix, known as Tucker3 approximation (Tucker, 1966) and calculates the rank based on this approximation (see Methods for more details). High consistency suggests that there is a trilinear variation in the data. Congruence (also known as similarity) on the other hand, is the result of subtracting the maximum average uncorrected correlation coefficient (UCC) between factors corresponding to different initialisations of the ALS algorithm from 1 (see Methods). This addresses the local minima problem of the ALS algorithm. The lowest the congruence, the more stable the CD approximation (it does not depend on ALS initialisation). In these cases, congruence is small or close to zero, which was the case in our data too (see below).

In short, we chose the optimal rank *R* such that consistency is high and congruence is low. This ensures the CD approximation reflects a trilinear variation in the data and is stable. Consistency and congruence are shown in the right and left vertical axes of the bar plots in the main panels of Fig. 3C. Consistency is shown using magenta bars, while congruence is shown using grey bars. Different bars correspond to different ranks. Rank is shown on the horizontal axes. Consistency values are shown on the left vertical axes, while congruence values are shown on the right vertical axes.

For cues presented at both $\theta = 0, 60$ degrees (left and right panels in Fig. 3C) we obtained high consistency values for *R=1,2* (magenta bars). The same was true also for other angles (Supplementary Fig. 2A). For all values of *R* in Fig. 3C and Supplementary Fig. 2A, congruence was very small (grey bars). Its order was 10^{-4} for $\theta = 0$ and 10^{-5} for $\theta = 60$ degrees . Thus, in what follows, we used the CD approximation with *R*=2. For all angles, this corresponded to high consistency and low congruence. For *R*=2, the neuron mode was a matrix of dimensionality $N_S \times 2$. Fixing $m = M*$, where $M* = 1, 2$ we obtained two vectors $\vec{b}_{jM*}$, *j =1,…,* N_S that approximate average LFPs across time and trials. These are the two *columns* of the neuron mode. Following (Williams et al., 2018), we call these vectors the 1st and 2nd neuron factors.

Recall that, each connectivity component is also a vector of length N_S. In (Pinotsis et al., 2017), we studied the first four connectivity components (similar to principal components in PCA). Here we focused on the first, as this explains most of the data variance similarly to the neuron factors that comprise the neuron mode. This explained about 35% of variance (Supplementary Fig. 2B). Keeping up to 4 components, variance explained increased to about 60%. We asked whether the two neuron factors (recall *R=2* above) were correlated with the first connectivity component. For cues presented at every angle ($\theta = 0, 60, 120, 180, 240, 300$ degrees), we computed the correlation coefficient and corresponding *p*-value between the 1st and 2nd neuron factors and the connectivity component averaged across trials. These are shown in Fig. 3D. We found that the average connectivity component was significantly correlated with both the 1st and 2nd neuron factors. Correlations were significant for all angles. *P*-values (grey bars) are shown on the left vertical axes of left (1st neuron factor) and right (2nd neuron factor) panels. Only the 2 largest *p*-values for $\theta = 60$ degrees are shown. All other *p*-values were much smaller than $p<10^{-5}$. Thus, they are not visible in the plot. The corresponding values of the correlation coefficient *r* are shown on the right vertical axes (burgundy bars). They are all high. Correlation coefficients were *r*>0.9 for all angles except $\theta = 60$ for which *r*>0.7. Note that the CD approximation assumes a trilinear variation in the data, while the autoencoder approach we used to obtain our components is nonlinear. Thus, the remaining dissimilarity can be explained by a nonlinear mixing of latent states afforded by an autoencoder.

To sum up, we compared our approach for performing dimensionality reduction to a high dimensional SVD approach, known as Canonical Decomposition (CD (Carroll and Chang, 1970; Williams et al., 2018)). We found that the effective connectivity components obtained using our approach correlates significantly with the neuron factors obtained using CD. This provides a second, independent validation of our approach.

All in all, we compared the effective connectivity components with results obtained using pairwise correlations and the latent states of a high dimensional SVD approach. Our components correlated with those found using alternative methods. Thus, all three methods found a similar structure of the latent space within which neural activity evolves, while neurons are maintaining cued angles.

3.5. *Stable electric fields emerge from neural ensembles that represent the same cued angle in different trials*

To sum so far, we first found the latent space associated with maintenance of a cued angle (connectivity components). We then mapped this space to a cortical patch occupied by a neural ensemble—and obtained the connectivity kernels. These describe the exchange of information during cue maintenance. We found that the corresponding connectivity weights correlated significantly with single trial ensemble indices obtained using pairwise correlations (Humphries, 2011) across a large percentage of trials.

Recall that the connectivity weights scale the input signal from other electrodes that targets a certain electrode measuring activity from a part of the neural ensemble. Having obtained these weights, we could then predict the Electric Field (EF) generated by the neural ensemble. The connectivity kernels describe how neurons communicate via electric signals sent from one part of the ensemble patch to the other. These electric signals generate the EF. Below we used the connectivity kernel and the deep neural field model to simulate EFs. We wanted to test if EFs were similar across trials where the same cued angle was maintained.

