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

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

[1]Centre for Mathematical Neuroscience and Psychology and Department of Psychology, City —University of London, London EC1V 0HB, United Kingdom

[2]The Picower Institute for Learning & Memory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Correspondence: Dimitris A. Pinotsis

Centre for Mathematical Neuroscience and Psychology and Department of Psychology, City —University of London, London EC1V 0HB, United Kingdom
pinotsis@mit.edu

KEYWORDS: memory engrams; neural ensembles; working memory; predictive coding; auto-encoders; effective connectivity

Acknowledgements. This work is supported by UKRI ES/T01279X/1, ONR MURI N00014-16-1-2832, and the JPB Foundation.

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. The electric fields, in turn, can act as "guard rails" that funnel higher dimensional variable neural activity along stable lower dimensional routes. We show that electric fields carry information about working memory content. 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.

Introduction

In the era of large scale electrophysiology [1], neural recordings of high dimensionality are abundant. Yet this has revealed that brain areas seems to exchange information in low dimensions, using few task-related variables (latent variables) [2]. Indeed, brain dynamics evolve in low, not high, dimensional spaces [3,4]. These spaces are found by dimensionality reduction, [5,6]. Low dimensionality underlies a variety of cognitive and motor tasks [7,8].

A key point is that low-dimension latent variables track information and task demands and are stable, highly correlated across trials [9]. This stands in contrast to higher-dimensional neural dynamics; while there is some overlap[10], the specific neurons and synapses activated are variable across trials [11,12,13]. 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 [14–16]. This low dimensional stability is important for normal cognition and behavior. Downstream neurons and networks presumably need some consistency from upstream networks even though those upstream networks are under continuous reconfiguration.

The continuous reconfiguration is known as representational drift [17]. It occurs at a time scale of days, minutes or seconds [18]. It helps ensure the robustness of brain circuits. If some neurons fail, others can do the same task [19]. Plus, neurons, especially in higher cortical areas, have mixed selectivity which adds computational horsepower and cognitive flexibility [20,21].

Representational drift may also be important for the brain computations needed for Predictive Coding [22] and Reinforcement Learning [23]. 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 [24–26]. 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 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 [27]. Taken together, the above 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.

Further, 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 act as "guard rails" that funnel the higher dimensional variable neural activity along stable lower-dimensional routes. According to the theory of Synergetics [28], the electric field can be viewed as a control variable similar to energy [29] and attention signals [30] that evolves more slowly than the latent variables that represent information. This leads to 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. This then offers an explanation of 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. The same data were earlier used to build brain computer interfaces [31]. 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 [8,32]. 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 [33], 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. This extra step also allowed us to reconstruct the electric field produced by the ensemble. Having a detailed description of electric signals within the patch, we computed the electric field near it, using a classic dipole model from electromagnetism [34]. To sum up, we predicted the electric field generated each time (trial) the same location had to be remembered. Then, we tested if it was the same. We found that the electric field was different for different remembered locations and highly consistent across trials even though specific neurons activated were variable across trials (representational drift).

Results

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 [35,36]. 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 Figure 1A).

Examples of neural activity for two different individual trials corresponding to the same task condition are shown in Figure 1A. 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 $K(u_a, v_a, u_b, v_b, t, t')$.

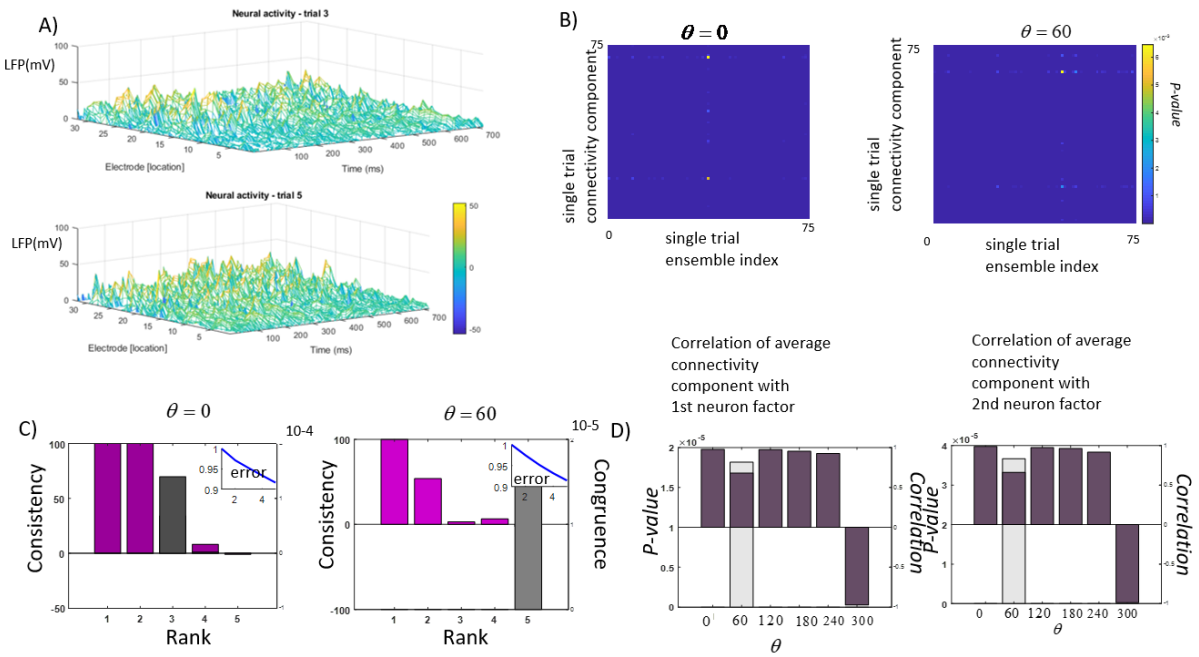


Figure 1. A. Examples of neural activity for two different individual trials corresponding to the same task condition. Local field potentials (LFPs, in *mV*) are shown on the vertical axis. The electrodes (location on the cortex) and time (in *ms*) are shown on the two horizontal axes. **B.** Significance (*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 location at θ=60 degrees are shown in the right panel. Estimates for all trials correlated perfectly ($p<10^{-2}$). **C.** Canonical Decomposition. Left and right panels show results for cued locations 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 locations 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).

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). Our starting point was different: we modelled each neural ensemble as a 2D neural network model of interacting excitatory and inhibitory populations (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 [37–39]. In turn, this model was transformed to a biophysical Gaussian Linear Model (GLM), that we called deep neural field. We trained the deep neural field to perform predictive coding (similar to [33]). 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, describing locations of excitatory and inhibitory populations into one. To sum up, the original 2D neural network model was first transformed to a 2D neural field and then to an 1D deep neural field considered in [33]. 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 neural ensembles, by comparing our results to those obtained using other methods.

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 [40] and a high dimensional SVD approach [41].

We first discuss connectivity components, denoted by A_k. 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 [42]. Specifically, we obtained single trial estimates in the following way. We trained a deep neural field model using a cost function considered in predictive coding and autoencoder networks. This training yielded single trial estimates of effective connectivity components A_k (Methods). Their

averages across trials were shown in Figure 2 of [33]. In that paper, we also showed that A_k 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.

Here, we validated the effective connectivity components obtained in [33] using two independent methods. First, using a correlation-based method, see [40]. 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 [40] did not yield effective connectivity per se, but one can map the ensemble index to effective connectivity components and kernels that we obtained. We will come back to this below.

We also used a second method. This is based on some old extension of high dimensional SVD, known as Canonical Decomposition (CD), see [41] and [42] for a recent application. This method does not provide single trial ensemble connectivity. It provides an estimate of average (across trials) connectivity that we had found in [33]. [42] 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 neuron mode with the average connectivity component. We will show that the two correlated perfectly. We will also show similar results using the first method above.