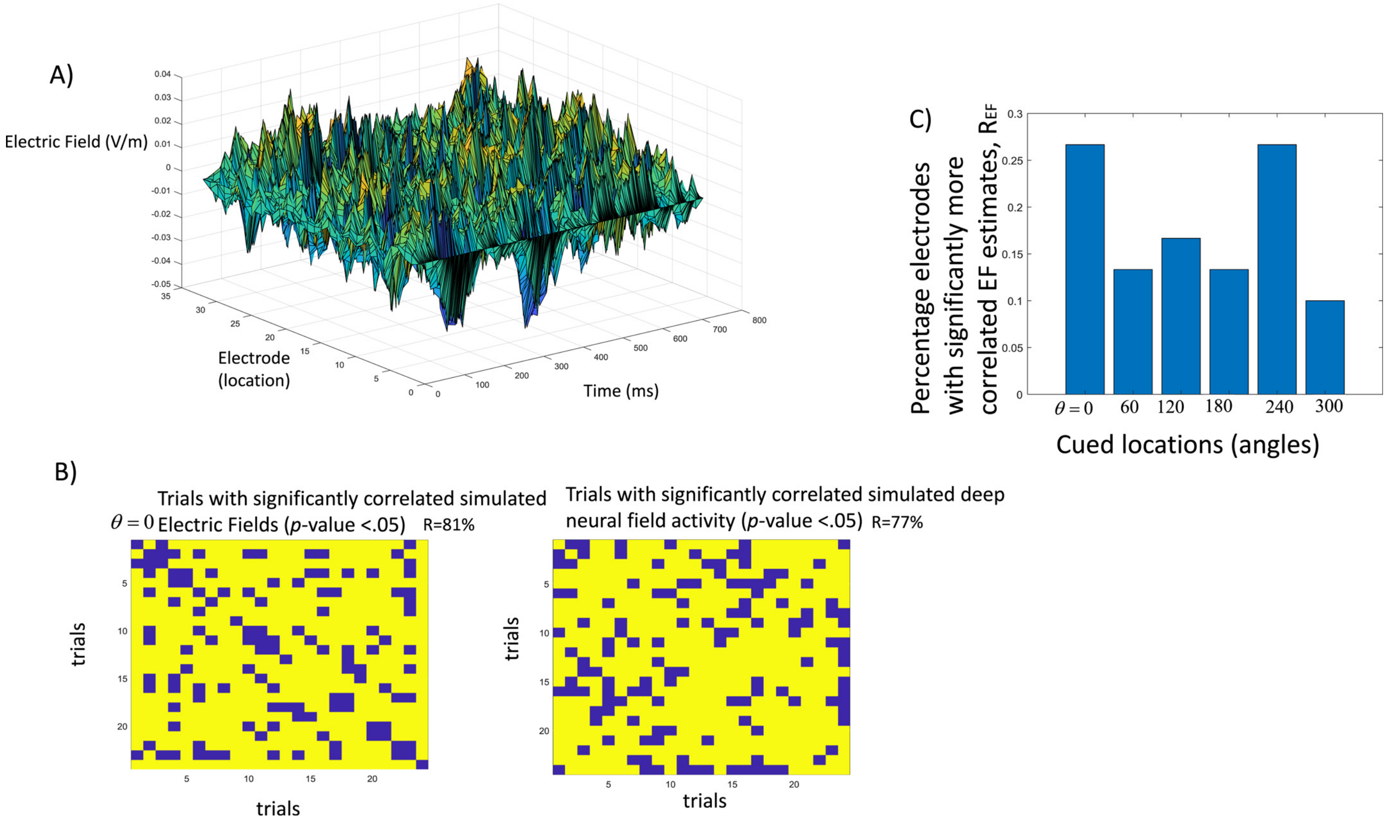

Fig. 4. A. Example of simulated electric field (EF) using the bidomain model. The EF amplitude is shown on the vertical axis (V/m), while the two horizontal axes show the electrode (location on the cortex) and time (ms). B. (Left) *P*-values of correlations between single trial EF amplitudes. These correspond to EFs generated by neural ensembles maintaining a cued angle at $\theta = 0$ degrees. Yellow entries in the correlation matrix denote significant *p*-values, *p <.05*. The percentage of significantly correlated single trial EF estimates is shown on the top right corner, *R*=80%. (Right) *P*-values of correlations between single trial deep neural field data. Yellow entries denote significant *p*-values as in Fig. 4B. Percentage of correlated trials is lower than Fig. 4B (Left). C. Percentage of electrodes where electric field estimates were correlated across a larger number of trials compared to neural activity estimates, for different stimuli (angles).

LFPs can be thought of as proxies to electric fields. However, it is not clear what their source is –and whether they are solely produced by neurons that participate in an ensemble or neighbouring neurons too. In other words, the ground truth regarding neural sources that produce the ensemble EF is unknown. Thus, we used our deep neural field model and the bidomain model to obtain predictions of ensemble EFs. The deep neural field model stands in for an in silico implementation of a neural ensemble. The bidomain model has been used to predict the electric field generated by biological tissues, like the cardiac muscle (Henriquez, 1993; Roth, 1997) and auditory brainstem (Goldwyn et al., 2017). To estimate the extracellular EF, the model requires only a measurement of the transmembrane potential V^m (Methods). The bidomain model neglects ephaptic coupling and electromagnetic wave effects – that are small compared to electric effects. It yields the EF in the extracellular space by computing the Fourier transform of V^m measurements and an analytical expression based on Bessel functions of the first and second kind.

Here, we obtained two EF estimates. First, EF estimates based on deep neural field model predictions of transmembrane potentials V^m. These are simulated potentials after training the deep neural field model with all available data. We called them *simulated EFs*. Second, EF estimates based on real LFPs. These did *not* use the deep neural field model. LFPs were used as proxies for transmembrane potentials and replaced the simulated transmembrane potential from the neural field model above. We called the EFs obtained using real LFPs and the bidomain model, *real EFs*.

An example simulated EF estimate is shown in Fig. 4A. The EF amplitude is shown on the vertical axis (V/m), while the two horizontal axes show the electrode number (ID; location on the cortex) and time (ms). *P*-values of correlations between EF amplitudes are shown in Fig. 4B. These correspond to EFs generated by neural ensembles maintaining a cued angle at $\theta = 0$ degrees. We here considered EF estimates from trials where our connectivity kernel correlated with the findings of (Humphries, 2011) (correlated trials). To obtain these estimates, we first simulated neural activity using our deep neural field model. Variance explained was about 40% for all stimuli (cued angles, Supplementary Fig. 4A and see also Fig. 9 in (Pinotsis et al., 2017); there we had used all trials, instead of correlated trials that we used here). After simulating neural activity, we computed EF estimates using the bidomain model. We asked whether they correlated across trials for the same electrode.

Yellow entries in the correlation matrix denote significant *p*-values, *p <.05*, for an electrode at the edge of the patch and cued angle $\theta = 0$. The percentage of significantly correlated single trial EF estimates is shown on the top right corner of the left panel in Fig. 4B, *R*=81%. Similarly, we found that EF amplitudes were also correlated across all trials and other angles with *R*= 70-80% (see Supplementary Fig. 3). We also computed the corresponding correlation using deep neural field activity estimates at the same electrode for the same cued angle. The percentage of significantly correlated trials was *R*=77% (top right corner of right panel in Fig. 4B). Note this is lower than the percentage of correlated trials computed using EF estimates obtained above.