The first method described above is a correlation based method developed by [40]. This allowed us to obtain neural ensembles in an unsupervised way, using a *k*-means algorithm. We adapted the original algorithm 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 [35,36], this process also assigns populations to ensembles. We then computed the correlation between the ensemble indices and the first connectivity components for different stimuli (angles). We asked whether the ensemble index correlated with the component strength for each electrode and trial. In [33], 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 using different trials as samples. Recall that the monkey performed a spatial delayed saccade task (Supplementary Figure 1B). The *p*-values across all 32 electrodes are shown in Figure 1B 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 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[43]. It also provides an independent validation of our effective connectivity estimates at the single trial level.

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. A second, independent validation of our approach was achieved by comparing the connectivity components to similar estimates obtained using a high dimensional SVD approach. This is the second method mentioned above, known as Canonical Decomposition (CD; see Methods). CD yields an approximation of the three-dimensional LFP array $Y \in \mathbb{R}^{N_T \times N_S \times T}$ (trials x electrodes x time) $\tilde{Y}_{ijk} \approx f_{i1} b_{j1} c_{k1} + f_{i2} b_{j2} c_{k2} + \ldots + f_{iR} b_{jR} c_{kR}$ in terms of a sum of combinations of elements $f_{i1}, b_{j1}, c_{k1}, f_{i2}, b_{j2}, c_{k2}, \ldots$. Taking together 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* in the mathematical literature [44]. 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 ($j=1,\ldots,N_S$) or time points ($k=1,\ldots,T$) for f_{im}, b_{jm} and c_{km} respectively. 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, $m=1,...,R$. Thus, R is equal to the number of factors in the CD approximation $\tilde{Y}_{ijk}$. Below, we will see that R can be estimated based on some measures from statistics.

We used the CD approximation to validate our effective connectivity estimates. Of particular interest for our current analysis is the $B=[b_{jm}]$ mode. This is an $N_S \times R$ matrix where fixing $m=M*$, we obtain a vector $\vec{b}_{jM*}$ of dimension N_S that approximates the LFPs measured at each electrode (averaged across time and trials). [42] used spiking data and suggested that one can think of $\vec{b}_{jM*}$ as a prototypical firing rate across neurons. They called this "neuron factor". Below, we call $B=[b_{jm}]$, the "neuron mode" because the word "factor" is commonly used in the mathematical literature to denote terms in the CD approximation [44]. In other words, our "neuron mode" is the "neuron factor" of [42]. According to these authors, a neuron mode corresponds to "*the synaptic weights from each latent input to each neuron"*. This is the *same definition* as that of the component strengths included in (Pinotsis et al., 2017): *"(component strengths) express the sum of all connectivity weights that target the neurons that contribute to the LFPs observed from each electrode"*. Thus, connectivity components and neuron modes are generalisations of principal components in three dimensions, and they have similar definitions.

We therefore asked whether connectivity components and neuron models were correlated. We considered connectivity components averaged across trials. First, we found the neuron mode $B=[b_{jm}]$, where $j=1,\ldots,N_S$ and $m=1,...,R$ by decomposing our LFP data using CD. To find the CD approximation $\tilde{Y}_{ijk} \approx \sum_{m=1}^{R} f_{im} b_{jm} c_{km}$, we used a standard iterative Alternating Least Squares (ALS) algorithm [45]. We assumed different values for R=1,…,5. For each value of R, we calculated the sum of squares reconstruction error $\left\| Y - \tilde{Y} \right\|_{Frob}$. This is plotted on the vertical axis appearing in the top right insets of the panels in Figure 1C. 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 Figure 2), the error reduced with an increasing number of factors (blue line in the insets). 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 an alternative way to obtain the rank in CD approximations [46]. It uses CD factors to compute an alternative approximation of the data matrix, known as Tucker3 approximation [47]. The Tucker3 approximation also contains (mixtures of) CD factors. Then 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 [46], this corresponds to consistency values between 50-100%. 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).

The optimal rank *R* should result in high values for both measures. These are shown in the right and left vertical axes of the bar plots in the main panels of Figure 1C. 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 Figure 1C) we obtained high consistency values for *R=1,2* (magenta bars). The same was true also for other angles (Supplementary Figure 2). For all values of *R* in Figure 1C and Supplementary Figure 2, congruence was very small (grey bars). Its order was 10^{-4} for $\theta = 0$ and 10^{-5} for $\theta = 60$ degrees (Figure 1C). 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 $B=[b_{jm}]$ of dimensionality $N_S \times 2$. Fixing $m=M*$, where $M*=1,2$ we obtained two vectors $\vec{b}_{jM*}$ that approximate average LFPs across time and trials. These are the two *columns* of the neuron mode $B=[b_{jm}]$. We call these vectors the 1st and 2nd neuron factors. The two neuron factors are the columns of the neuron mode.

Recall that, each connectivity component is also a vector of length N_S. In [33], 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. 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 Figure 1D. We found that the 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). Correlation coefficients were $r>0.7$ for all angles.

To sum up, we compared our approach for performing dimensionality reduction to a high dimensional SVD approach, known as Canonical Decomposition (CD[41,42]). We found that the effective connectivity component obtained using our approach correlates significantly with the neuron factors obtained using CD). This means that both methods find a very similar structure of the latent space within neural activity evolves. This provides a second, independent validation of our approach.

The approach introduced in [33] has two steps. Step 1 is dimensionality reduction that we considered above. This yields the connectivity components. These are the variables that describe the structure of the latent space. Step 2 is considered below. It involves mapping the latent space to a cortical patch. This yields the connectivity kernel. This extra step distinguishes our approach from other dimensionality reduction approaches, e.g. [5,7]. The mapping to a

cortical patch is constructed by obtaining the connectivity kernel $K(u_a,u_b,t,t')$ from the connectivity components A_k . Recall that the entries of the connectivity components, called *component strengths,* are similar to principal components in PCA. They are *sums* of all *connectivity weights* that target the neurons that contribute to the LFPs observed from each electrode, see Methods and [33] for more details. The corresponding *connectivity weights* are the entries of the connectivity kernel $K(u_a,u_b,t,t')$ of a deep neural field model. We will see below that this kernel can also be thought of as the probability of connections between any two neuronal populations that are part of the neural ensemble.

The connectivity weights quantify the strength of the effective connections between the recording sites within each cortical area, see Supplementary Figure 1A. 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. Note that although the connectivity components A_k we used here correspond to an 1D deep neural field model, they are linearly related to the connectivity components $z_{kl}^{PP'}$ of a 2D neural network model with distinct excitatory and inhibitory populations (Equation 7 in Methods). This means if we had separate multiunit activity from excitatory and inhibitory neurons, we could have used the same approach used here to obtain separately the effective connectivity of excitatory and inhibitory populations $z_{kl}^{PP'}$. Because we here used aggregate neural activity from both kinds of populations (LFPs), we obtained effective connectivity components A_k for both populations together. This is similar to other dimensionality reduction approaches, e.g. [8,42].

In brief, after obtaining the effective connectivity components, we used them to obtain the connection weights (the connectivity kernel, see Methods). This is important because it enabled us to map the latent space to a cortical patch, that the neural ensemble occupies. Later, we will also use this mapping and the connectivity kernel to predict the electric field generated by the ensemble. First, we assessed whether the connectivity kernels could also identify neural ensembles, similarly to connectivity components above.

We assumed a Gaussian connectivity profile over space and obtained single trial estimates for the connectivity kernel

$K(u_a,u_b)=g_C(u_a,u_b+\bar{u})$, where $g_C(u_a,u_b+\bar{u})=\left(C\sqrt{2\pi}\right)^{-1}\exp\left\{-(u_a-u_b-\bar{u})^2/2C^2\right\}$. This is not the only expression for the kernel we could have considered. In fact, there are several ways to obtain $K(u_a,u_b)$. To understand what they are, we revisit its definition. Mathematically, the connection weights (kernel) can be thought of as the probability of having connections between neural populations forming a neural ensemble. In [33], this probability (connection weights) were obtained from the latent variables.

In brief, the novelty of our approach in comparison to other dimensionality reduction approaches is that we map the latent space (latent variables) to a cortical patch (connection weights). This is based on considering the connection kernel as a probability function. There are other methods to obtain the probability function, including splines [48], and tools from complex systems [49]. We will systematically consider these methods elsewhere.