We then asked if the same result holds across many electrodes: that is, if the percentage of correlated trials was lower when using single trial neural activity (deep neural field simulated data) compared to EF estimates. If it was, that would mean that neural activity was more variable than EF recordings. In other words, several distinct configurations of neural sources led to the same field.

To sum up, we asked whether for the same electrode (location on the ensemble patch), there was a significant difference in the percentage, *R,* of correlated trials obtained using: (1) reconstructed neural activity, which we called, R_{NA} and (2) reconstructed EF estimates, called R_{EF}. We repeated this for all electrodes and summarized results for each cued angle. Results are shown in Fig. 4C. For all stimuli, a larger number of electrodes had reconstructed single trial EF estimates that were correlated across trials, $R_{EF,}$ compared to reconstructed neural activity estimates, $R_{NA,}$:

Bars in Fig. 4C show the percentage *electrodes, Q,* where R_{EF} was significantly higher than $R_{NA,}$ *Q*=11-27%. To test for statistical significance, we used a Fischer exact test. This allows one to find differences between binomial distributions. Here, the binary variable describes whether a single trial EF estimate or neural activity estimate was correlated or not (the entries of matrices in Fig. 4B). The null hypothesis was that there was no difference in the percentage of correlated trials (at the 5% significance level). We repeated the analysis for each electrode.

We found that a large part of electrodes had $R_{EF} > R_{NA}$ [2] . In other words, electric field estimates were more often correlated across trials, i.e. more stable, compared to neural activity estimates[3]. In the next section, we will see that stronger stability of the EF compared to neural activity was also confirmed by decoding analyses. Training accuracy based on neural activity was significantly lower than accuracy based on EF estimates. This also suggests that information contained in delay neural activity was less stable than that contained in the electric field.

Not all electrodes had significant differences, $R_{EF} > R_{NA}$, because different stimuli activate different parts of the patch. Differences in connectivity components between stimuli are localised within those parts (see Fig. 2 of (Pinotsis et al., 2017) and relevant discussion). This suggests that only parts of the patch (certain electrodes) will be sensitive to changes of stimuli. EF and neural activity estimates measured at those electrodes will be correlated, that is, stable across trials (not the whole patch).

To sum, we found that electric fields were correlated across trials where the same cue was maintained. Further, the number of electrodes (locations) where this happens was larger than the corresponding number when neural activity was correlated. All in all, the above results suggest that stable electrical fields emerge from high-dimensional ever-shifting neuronal activity patterns of neural ensembles during trials where the same cue was maintained in memory networks. Having shown that electric fields are stable, we turned to the information carried by them and asked if it was stable too.

3.6. *Emergent electric fields carry unique information about working memory content*

Finally, we asked whether EF produced by neural ensembles carried information about working memory content. We assessed whether EF estimates obtained using our approach were consistently different among neural ensembles that maintained different cued angles. In other words, we tested if we could distinguish between memorized cues based on EFs. If we could, this means that EFs can be uniquely associated with different working memories that are used to perform the task. To formally test this hypothesis, we used the EF estimates as classification features of different trials by cued angle. We used 450 trials and held out 20% of the data as a test set. We used simulated and real EFs as classification features and two different algorithms, Naïve Bayes and diagonal LDA. These are among the most commonly used.

The results of our analyses are shown in Fig. 5 (using Naïve Bayes) and Supplementary Figure 6 (using diagonal LDA). Decoding accuracy values are shown on the vertical axis, while the corresponding electrodes (patch locations) are shown on the horizontal axis. We performed permutation tests, after shuffling class labels (cued angles) around. Blue bars denote observed accuracy values. Orange bars denote the maximum of the shuffled distribution. If blue bars are larger than orange, the observed accuracy is significantly higher than chance (max of shuffled estimates) at the *p*=0.01 level. This was the case for over half of the electrodes and accuracies obtained using simulated EFs (Fig. 5A). The corresponding train and test confusion matrices are shown in Supplementary Fig. 5A. These are averages over all electrodes. Accuracies were very similar for all stimuli[4].

Recall that simulated EFs above were obtained from connectivity components, which, in turn, were obtained after training the neural field model on the whole dataset. Thus, decoding features contain some previous information from the data, something often referred to as data leakage. To address this, we computed the decoding accuracy using real EFs as features. Recall also that these were obtained after using LFPs as proxies for transmembrane potential. Thus, the corresponding accuracy will not be biased and includes out-of-sample validation based on a 20% held out test set. Similarly to simulated EFs, a permutation test confirmed accuracy significantly higher than chance at the $p<0.01$ level for over half of the electrodes (Fig. 5B). The corresponding average confusion matrices are shown in Supplementary Fig. 5B. Accuracies based on simulated EFs are similar to those obtained using LFPs (real EFs). To test for their equivalence, we used the TOST procedure (Lakens et al., 2018). We found that accuracies were the same, *t(31)*=-2.05, *p*=0.02 (assuming that a meaningful difference would be larger than 2%).

To summarize, we found that simulated and real EF estimates differed systematically depending on the exact cued angle; they were uniquely associated with the remembered stimulus. The above results confirm our earlier result that EFs were stable across trials where the same cued angle was maintained. They contained unique information about the remembered stimulus, that seems to be preserved across trials.

The theory of electromagnetism suggests that if EFs are stable, then the differences of the corresponding extracellular potentials should also be stable. These are known as *Gauge functions* (Methods). They are obtained by subtracting real LFPs recorded in different trials where the same cued angle was maintained. We thus asked if we could distinguish cued angles when using Gauge functions as decoding features. If we could, this would provide an alternative confirmation of our results. Crucially, Gauge functions do not rely on the validity of neither bidomain nor the deep neural field model. Thus, if they can distinguish between cued angles this is a confirmation of our result independent of these models.

The results of our analyses are shown in Fig. 5C. As before, blue and orange bars correspond to observed accuracy and chance accuracy (maximum of the shuffled distribution) respectively. Accuracy obtained Gauge is similar to the results in Fig. 5A and B. Thus, Gauge functions are also stable and contain information about the cued angle.