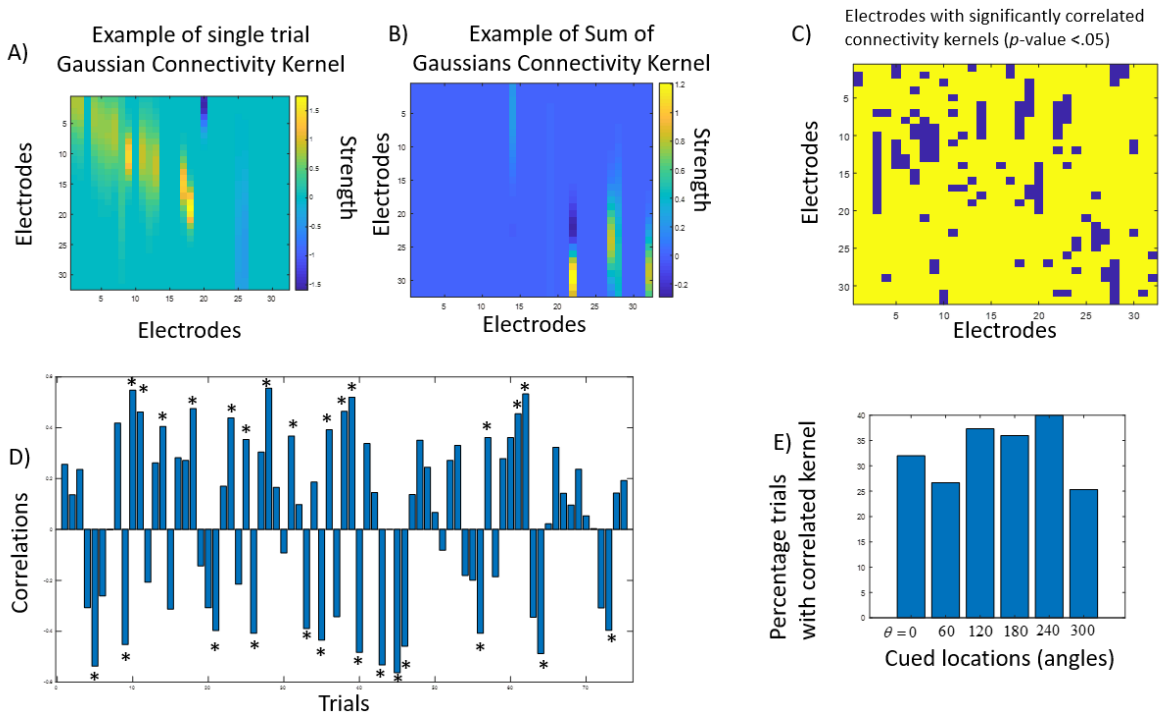


Figure 2. A. Example of effective connectivity kernel with a Gaussian profile. This describes the weights which scale neural activity propagating between any pair of populations located near one of the electrodes. This kernel characterizes information flow at the single trial level. **B.** Example of an alternative expression of the effective connectivity kernel obtained as a sum of Gaussian functions using a series expansion. **C.** Correlations between

the connectivity kernels in panels A. and B. 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. **D.** Correlations between single trial ensemble indices and connectivity kernels for cued location 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. **E.** Percentage of significantly correlated trials for each cued location. Locations are shown on the horizontal axis. Overall, 25-40% of single trial kernel estimates correlated with ensemble indices for different angles.

Here, as an alternative to the Gaussian profile, we also considered a series expansion [50]. This provided an expression for $K(u_a, u_b)$ that involved sums of Gaussian functions $g_C(u_a, u_b)$ weighted by known Hermite polynomials $H_n(u)$ [51]. Using the first three Hermite polynomials, we show in Methods that the connectivity kernel is given by

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

$$g_C(u,u') = \exp\left\{-(u-u'-\bar{u})^2/2C^2\right\}$$
$$\bar{u} = A_1 / A_0$$
$$C = \frac{\sqrt{A_0 A_2 - A_1^2}}{A_0}$$

In the above expression, the connectivity kernel is given in terms of the connectivity components A_k (see Methods for more details). Example connectivity kernels obtained using data from a random trial are shown in Figure 2A and 2B. Figure 2A shows a single Gaussian kernel $g_C(u_a, u_b)$, while Figure 2B shows sums of Gaussian functions given by the expression above. Note that because of the Gaussian profile, only elements around the main diagonal are non-zero. Figure 2C shows correlations between the two expressions obtained. R=87% of connectivity weights of the kernels shown in Figures 2A and 2B were correlated.

For simplicity, in the analyses below we used the expression involving a single Gaussian kernel

$g_C(u_a, u_b)$. Similar analyses can be carried out using alternative expressions in a similar way. 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. These correlations were obtained between estimates obtained for each trial using our and the method of (Humphries,2011). They 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. The difference to correlations with connectivity components above was that the correlations with the connectivity components were significant only for some of the trials where the same cued location was maintained, not all. Note that the connectivity kernels were a priori constrained to have a Gaussian (parametric) form, $g_C(u_a, u_b)$, 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 (Humphries,2011) that requires this assumption—which, on the other hand, was intrinsic to the parametric form for the connectivity kernel that we used. If ensemble indices are not Gaussian, there are no significant correlations. Correlations between ensemble indices and connectivity kernels for cued location at $\theta = 240$ degrees are shown in Figure 2D . 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 location is shown in Figure 2E.

To sum so far, in the second step of our approach described above, we mapped the latent space to a cortical patch occupied by a neural ensemble. We also described the exchange of information during cue maintenance using a connectivity kernel. We found that the corresponding weights correlated significantly with single trial ensemble indices obtained using the method of (Humphries,2011) across a large percentage of trials. These effective connectivity weights identify the same neural ensembles obtained by (Humphries,2011).

Having obtained the connectivity of neural ensembles, we could then predict the Electric Field (EF) generated by them. The connectivity kernels describe how neurons communicate via electric signals sent from one part of the patch occupied by the neural ensemble 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 location was maintained. We used the bidomain model that has been used to predict the electric field generated by biological tissues, like the cardiac muscle [52,53]. To estimate the extracellular EF, the model requires only a measurement of the transmembrane potential V^m. 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 (Methods).

Here, we used the bidomain model to obtain three EF estimates. This model requires transmembrane potential V^m measurements. Thus, we used three proxies for V^m : First, simulated potentials after training the deep neural field model with all available data, which we called simulated EFs. Second, simulated transmembrane potentials after training the deep neural field model with data from those trials, where single trial connectivity kernels correlated with single trial ensemble indices ("correlated trials"; cf. Figure 2D and 2E and relevant discussion above). We called the corresponding EF estimates, simulated EFs across correlated trials. Third, real LFPs that were obtained from recordings and did not use the deep neural field model. These were used as proxies for transmembrane potentials and replaced the simulated transmembrane potential from the neural field model above. We called the corresponding EFs obtained used real LFPs and the bidomain model, real EFs.

An example simulated EF estimate is shown in Figure 3A. 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*). *P*-values of correlations between EF amplitudes are shown in Figure 3B. These correspond to EFs generated by neural ensembles maintaining a cued location at $\theta = 0$ degrees. We have here considered EF estimates from trials where our connectivity kernel correlated with the findings of (Humphries, 2011; correlated trials). 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%. Similarly, we found that EF

amplitudes were also correlated across all trials and other angles with R= 70-80% (see Supplementary Figure 3).