Finally, we repeated the decoding analyses using simulated neural activity (from the deep neural field model). Permutation test results are shown in Fig. 5D. Accuracy was higher than chance ($p<0.01$) for over half of the electrodes. A one sided, *Welch* test also found that training accuracy based on neural activity was significantly smaller than accuracy obtained using real EFs *t(31)*=-8.2, $p<0.001$. The corresponding confusion matrices are shown in Supplementary Fig. 6C. Correctly classified trials were fewer than those obtained using real and simulated EFs. Thus, neural activity did not contain the same stable information as the electric field. This is in accord with our earlier result (Fig. 4).

[2] For all cued angles except θ=60 degrees (Supplementary Fig. 4B).

[3] During memory delay, some part of neural activity will be stable (attractor dynamics). This is not always picked up by EF estimates measured at certain locations (electrodes) due to assumptions in the bidomain model (isotropic field, homogeneous resistivity, infinite neural source etc; Supplementary Fig. 4B).

[4] Except for θ=0 degrees, which is slightly higher.

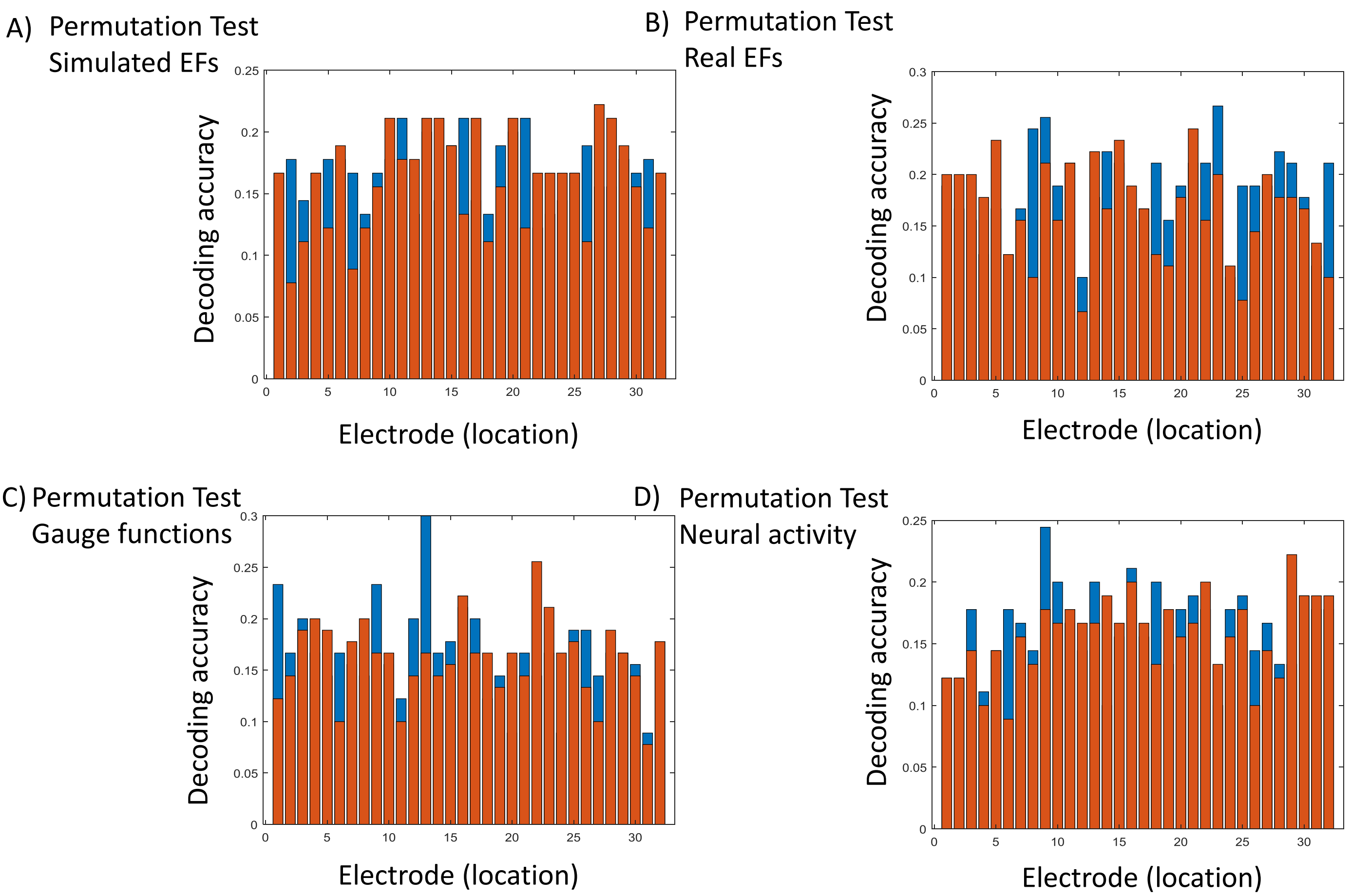

Fig. 5. A. Permutation test of decoding accuracy based on simulated EF estimates. Accuracy values are shown on the vertical axes, while the corresponding electrodes (patch locations) are shown on the horizontal axes. Blue bars show observed accuracy estimates. Orange bars show the maximum obtained accuracy after performing N=100 permutations. For those electrodes that unshuffled estimates are higher than the maximum of the distribution after shuffling, decoding accuracy is significantly higher than chance at the p=0.01 level. Over half of the electrodes have higher accuracy (blue bars) than the maximum of the distribution obtained after shuffling (orange bars). B. Same as in A. after replacing simulated EFs by real EFs. An equivalence test found that simulated and real EF accuracy estimate are the same (see text). C. Same as in A. after replacing simulated EFs by Gauge functions that do not depend on the neural field or dipole models. D. Same as in A. after replacing simulated EFs by neural activity estimates. A Welch test found that training accuracy obtained using neural activity was smaller than the corresponding accuracy obtained using real EFs (see text). Results shown in all panels were obtained using a Naïve Bayes classifier.

All in all, we found that electric fields provided higher than chance decoding accuracy in predicting the remembered stimuli (cued angles). Thus EFs contained unique information about working memory content needed to perform the task.

4. Discussion

We analyzed monkey LFP data from a spatial working memory task (Jia et al., 2017; Pinotsis et al., 2017). We found that stable electrical fields emerge from high-dimensional ever-shifting neuronal activity patterns of neural ensembles in the brain. We trained a biophysical neural network model as an autoencoder that learned to maintain spatial locations. This provided latent variables describing the connectivity of neural ensembles, which we called 'connectivity components'. We also reconstructed single trial effective connectivity estimates, ' connectivity kernels' (Pinotsis et al., 2017). These describe information flow within the neural ensemble; in other words, the exchange of electric signals between neurons forming an ensemble. Crucially, this distinguishes our approach from other dimensionality reduction approaches (Jazayeri and Ostojic, 2021; Cunningham and Byron, 2014). Our approach maps the latent space to a cortical patch. It goes beyond dimensionality reduction and reconstructs information flow.

Mathematically, the connection weights (kernel) can be thought of as the probability of having connections between neural populations forming a neural ensemble (Pinotsis et al., 2017). Other methods to obtain the probability function include splines (Gehringer and Redner, 1992) and tools from complex systems (Heinz, 2013). We will systematically consider these methods elsewhere. We here used a Restricted Maximum Likelihood (ReML) algorithm for obtaining the connectivity components (Pinotsis et al., 2017). This optimizes the same cost function used in variational autoencoders, called Free Energy (FE; also known as Evidence Lower Bound, ELBO). ReML does not require an explicit cross validation step (the E-step is embedded in the M-step after substituting the posterior variance). While cross validation (CV) partitions data in test and training sets, ReML prevents overfitting by penalizing for model complexity. The relationship between CV and FE for assessing source reconstruction error in the context of neuroimaging data has been systematically studied in several studies including (Troebinger et al., 2014; Little et al., 2018). CV error and FE are correlated (Bonaiuto et al., 2018).

We found that the connectivity components were highly correlated with latent factors extracted by Canonical Decomposition (high dimensional SVD) (Carroll and Chang, 1970; Williams et al., 2018; Kiers, 1998). We also found that connectivity components and kernels were correlated with cluster indices obtained using unsupervised clustering (Humphries, 2011).

Connectivity components and kernels describe the effective connectivity between different neurons forming a neural ensemble: How electric signals and information are exchanged between them. Using kernels and the classic dipole theory of electromagnetism, we reconstructed the electric fields (EFs) produced by a neural ensemble. We reconstructed the electric fields using four steps (Fig. 1A). We first reformulated a neural network (described by Wilson Cowan equations) as a neural field model and then a Gaussian Linear Model (GLM; step 1). We trained this model as an autoencoder using LFP data. This allowed us to obtain the latent states (connectivity components; step 2). Then, using inverse problem theory, we obtained the corresponding connectivity kernels (step 3). Finally, electromagnetism (dipole theory) allowed us to predict the electric field generated by an ensemble (step 4).

We found that different remembered locations resulted in different electric fields. These fields were highly stable across trials yet, at the level of specific circuits there was more variability (representational drift (Driscoll et al., 2017; Deitch et al., 2020)). We trained a neural field model using single trial LFP data and obtained its connectivity. This model described a neural ensemble. We then reconstructed electric field and neural activity estimates generated by the ensemble during delay when the same location was remembered and looked at the percentage of correlated trials. The percentage of electric field estimates that were significantly correlated across trials was higher than the corresponding percentage obtained using neural activity estimates and this was replicated across many electrodes and stimuli.

This result is also supported by the theory of electromagnetism. The same electric field can arise from different combinations of specific neurons and networks (electromagnetic sources and sinks (Perkins and Perkins, 2000)). This is known as non-uniqueness of the electromagnetic inverse problem: One cannot find the exact sources by measuring electric fields alone (Jackson, 1999). This non-uniqueness implies that neural sources changed between trials but the electric field was stable: Across like trials, where the same memory was maintained, the inputs entering a given network changed. Electromagnetism predicts that neural sources will reconfigure themselves to accommodate these inputs but the overall electric field will be the same. When inputs change, the neural sources change but the electric field will not. This can explain the observed variability in the patterns of neurons forming a neural ensemble. Here, we confirmed this hypothesis using LFP data and computational modeling. In future work, we will experimentally test the stability of the electric field.

Finally and importantly, different EFs were uniquely associated with different working memories needed to perform the experimental task successfully. To support this, it was shown that the EF estimates provided higher-than-chance decoding accuracies in predicting the remembered stimuli (i.e., cued angles). Further, training accuracy based on neural activity was lower and correctly classified trials were fewer than those obtained using EFs. Neural activity is less stable than the electric field.

Our model assumes that LFPs contain information about the excitation to inhibition (E/I) balance, despite being an aggregate measure of neural activity obtained from both excitatory and inhibitory populations. This is supported by both computational (Mazzoni et al., 2013; Glomb et al., 2021; Kang et al., 2020) and empirical (Trakoshis et al., 2020; Haider et al., 2006) studies, see also (Gao et al., 2017) for a recent discussion. In particular, a large body of work by us and others using Dynamic Causal Models (DCM) has shown that it is possible to infer E/I ratios assuming that LFPs arise as a result of certain synaptic currents, usually AMPA and GABA_A currents, see e.g. (Pinotsis et al., 2016; Pinotsis et al., 2017; Friston et al., 2015; Legon et al., 2016; Hamburg et al., 2019; Pinotsis and Miller, 2020). In future work, we will use separate recordings (depolarization or spike rates) from excitatory and inhibitory populations, to reconstruct excitatory and inhibitory activity separately.

In general, there are three different ways one can reduce the dimensionality in large, brain imaging datasets. Because these datasets involve three-way matrices (tensors) with dimensions (*time* x *neurons* x *trials*), three different sets of principal components (PCTs) can be obtained, in either (i) time; (ii) neurons (or channels) or (iii) trials domain. The outputs of this process are trajectories – i.e. collections of points– in domains spanned by the corresponding PCTs. For example, in (Wang et al., 2018) the output was motor timing (i.e. trajectories in a low dimensional domain spanned by time PCTs—ie. temporal evolution of population activity); while (Mante et al., 2013) obtained trajectories traced out by neurons in the motion and color domains (because PCTs along the second dimension, neurons, correspond to behaviourally relevant variables; neurons are grouped into PCTs depending on their tuning preferences). Finally, PCTs can be defined in the trial domain and the corresponding trajectories can then be used to obtain estimates of trial to trial variability. This can e.g. reveal changes in excitability of neural populations due to attention (Kanashiro et al., 2017) and ongoing cognitive variables in general (Nienborg et al., 2012).