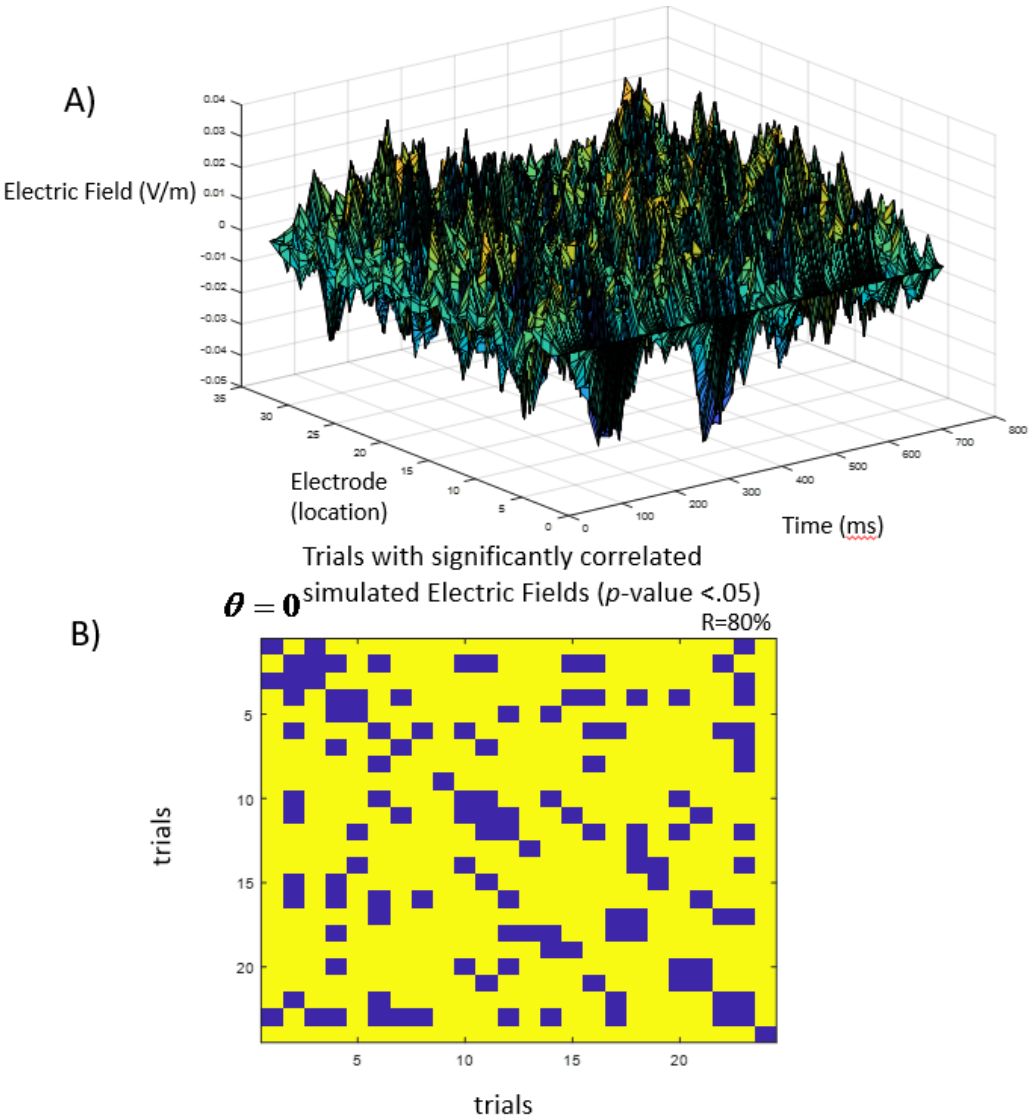


Figure 3. A. Example of simulated electric field (EF) using the bidomain model. The EF amplitude is shown on the vertical axis (*V/m*), while the two horizontal axes show the electrode (location on the cortex) and time (*ms*). **B.** *P*-values of correlations between single trial EF amplitudes. These correspond to EFs generated by neural ensembles maintaining a cued location 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%.

Finally, we asked whether EF estimates were consistently different for neural ensembles that maintain different cued locations. We wanted to test if we could distinguish between memorized cues based on simulated and real EFs. To formally answer this, we used the EF estimates as classification features of different trials by cued location. We analysed simulated EFs across all and only across correlated trials as well as real EFs. We used two different

methods for classification, Naïve Bayes and diagonal LDA. These are among the most commonly used classification algorithms.

The results of our analyses are shown in Figure 4. Decoding accuracy is shown on the vertical axis, while the corresponding electrodes are shown on the horizontal axis. Red horizontal lines depict chance level (0.1667). In Figure 4A, we show results obtained using diagonal LDA (yellow) and Naïve Bayes (orange). Simulated EF estimates across all electrodes show high accuracy for both methods. Decoding accuracy was up to five times higher than chance. Simulated EF estimates across correlated trials showed accuracy higher than chance for half of electrodes using Naïve Bayes (Figure 4B; results using diagonal LDA overlap closely, not shown). Weaker accuracy is due to the smaller number of trials used for training the classifiers (25-40% of all trials, see Figure 2C, above). Real EF estimates also showed high decoding accuracy across electrodes (Figure 4C). Results obtained using diagonal LDA are shown in orange and Naïve Bayes in blue.

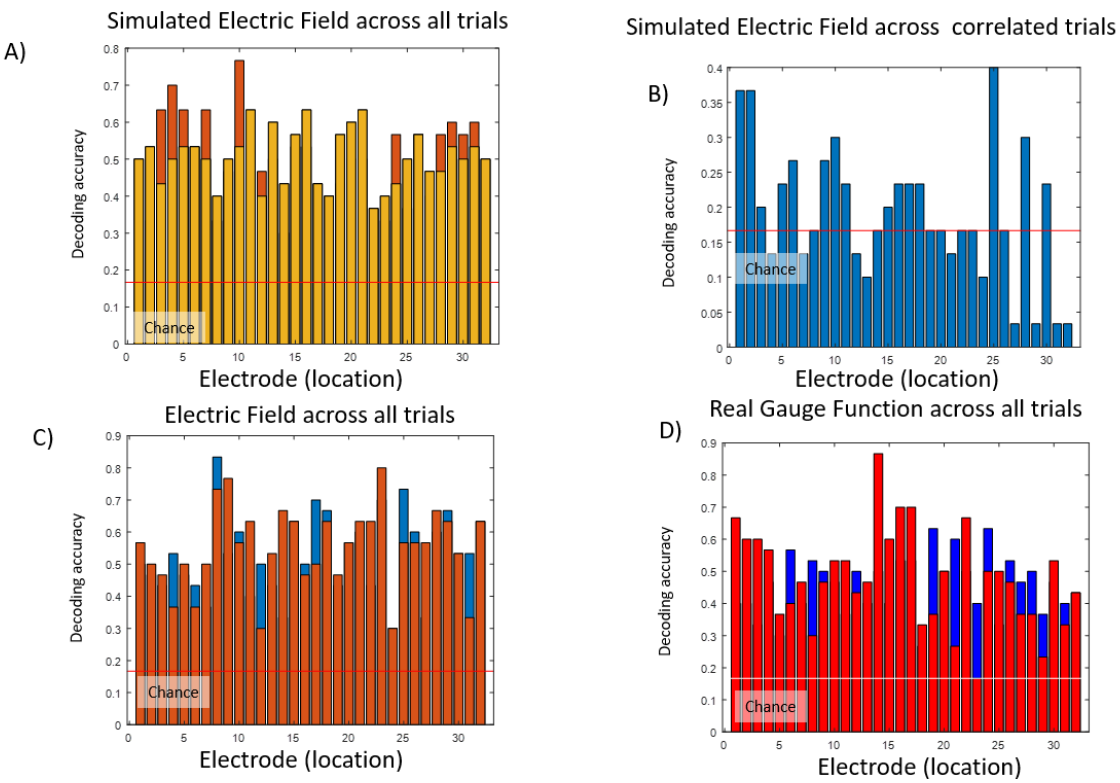


Figure 4. A. Decoding accuracy obtained using diagonal LDA (yellow bars) and Naïve Bayes (orange bars). Accuracy values are shown on the vertical axes, while the corresponding electrodes are shown on the horizontal axes. Red horizontal lines in this and subsequent panels depict chance level (0.1667). Classification features included EF estimates using the bidomain model and simulated transmembrane potentials from the deep neural field model. **B.** Bars depict decoding accuracy estimates obtained using Naïve Bayes. Estimates using diagonal

LDA overlap closely and are not shown. In this panel, classification features included EFs obtained only from the trials where the connectivity kernel correlated with the ensemble indices of (Humphries, 2011). **C.** Decoding accuracy based on classification features that included EFs obtained from the bidomain model and real LFPs – used as proxies of transmembrane potentials. Accuracy estimates based on diagonal LDA are depicted by orange bars and estimates obtained via Naïve Bayes are depicted by blue bars. **D.** Decoding accuracy based on classification features that included gauge functions as opposed to EFs. These do not require the bidomain or neural field models. Accuracy estimates using diagonal LDA are depicted by red bars and estimates using Naïve Bayes are depicted by blue bars.