We here characterized the latent states during memory maintenance using biophysically informed models, neural fields. Because these models are defined in the time and neuron (i.e. space) domain, this reduction provides insights in both those domains. This, in turn, can help one understand the relation between representational drift and properties, like criticality (Maturana et al., 2020; Bak et al., 1988). Cortical dynamics in critical regimes are characterised by a co-occurrence of different temporal frequencies at different spatial scales (Freeman, 2003). Both frequencies and spatial scales can be described by the connectivity components and principal axes obtained after training a neural field model (Pinotsis et al., 2017). In that earlier work, we showed that single trial principal axes predict the characteristic Lyapunov exponents that determine the timescales at which the system returns to equilibrium after perturbations, commonly known as critical slowing (Grindrod and Pinotsis, 2011; Pinotsis and Friston, 2011). The connectivity components describe different neural ensembles, i.e. spatial patterns or combinations of neurons that maintain cued angles. These change between trials (representational drift). Thus, by studying single trial estimates of components and principal axes, one can link critical slowing with ensembles and representational drift. This will be considered elsewhere.

In short, we found that stable EFs emerge from high-dimensional ever-shifting neuronal activity patterns of neural ensembles in the brain. These EFs were robust across experimental trials where the same location was maintained, despite the continually changing neuronal activity, something known as the 'representational drift'. Also, the low-dimensional emergent electrical fields carry information about working memories.

The stability of the electric field can allow the brain to control the latent variables (e.g., oscillations) that give rise to the same memory. We suggest that the electric field does not just emerge from the representational drift. It also helps sculpt and herd that general pattern of traffic. In other words, electric fields can act as “guard rails” that funnel the higher dimensional variable neural activity along stable lower-dimensional routes. We will test this hypothesis elsewhere. The low-dimensional stability in electric fields might help the brain perform computations, by allowing latent states to be reliably transferred between brain areas, in accord with modern engram theory (Ryan et al., 2015). This is also in accord with the theory of Synergetics (Basar et al., 1983, Fuchs et al., 2000; Haken, 2006; Jirsa and Kelso, 2000). The electric field can be viewed as a control variable similar to energy (Haken, 1985) and attention signals (Ditzinger and Haken, 1989) that evolves more slowly than the latent variables that represent information. In other words, there might be a temporal hierarchy comprising the timescales of control parameters (e.g. electric field), order parameters (e.g. latent variables (Gallego et al., 2020; Yu et al., 2008)) and enslaved parts (e.g. oscillations/spiking (Haken, 2006)).

All in all, our results and related work suggest that the electric field is conserved in memory networks and allows latent variables from different brain areas to interact and produce behavior. Although the exact neurons forming a neural ensemble differ from trial to trial (representational drift), the electric field is stable and contains unique information about the remembered stimulus, that seems to be preserved across trials.

Data and code availability statement

The data and analysis tools are available from the corresponding author upon reasonable request.

Credit authorship contribution statement

Dimitris A. Pinotsis: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Earl K. Miller:** Conceptualization, Validation, Investigation, Data curation, Resources, Writing – review & editing, Visualization, Funding acquisition.

Acknowledgements

This work is supported by UKRI ES/T01279X/1, ONR MURI N00014-16-1-2832, and the JPB Foundation. We are also grateful to Dr Jay Myung, Jorge Chang and anonymous reviewers for useful comments.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119058.

References

Steinmetz, N.A., Koch, C., Harris, K.D., Carandini, M., 2018. Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. Curr. Opin. Neurobiol. 50, 92–100.

Katlowitz, K.A., Picardo, M.A., Long, M.A., 2018. Stable sequential activity underlying the maintenance of a precisely executed skilled behavior. Neuron 98, 1133–1140 e3.

Gallego, J.A., et al., 2018. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. Nat. Commun. 9, 1–13.

Mastrogiuseppe, F., Ostojic, S., 2018. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. Neuron 99, 609–623 e29.

Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. arXiv preprint arXiv:2107.04084 (2021).

Urai, A.E., Doiron, B., Leifer, A.M. & Churchland, A.K. Large-scale neural recordings call for new insights to link brain and behavior. arXiv preprint arXiv:2103.14662 (2021).

Cunningham, J.P., Byron, M.Y., 2014. Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. 17, 1500–1509.

Pang, R., Lansdell, B.J., Fairhall, A.L., 2016. Dimensionality reduction in neuroscience. Curr. Biol. 26, R656–R660.

Pandarinath, C., et al., 2018. Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods 15, 805–815.

Churchland, M.M., et al., 2010. Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat. Neurosci. 13, 369–378.

Mongillo, G., Rumpel, S., Loewenstein, Y., 2017. Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory. Curr. Opin. Neurobiol. 46, 7–13.

Attardo, A., Fitzgerald, J.E., Schnitzer, M.J., 2015. Impermanence of dendritic spines in live adult CA1 hippocampus. Nature 523, 592–596.

Ziv, N.E., Brenner, N., 2018. Synaptic tenacity or lack thereof: spontaneous remodeling of synapses. Trends Neurosci. 41, 89–99.

Clopath, C., Bonhoeffer, T., Hübener, M., Rose, T., 2017. Variance and invariance of neuronal long-term representations. Philos. Trans. R. Soc. B Biol. Sci. 372, 20160161.

Lu, J., Zuo, Y., 2021. Shedding light on learning and memory: optical interrogation of the synaptic circuitry. Curr. Opin. Neurobiol. 67, 138–144.

Kozachkov, L., Lundqvist, M., Slotine, J.J., Miller, E.K., 2020. Achieving stable dynamics in neural circuits. PLoS Comput. Biol. 16, e1007659.

Driscoll, L.N., Pettit, N.L., Minderer, M., Chettih, S.N., Harvey, C.D., 2017. Dynamic reorganization of neuronal activity patterns in parietal cortex. Cell 170, 986–999 e16.

Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. bioRxiv (2020).

Marder, E., Goeritz, M.L., Otopalik, A.G., 2015. Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. Curr. Opin. Neurobiol. 31, 156–163.

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

Rigotti, M., et al., 2013. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585.

Rule, M.E., O'Leary, T., Harvey, C.D, 2019. Causes and consequences of representational drift. Curr. Opin. Neurobiol. 58, 141–147.