The above results suggest that EFs were stable across trials where the same cued location was maintained. EFs also differed systematically depending on the exact location and seemed to reflect the remembered stimulus. Although the exact neurons forming an ensemble differed from trial to trial, the corresponding real and simulated EFs were the same. They contained unique information about the remembered stimulus, that seems to be preserved across trials.

To confirm that the stability of the EFs found above is not due to the EF model used (bidomain model), we repeated the above analysis using some different classification features that are independent of this model. These features where the values of some functions known from electromagnetism as *gauge functions* (Methods). They are obtained by subtracting real LFPs recorded in different trials where the same cued location was maintained. Similarly to our analyses above, we used LFPs as proxies for extracellular potentials (EPs). Following the theory of electromagnetism EF is stable if and only if the gauge function is stable (and the other way around). The results of our analyses are shown in Figure 4D. The bars show decoding accuracy obtained via diagonal LDA (red bars) and Naïve Bayes (blue bars). Accuracy obtained using gauge functions as features is high, similar to the results in Figure 4C. Thus, gauge functions, and therefore the EFs, remained stable. This conclusion follows from using the LFP data directly—that is, without using the EF and deep neural field models.

Methods

Experimental Data and Recording Setup.

Two adult male monkeys (monkey C, Macaca fascicularis, 9kg; monkey J, Macaca mulatta, 11kg) were handled in accordance with National Institutes of Health guidelines and the

Massachusetts Institute of Technology Committee on Animal Care. They were trained to perform an oculomotor spatial delayed response task (Supplementary Figure 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 location, the target was presented with a green highlight and a water reward was delivered otherwise the target was presented with a red highlight and reward was withheld. Three 32-electrode chronic arrays were implanted unilaterally in PFC, SEF and FEF in each monkey (Supplementary Figure 1C). Each array consisted of a 2 x 2 mm square grid, where the spacing between electrodes was 400 um. The implant channels were determined prior to surgery using structural magnetic resonance imaging and anatomical atlases. From each electrode, we acquired local field potentials (extracted with a fourth order Butterworth low-pass filter with a cut-off frequency of 500Hz, and recorded at 1 kHz) using a multichannel data acquisition system (Cerebus, Blackrock Microsystems). We analyzed local field potentials (LFPs) during the delay period when monkeys held the cued locations in memory.

Neural network model of neural ensembles.

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. [39,54,55]. Let $x_E^a(u_a, v_a, t)$ and $x_I^b(u_b, v_b, t)$ 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 $x_E^a(u_a, v_a, t)$ and $x_I^b(u_b, v_b, t)$ is given by the following neural network equations

$$\dot{x}_E^a(u_a, v_a, t) = -\tau_E x_E^a(u_a, v_a, t) + \sum_c K_{PP'}(u_a, v_a, u_c, v_c) f[x_P^c(u_c, v_c, t)] + S \circ U_E, \quad P = E; P' = \{E, I\}$$

and (1)

$$\dot{x}_I^b(u_b, v_b, t) = -\tau_I x_I^b(u_b, v_b, t) + \sum_c K_{PP'}(u_b, v_b, u_c, v_c) f[x_P^c(u_c, v_c, t)] + S \circ U_I, \quad P = I; P' = \{E, I\}$$

where $S:\mathbb{R}^n \to \mathbb{R}^n$ maps exogenous inputs to depolarization and 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 $\upsilon \in W$.

We can then take the continuum limit of Equations (1). This 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_i^a with dimensions $(\Delta\upsilon, \Delta v)$ $i \in 1,\ldots,N$ and $j \in 1,\ldots,L$. Thus the subgroup of neurons in the square $T_{ij}^a = \{[i\Delta u,(i+1)\Delta u),[j\Delta v,(j+1)\Delta v)\}$ of M_A is given by $\rho_{ij}^a = \zeta_{ij}^a \Delta\upsilon\Delta v$. For mathematical convenience, consider a copy M_B of manifold M_A. The interaction between neurons in cortical patches $T_{ij}^a \in M_A$ and $T_{kl}^b \in M_B$ only depends on the duplets (i,j) and (k,l). A neuron at location (u_a, v_a) inside square T_{ij}^a receives input from all neurons in square T_{kl}^b with strength $K_{PP'}(i,j,k,l) = \tilde{K}_{PP'}(i\Delta u, j\Delta v, k\Delta u', l\Delta v')$, 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 $\Delta\upsilon$ 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 $X_P(i\Delta u, j\Delta v, t) = (\rho_{ij}^a)^{-1} \sum_{(i,j)\in T_{ij}^a} x_P(i,j,t)$ and consider the continuum limit $\Delta u, \Delta v, \Delta u', \Delta v' \to 0$: all nodes within the patches T_{ij}^a and T_{ij}^a 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

$$\dot{X}_E(u,v,t) = -\tau_E X_E(u,v,t) + \iint_{M_B} K_{PP'}(u,v,u',v') f[X\ (u',v',t)]du'dv' + S \circ U_E$$

and (2)

$$\dot{X}_I(u,v,t) = -\tau_I X_I(u,v,t) + \iint_{M_B} K_{PP'}(u,v,u',v') f[X\ \ (u',v',t)]du'dv' + S \circ U_I$$

Similarly to (Pinotsis et al., 2017), we consider perturbations $\hat{X}_P$ of membrane potentials around baseline $X_P(u,v,t) = X_{0P}(u,v) + \hat{X}_P(u,v,t)$, with respect to the variables $u,v \in M_A$ that denote locations on the cortical manifold M_A . Thus, we obtain

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

where $\varepsilon \sim (0, s_s^{\,2} I)$ and

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

In [33], we called the functions G_k *principal axes* and the latent variables $z_{kl}^{PP'}$ *connectivity components*—to resemble standard PCA terminology. In that earlier work, we found that the principal axes contained temporal information, while the connectivity components contained spatial information. Similarly to that earlier work, connectivity components are defined by the following equations

$$\begin{aligned}
&z_0^{PP'}(u,v) = df_0 \tau_P^{\,-1} \iint K_{PP'}(u,v,u',v')du'dv' \\
&z_{kl}^{PP'}(u,v) = \tfrac{df_0}{k!}\, \tau_P^{\,-1} \iint K_{PP'}(u,v,u',v')(u-u')^k (v-v')^l du'dv'
\end{aligned} \tag{5}$$

If we had separate data (depolarization or spike rates) for the excitatory and inhibitory populations, we could use Equations (2) 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 [56,57]. 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:

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

(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 Equations (3), 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 excitatory and $\tilde{v}$ for the location of the inhibitory populations into a single variable. Then, adding Equations (2a) and (2b), we obtain

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

where the aggregate connectivity components $\tilde{A}^P$ are two factor sums of $z^{PP'}$ defined by Equations (5) :

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

Letting $H = \left[G_{10}^E, G_{01}^I, G_{20}^E, G_{02}^I, \ldots, G_{i\cdots 0}^E, G_{0\cdots i}^I \right]^T$, Equation (6) is a *deep neural field*, and can be rewritten in the general form of a Gaussian Linear Model (GLM; cf. Equation (1) in [33]),

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

where for simplicity of notation we have relabelled, $A_k = [\tilde{A}_0^P, \tilde{A}_{01}^P, \tilde{A}_{10}^P, \ldots, \tilde{A}_{10}^P, \tilde{A}_{0i}^P]$ and have

dropped the superscript P, because we do not distinguish between neural populations in what follows. This simply relabels components with two subindices as components with a single subindex. Since there is only one spatial variable, only one sub-index was needed. We have also assumed that cortical activity $\widehat{X}(\tilde{u},\tilde{v},t) \in \tilde{X}$ was sampled from a random process $\tilde{X}$ and $Y = \hat{X} - m$.

The above reduction was obtained under some mathematical assumptions. To validate them, we compared our effective connectivity estimates against two established approaches (see Methods subsection 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.

To sum up, starting from a neural network model for coupled excitatory and inhibitory populations (Equation 1), we have shown how it can be reformulated as a deep neural field model (Equation 6) – and then a GLM (Equation 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.

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,

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

using a Restricted Maximum-Likelihood (ReML) algorithm. 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 [33] for more details. Note that the cost function defined by Equation (9), is the same cost function like the one used in Predictive Coding.

To sum up, Equations (4) define the connectivity components $z^{PP'}$ of the neural network (2). Similarly, Equations (7) define the connectivity components A_k of the deep neural field (Equation (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 Figure 2 of [33]. 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 and 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 (4) and (7), 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 but continues and thus depends on continuous variables (u, v).

In [33], after obtaining A_k, we assumed that cortical connectivity has a Gaussian profile and computed $K(u_a, u_b) = \left(C\sqrt{2\pi}\right)^{-1} \exp\left\{-(u_a - u_b - \bar{u})^2 / 2C^2\right\}$ and 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 focused on the corresponding single trial estimates of $K(u_a, u_b)$ and also compared them to estimates obtained without assuming a Gaussian profile.

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 [40] . 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 [58]. 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. The method provides a spectral decomposition of the modularity matrix using a stochastic algorithm [58]. It employs a consensus algorithm to ensure that the same clustering is obtained for different initialisations [59]. This method has been applied to neural activity in visual cat areas [40] and the Aplysia pedal ganglion [60].

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

$$\tilde{Y}_{ijk} \approx \sum_{m=1}^{R} f_{im} b_{jm} c_{km}, \tag{10}$$

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 the three modes with dimensions trials (N_T) x electrodes (N_S) x time (T) and R is known as the rank of Y . 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 $\otimes$ 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 [45]. CD has recently been applied to analyse spiking data and identify neural ensembles in [42]. Here, we adapted this work to identify neural ensembles using LFPs. The CD approach by [42] 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 is a mathematical construct that uses model comparison where different models comprise mixtures of factors appearing in a CD approximation, similar to a so-called *Tucker3* approximation [47] . It quantifies the degree that the LFP data contain a trilinear variation, see [46] for more details. Consistency 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 [61]. To ensure that results reflect a trilinear variation in the data and are stable, below we chose a rank R with high consistency and low congruence.

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 [34]. This assumes that the neural

tissue can be represented by a cylindrical fiber of radius d . This means that the problem has rotational symmetry and 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 [62] and Supplementary Figure 4. 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 . 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 [27]

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

Here r is the brain resistivity with $r = 2.2$ Ohm [63] 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 σ^e is the conductivity of the extracellular space, and R is the distance between the

current source at the point (d,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, $R=\sqrt{(u^e-u^a)^2+(\rho-d)^2}$, see Supplementary Figure 4. Then, according to the bidomain model, Equation (14) can be written as [52,53]

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

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,

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

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 [51],

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

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

Gauge transformations of electric potentials.

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 [27]. 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. Below, 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 used to obtain estimated of the EP at an arbitrary trial j, $V^e\{\text{trial } j\}$, using Equations (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 two ways: First, we asked if EF estimates were consistently different for neural ensembles that maintain different cued locations. 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[33]. Second, we used gauge functions that connect the recorded LFPs (considered as proxies for extracellular EPs). If the EF was stable, the EPs are related by called *gauge transformations* [27]

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

where the function $\chi = \chi(\rho, \zeta, t)$ is called *gauge function.* In brief, a second way to test the stability of EFs is to test if the gauge functions can be used to distinguish between different cued locations (see Results). According to Equation (18), the time derivative of the gauge function $\partial\chi / \partial t$ is equal to the difference of EPs and LFPs corresponding to any two trials. Equation (18) should hold for any arbitrary pair of trials. Thus, we asked whether we could decode cued locations using gauge function derivatives $\partial\chi / \partial t$ as classification features. An independent experimental validation could also be carried out using intracellular recordings: If Equation (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 $V^i\{\text{ trial } j\}$.

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 [33]. 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 Equation (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 sub-indices denoting location have been replaced by continuous spatial variables that lie on a patch $u,v \in M_A$. Then the problem is the following. Given the connectivity components A_0, A_{kl} (Equation 5) find $K_{PP'}$. In mathematical terms, the kernels are probability distribution functions and can be estimated using a variety of methods, including splines [48], series expansions [50] and other methods [49,64].

We here considered a Gaussian connectivity profile used in [33] and an alternative expression for the connectivity kernel that includes sums of Gaussian profiles weighted by polynomial factors known as Hermite polynomials, H_n [51]. These sums are known as Gram-Charlier series. In brief, the connectivity kernel of the neural ensemble can be approximated by

$$
\begin{aligned}
&K(u,u') = \sum_{0}^{k} d_n H_n(u) g_C(u,u') \\
&g_C(u,u') = \exp\left\{-(u-u'-\bar{u})^2 / 2C^2\right\} \\
&\bar{u} = A_1 / A_0 \\
&C = \frac{\sqrt{A_0 A_2 - A_1^2}}{A_0}
\end{aligned}
\tag{19}
$$

where the Hermite polynomials, H_n 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, Equation (20) using the binomial theorem and the definition of connectivity components gives

$$
\begin{aligned}
&A_k(u,u') = \gamma \int \sum_{1}^{k} d_n H_n(u) g_C(u,u') \sum_{0}^{k} (-1)^n \binom{k}{n} u^k u'^{k-n} du' \\
&\gamma = \tfrac{df_0}{k!} \tau_P^{-1}
\end{aligned}
\tag{21}
$$

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

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

Substituting the above expressions and the expressions for Hermite polynomials into Equation (19), we obtain an alternative expression for the connectivity kernel $K(u,u')$ that involves sums of Gaussian functions (keeping the first three terms):

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

Discussion

We found that in a spatial working memory task, different remembered locations resulted in different electric fields that were highly stable across trials yet, at the level of specific circuits there was variability (representational drift[17,18]). This follows directly from theory of electromagnetism. The same electric field can arise from different combinations of specific neurons and networks (electromagnetic sources and sinks[65]). This is known as non-uniqueness of the electromagnetic inverse problem: One cannot find the exact sources by measuring electric fields alone [27].

Across like trials, the same memory was maintained even though 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 of the observed variability in the patterns of neurons forming a neural ensemble. Here, we tested this prediction using LFP data and computational modeling. In future work, we will experimentally test the stability of the electric field.

Does low-dimensional stability in electric fields help the brain perform computations? We argue: Yes, by allowing latent states to be reliably transferred between brain areas, in accord

with modern engram theory [66]. To see how, consider the theory of Synergetics [28,67–69]. In Synergetics, latent states are called order parameters. The theory posits that order parameters exist in all self-organized dynamical systems (e.g. molecules, fluids) and therefore the brain. They emerge as a result of self-organization and capture collective dynamics of a large part of the system's individual parts. Importantly, order parameters evolve slowly, that is, at a different timescale than the activity of parts, e.g. oscillations/spiking [28]. In brief, Synergetics suggest that neural ensembles are determined by parameters that evolve slowly – at a timescale determined by behavior. Thus, in modern language, order parameters are similar to latent variables [70,71] and the effective connectivity estimates that we found here.

In Synergetics, order parameters evolve at a slower timescale than latent variables and neuronal spiking. This separation of timescales follows from the center manifold theorem. Haken pointed out this separation is crucial for consciousness. Order parameters evolve slowly and this "can be interpreted as a phase transition from subliminal to conscious phase". They sent essential information to other brain areas. This results in conscious experience. This hypothesis resonates with modern engram theories which say that memories are the result of coordinated activity within a neural ensemble and across ensembles[72]. For example, the frequency of oscillations in a certain brain area (e.g. PFC, FEF and LIP) is an order parameter that controls the spiking of a large number of neurons. In [73], we showed that during a working memory task, when the cognitive capacity limit was exceeded, synchrony between oscillatory responses in PFC, FEF and LIP broke down and the monkey made errors. That is, order parameters were interacting differently with each other when the monkey could vs. when he could not remember.