Kappel, D., Legenstein, R., Habenschuss, S., Hsieh, M. & Maass, W. A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning. Eneuro 5, (2018).

Park, H.J., Friston, K., 2013. Structural and functional brain networks: from connections to cognition. Science 342.

Shine, J.M., et al., 2016. The dynamics of functional brain networks: integrated network states during cognitive task performance. Neuron 92, 544–554.

Westphal, A.J., Wang, S., Rissman, J., 2017. Episodic memory retrieval benefits from a less modular brain network organization. J. Neurosci. 37, 3523–3531.

Jackson, J.D., 1999. Classical Electrodynamics. American Association of Physics Teachers.

Jia, N., et al., 2017. Decoding of intended saccade direction in an oculomotor brain–computer interface. J. Neural Eng. 14, 046007.

Pinotsis, D.A., Brincat, S.L., Miller, E.K., 2017. On memories, neural ensembles and mental flexibility. Neuroimage 157, 297–313.

Gao, P., Ganguli, S., 2015. On simplicity and complexity in the brave new world of large-scale neuroscience. Curr. Opin. Neurobiol. 32, 148–155.

Humphries, M.D., 2011. Spike-train communities: finding groups of similar spike trains. J. Neurosci. 31, 2321–2336.

Schwartz, B.L., Chauhan, M., Sadleir, R.J., 2016. Analytic modeling of neural tissue: I. A spherical bidomain. J. Math. Neurosci. 6, 1–20.

Ermentrout, G.B., Cowan, J.D., 1979. A mathematical theory of visual hallucination patterns. Biol. Cybern. 34, 137–150.

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

Wilson, H.R., Cowan, J.D., 1973. Mathematical theory of functional dynamics of cortical and thalamic nervous-tissue. Kybernetik 13, 55–80.

Grossberg, S., 1967. Nonlinear difference-differential equations in prediction and learning theory. Proc. Nat. Acad. Sci. U. S. A. 58, 1329.

van Hemmen, J.L., 2004. Continuum limit of discrete neuronal structures: is cortical tissue an "excitable" medium? Biol. Cybern. 91, 347–358.

Eccles, J.C., Fatt, P., Koketsu, K., 1954. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. J. Physiol. 126, 524–562.

Song, H.F., Yang, G.R., Wang, X.J., 2016. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. PLoS Comput. Biol. 12, e1004792.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72, 320–338.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.

Lancichinetti, A., Fortunato, S., 2012. Consensus clustering in complex networks. Sci. Rep. 2, 1–7.

Bruno, A.M., Frost, W.N., Humphries, M.D., 2015. Modular deconstruction reveals the dynamical and physical building blocks of a locomotion motor program. Neuron 86, 304–318.

Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika 35, 283–319.

Kolda, T.G., Bader, B.W., 2009. Tensor decompositions and applications. SIAM Rev. 51, 455–500.

ten Berge, J.M., 1993. Least Squares Optimization in Multivariate Analysis. DSWO Press, Leiden University Leiden.

Williams, A.H., et al., 2018. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. Neuron 98, 1099–1115 e8.

Bro, R., Kiers, H.A., 2003. A new efficient method for determining the number of components in PARAFAC models. J. Chemometr. A J. Chemometr. Soc. 17, 274–286.

Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311.

Kiers, H.A., 1998. A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. J. Chemometr. A J. Chemometr. Soc. 12, 155–171.

Goldwyn, J.H., McLaughlin, M., Verschooten, E., Joris, P.X., Rinzel, J., 2017. Signatures of somatic inhibition and dendritic excitation in auditory brainstem field potentials. J. Neurosci. 37, 10451–10467.

Mc Laughlin, M., Verschooten, E., Joris, P.X., 2010. Oscillatory dipoles as a source of phase shifts in field potentials in the mammalian auditory brainstem. J. Neurosci. 30, 13472–13487.

Plonsey, R., 1974. The active fiber in a volume conductor. IEEE Trans. Biomed. Eng. 371–381.

Rush, S., Driscoll, D.A., 1969. EEG electrode sensitivity-an application of reciprocity. IEEE Trans. Biomed. Eng. 15–22.

Henriquez, C.S., 1993. Simulating the electrical behavior of cardiac tissue using the bidomain model. Crit. Rev. Biomed. Eng. 21, 1–77.

Roth, B.J., 1997. Electrical conductivity values used with the bidomain model of cardiac tissue. IEEE Trans. Biomed. Eng. 44, 326–328.

Abramowitz, M., Stegun, I.A., Romer, R.H., 1988. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. American Association of Physics Teachers.

Darrigol, O., 2003. Electrodynamics from Ampere to Einstein. Oxford University Press.

Maxwell, J.C., 2021. On Faraday's Lines of Force. Good Press.

Gehringer, K.R., Redner, R.A., 1992. Nonparametric probability density estimation using normalized b–splines. Commun. Stat.-Simul. Comput. 21, 849–878.

Amindavar, H., Ritcey, J.A., 1994. Padé approximations of probability density functions. IEEE Trans. Aerosp. Electron. Syst. 30, 416–424.

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

Mersmann, A., 1995. Crystallization technology handbook. Drying Technol. 13, 1037–1038.

Pinotsis, D.A., Siegel, M., Miller, E.K., 2019. Sensory processing and categorization in cortical and deep neural networks. Neuroimage 202, 116118.

Siegel, M., Buschman, T.J., Miller, E.K., 2015. Cortical information flow during flexible sensorimotor decisions. Science 348, 1352–1355.

Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. 13, 407–420.

Lindén, H., et al., 2011. Modeling the spatial reach of the LFP. Neuron 72, 859–872.

Bojak, I., Liley, D.T.J., 2010. Axonal velocity distributions in neural field equations. PLoS Comput. Biol. 6, e1000653.

Atay, F.M., Hutt, A., 2006. Neural fields with distributed transmission speeds and long-range feedback delays. SIAM J. Appl. Dyn. Syst. 5, 670–698.

Deco, G., Jirsa, V.K., Robinson, P.A., Breakspear, M., Friston, K., 2008. The dynamic brain: from spiking neurons to neural masses and cortical fields. PLoS Comput. Biol. 4.

Coombes, S., 2010. Large-scale neural dynamics: simple and complex. Neuroimage 52, 731–739.

Appelle, S., 1972. Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. Psychol. Bull. 78, 266.