But why do order parameters appear and how do they evolve? Synergetics suggests that the goal of order parameters is to control individual parts. Here, individual parts are neurons. Fast neuronal firing is determined by the order parameters because individual neurons are enslaved. Neurons give rise to the ensemble and this, in turn, determines the function of each neuron (e.g. frequency of oscillations/spiking). This is known as slaving principle. Individual neurons are enslaved to the neural ensemble whose goal is to maintain memories. This is an example of circular causality. Order parameters emerge at a slower timescale from the collective activity of enslaved parts that evolve fast and control them. At the same time, order parameters also interact with each other, and then change, which also changes the behavior of enslaved parts. According to Synergetics, enslavement is not rigid. The same order parameter may give rise to

different firing patterns. This explains the representational drift as a result of the circular causality between order parameter—enslaved parts—order parameter.

In the above discussion, we saw that Synergetics explains neural recordings and behavior in terms of latent variables, similarly to recent work based on dimensionality reduction approaches. Interestingly though, Synergetics goes one step further. Besides latent variables and enslaved parts, the theory predicts that there is another set of parameters that characterize brain dynamics, called control parameters. These evolve at an even slower timescale than latent variables and spiking and are characteristic for each state of the brain. They are distinct from order parameters. Synergetics suggests that control parameters are somehow fixed in the sense that when they change, the brain goes to a different stable state, similar to phase transitions in thermodynamics [74]. Order parameters can become control parameters – but crucially, there is a temporal hierarchy comprising the timescales of control parameters, order parameters and enslaved parts.

We therefore suggest that the electric field is a control parameter. Similar parameters in Synergetics are energy [28,29], or feedback attention signals in a binocular rivalry task [30]. Importantly, considering the electric field as a control parameter implies two conclusions: first that it is stable for each brain state (memory). Second, there are three, not two, timescales that characterize the evolution of neural activity: spiking, latent variables like oscillations, resting state networks etc. and the timescale of electric field changes.

The stability of the electric field allows the brain to control the latent variables (e.g., oscillations) that give rise to the same memory. It does not represent information, but governs the latent variables that do so. In other words, the electric field is a physical quantity that 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.

References

1. Steinmetz, N. A., Koch, C., Harris, K. D. & Carandini, M. Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Current opinion in neurobiology* **50**, 92–100 (2018).
2. Katlowitz, K. A., Picardo, M. A. & Long, M. A. Stable sequential activity underlying the maintenance of a precisely executed skilled behavior. *Neuron* **98**, 1133-1140. e3 (2018).
3. Gallego, J. A. *et al.* Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature communications* **9**, 1–13 (2018).
4. Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609-623. e29 (2018).
5. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *arXiv preprint arXiv:2107.04084* (2021).
6. 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).
7. Cunningham, J. P. & Byron, M. Y. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience* **17**, 1500–1509 (2014).
8. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Current Biology* **26**, R656–R660 (2016).
9. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods* **15**, 805–815 (2018).
10. Churchland, M. M. *et al.* Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience* **13**, 369–378 (2010).
11. Mongillo, G., Rumpel, S. & Loewenstein, Y. Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory. *Current opinion in neurobiology* **46**, 7–13 (2017).

12. Attardo, A., Fitzgerald, J. E. & Schnitzer, M. J. Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature* **523**, 592–596 (2015).
13. Ziv, N. E. & Brenner, N. Synaptic tenacity or lack thereof: spontaneous remodeling of synapses. *Trends in neurosciences* **41**, 89–99 (2018).
14. Clopath, C., Bonhoeffer, T., Hübener, M. & Rose, T. Variance and invariance of neuronal long-term representations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160161 (2017).
15. Lu, J. & Zuo, Y. Shedding light on learning and memory: optical interrogation of the synaptic circuitry. *Current Opinion in Neurobiology* **67**, 138–144 (2021).
16. Kozachkov, L., Lundqvist, M., Slotine, J.-J. & Miller, E. K. Achieving stable dynamics in neural circuits. *PLoS computational biology* **16**, e1007659 (2020).
17. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986-999. e16 (2017).
18. Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. *bioRxiv* (2020).
19. Marder, E., Goeritz, M. L. & Otopalik, A. G. Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Current opinion in neurobiology* **31**, 156–163 (2015).
20. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology* **37**, 66–74 (2016).
21. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585 (2013).
22. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Current opinion in neurobiology* **58**, 141–147 (2019).

23. 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).
24. Park, H.-J. & Friston, K. Structural and functional brain networks: from connections to cognition. *Science* **342**, (2013).
25. Shine, J. M. *et al.* The dynamics of functional brain networks: integrated network states during cognitive task performance. *Neuron* **92**, 544–554 (2016).
26. Westphal, A. J., Wang, S. & Rissman, J. Episodic memory retrieval benefits from a less modular brain network organization. *Journal of Neuroscience* **37**, 3523–3531 (2017).
27. Jackson, J. D. *Classical electrodynamics*. (American Association of Physics Teachers, 1999).
28. Haken, H. Synergetics of brain function. *International journal of psychophysiology* **60**, 110–124 (2006).
29. 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).
30. Ditzinger, T. & Haken, H. Oscillations in the perception of ambiguous patterns a model based on synergetics. *Biological Cybernetics* **61**, 279–287 (1989).
31. Jia, N. *et al.* Decoding of intended saccade direction in an oculomotor brain–computer interface. *Journal of neural engineering* **14**, 046007 (2017).
32. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology* **32**, 148–155 (2015).
33. Pinotsis, D. A., Brincat, S. L. & Miller, E. K. On memories, neural ensembles and mental flexibility. *NeuroImage* **157**, 297–313 (2017).
34. Schwartz, B. L., Chauhan, M. & Sadleir, R. J. Analytic Modeling of Neural Tissue: I. A

Spherical Bidomain. *The Journal of Mathematical Neuroscience* **6**, 1–20 (2016).

35. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature reviews neuroscience* **13**, 407–420 (2012).
36. Lindén, H. *et al.* Modeling the spatial reach of the LFP. *Neuron* **72**, 859–872 (2011).
37. Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol Cybern* **27**, 77–87 (1977).
38. Coombes, S. Mathematical neuroscience. *Journal of Mathematical Biology* **54**, 305–307 (2007).
39. Wilson, H. R. & Cowan, J. D. Mathematical Theory of Functional Dynamics of Cortical and Thalamic Nervous-Tissue. *Kybernetik* **13**, 55–80 (1973).
40. Humphries, M. D. Spike-train communities: finding groups of similar spike trains. *Journal of Neuroscience* **31**, 2321–2336 (2011).
41. Carroll, J. D. & Chang, J.-J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283–319 (1970).
42. Williams, A. H. *et al.* Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron* **98**, 1099-1115. e8 (2018).
43. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
44. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM review* **51**, 455–500 (2009).
45. ten Berge, J. M. *Least squares optimization in multivariate analysis*. (DSWO Press,

Leiden University Leiden, 1993).