Amari, S., 1977. Dynamics of pattern formation in lateral-inhibition type neural fields. Biol. Cybern. 27, 77–87.

Coombes, S., 2007. Mathematical neuroscience. J. Math. Biol. 54, 305–307.

Pinotsis, D.A. & Miller, E.K. New approaches for studying cortical representations. in AAAI Spring Symposium-Technical Report 613–615 (AAAI, 2017).

Gray, C.M., König, P., Engel, A.K., Singer, W., 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338, 334–337.

Wang, J., Narain, D., Hosseini, E.A., Jazayeri, M., 2018. Flexible timing by temporal scaling of cortical responses. Nat. Neurosci. 21, 102.

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

Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological research: A tutorial. Adv. Methods Pract. Psychol. Sci. 1, 259–269.

Troebinger, L., López, J.D., Lutti, A., Bestmann, S., Barnes, G., 2014. Discrimination of cortical laminae using MEG. Neuroimage 102, 885–893.

Little, S., et al., 2018. Quantifying the performance of MEG source reconstruction using resting state data. Neuroimage 181, 453–460.

Bonaiuto, J.J., et al., 2018. Non-invasive laminar inference with MEG: Comparison of methods and source inversion algorithms. Neuroimage 167, 372–383.

Perkins, D.H., Perkins, D.H., 2000. Introduction to high energy physics. CAMBRIDGE university press.

Mazzoni, A., Logothetis, N.K. & Panzeri, S. Information content of local field potentials. Principles of neural coding 411–430 (2013).

Glomb, K., et al., 2021. Computational models in Electroencephalography. Brain Topogr. 1–20.

Kang, S., Hayashi, Y., Bruyns-Haylett, M., Delivopoulos, E., Zheng, Y., 2020. Model-predicted balance between neural excitation and inhibition was maintained despite of age-related decline in sensory evoked local field potential in rat barrel cortex. Front. Syst. Neurosci. 14, 24.

Trakoshis, S., et al., 2020. Intrinsic excitation-inhibition imbalance affects medial prefrontal cortex differently in autistic men versus women. Elife 9, e55684.

Haider, B., Duque, A., Hasenstaub, A.R., McCormick, D.A., 2006. Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. J. Neurosci. 26, 4535–4545.

Gao, R., Peterson, E.J., Voytek, B., 2017. Inferring synaptic excitation/inhibition balance from field potentials. Neuroimage 158, 70–78.

Pinotsis, D.A., Perry, G., Litvak, V., Singh, K.D., Friston, K.J., 2016. Intersubject variability and induced gamma in the visual cortex: DCM with empirical B ayes and neural fields. Hum. Brain Mapp. 37, 4597–4614.

Pinotsis, D.A., et al., 2017. Linking canonical microcircuits and neuronal activity: Dynamic causal modelling of laminar recordings. Neuroimage 146, 355–366.

Friston, K.J., Bastos, A.M., Pinotsis, D., Litvak, V., 2015. LFP and oscillations—what do they tell us? Curr. Opin. Neurobiol. 31, 1–6.

Legon, W., et al., 2016. Altered prefrontal excitation/inhibition balance and prefrontal output: markers of aging in human memory networks. Cereb. Cortex 26, 4315–4326.

Hamburg, S., Rosch, R., Startin, C.M., Friston, K.J., Strydom, A., 2019. Dynamic causal modeling of the relationship between cognition and theta-alpha oscillations in adults with down syndrome. Cereb. Cortex 29, 2279–2290.

Pinotsis, D.A., Miller, E.K., 2020. Differences in visually induced MEG oscillations reflect differences in deep cortical layer activity. Commun. Biol. 3, 1–12.

Kanashiro, T., Ocker, G.K., Cohen, M.R., Doiron, B., 2017. Attentional modulation of neuronal variability in circuit models of cortex. Elife 6, e23978.

M. Nienborg, H., Cohen, R., Cumming, B.G, 2012. Decision-related activity in sensory neurons: correlations among neurons and with behavior. Annu. Rev. Neurosci. 35, 463–483.

Maturana, M.I., et al., 2020. Critical slowing down as a biomarker for seizure susceptibility. Nat. Commun. 11, 1–12.

Bak, P., Tang, C., Wiesenfeld, K., 1988. Self-organized criticality. Phys. Rev. A 38, 364.

Freeman, W.J. A neurobiological theory of meaning in perception. in Neural Networks, 2003. Proceedings of the International Joint Conference on vol. 2 1373–1378 vol. 2 (2003).

Grindrod, P., Pinotsis, D.A., 2011. On the spectra of certain integro-differential-delay problems with applications in neurodynamics. Phys. D 240, 13–20.

Pinotsis, D.A., Friston, K.J., 2011. Neural fields, spectral responses and lateral connections. Neuroimage 55, 39–48.

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

Basar, E., Flohr, H., Haken, H. & Mandell, A.J. Synergetics of the Brain: Proceedings of the International Symposium on Synergetics at Schloß Elmau, Bavaria, May 2–7, 1983. vol. 23 (Springer Science & Business Media, 2012).

Fuchs, A., Jirsa, V.K., Kelso, J.A.S., 2000. Theory of the relation between human brain activity (MEG) and hand movements. Neuroimage 11, 359–369.

Haken, H., 2006. Synergetics of brain function. Int. J. Psychophysiol. 60, 110–124.

Jirsa, V.K., Kelso, J.A.S, 2000. Spatiotemporal pattern formation in neural systems with heterogeneous connection topologies. Phys. Rev. E 62, 8462–8465.

Haken, H. Complex Systems—Operational Approaches in Neurobiology, Physics, and Computers: Proceedings of the International Symposium on Synergetics at Schloß Elmau, Bavaria, May 6–11, 1985. vol. 31 (Springer Science & Business Media, 2012).

Ditzinger, T., Haken, H., 1989. Oscillations in the perception of ambiguous patterns a model based on synergetics. Biol. Cybern. 61, 279–287.

Gallego, J.A., Perich, M.G., Chowdhury, R.H., Solla, S.A., Miller, L.E., 2020. Long-term stability of cortical population dynamics underlying consistent behavior. Nat. Neurosci. 23, 260–270.

Yu, B.M., et al., 2008. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. Adv. Neural Inf. Process. Syst. 21, 1881–1888.