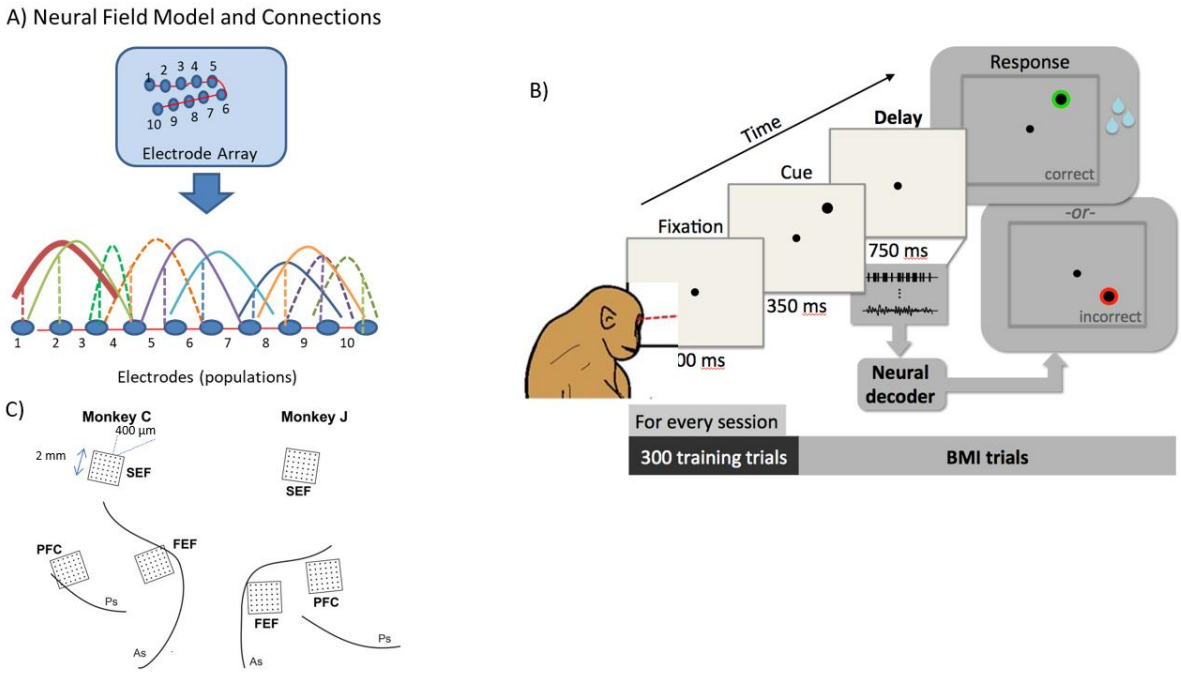
46. Bro, R. & Kiers, H. A. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics: A Journal of the Chemometrics Society* **17**, 274–286 (2003).
47. Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311 (1966).
48. Gehringer, K. R. & Redner, R. A. Nonparametric probability density estimation using normalized b–splines. *Communications in Statistics-Simulation and Computation* **21**, 849–878 (1992).
49. Heinz, S. *Statistical mechanics of turbulent flows*. (Springer Science & Business Media, 2013).
50. Amindavar, H. & Ritcey, J. A. Padé approximations of probability density functions. *IEEE Transactions on Aerospace and Electronic Systems* **30**, 416–424 (1994).
51. Abramowitz, M., Stegun, I. A. & Romer, R. H. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. (American Association of Physics Teachers, 1988).
52. Henriquez, C. S. Simulating the electrical behavior of cardiac tissue using the bidomain model. *Critical reviews in biomedical engineering* **21**, 1–77 (1993).
53. Roth, B. J. Electrical conductivity values used with the bidomain model of cardiac tissue. *IEEE Transactions on Biomedical Engineering* **44**, 326–328 (1997).
54. Ermentrout, G. B. & Cowan, J. D. A mathematical theory of visual hallucination patterns. *Biol Cybern* **34**, 137–50 (1979).
55. Pinotsis, D. A., Moran, R. J. & Friston, K. J. Dynamic causal modeling with neural fields. *Neuroimage* **59**, 1261–1274 (2012).
56. Eccles, J. C., Fatt, P. & Koketsu, K. Cholinergic and inhibitory synapses in a pathway

from motor-axon collaterals to motoneurones. *The Journal of physiology* **126**, 524–562 (1954).

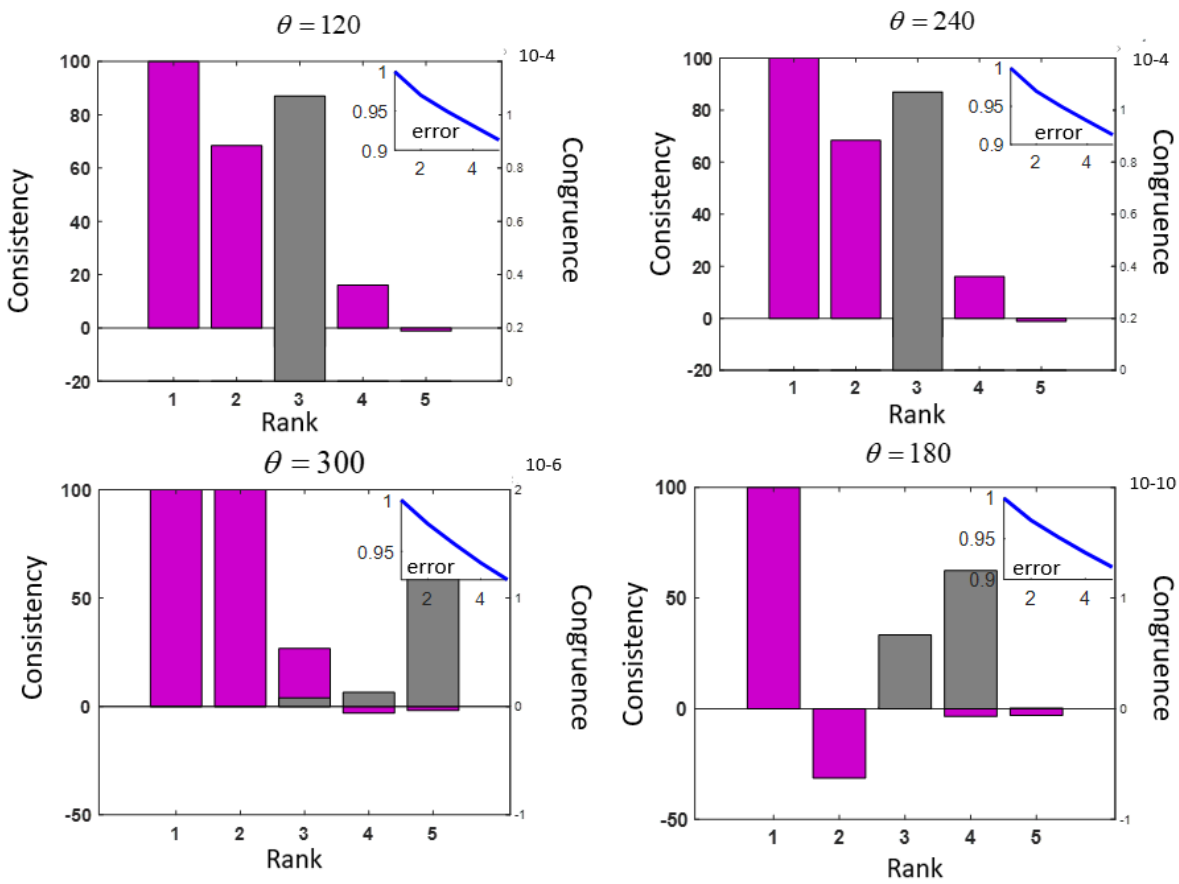
57. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology* **12**, e1004792 (2016).
58. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Physical review E* **69**, 026113 (2004).
59. Lancichinetti, A. & Fortunato, S. Consensus clustering in complex networks. *Scientific reports* **2**, 1–7 (2012).
60. Bruno, A. M., Frost, W. N. & Humphries, M. D. Modular deconstruction reveals the dynamical and physical building blocks of a locomotion motor program. *Neuron* **86**, 304–318 (2015).
61. Kiers, H. A. A three–step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. *Journal of Chemometrics: A Journal of the Chemometrics Society* **12**, 155–171 (1998).
62. Plonsey, R. The active fiber in a volume conductor. *IEEE transactions on biomedical engineering* 371–381 (1974).
63. Rush, S. & Driscoll, D. A. EEG electrode sensitivity-an application of reciprocity. *IEEE transactions on biomedical engineering* 15–22 (1969).
64. Mersmann, A. Crystallization technology handbook. *Drying Technology* **13**, 1037–1038 (1995).
65. Perkins, D. H. & Perkins, D. H. *Introduction to high energy physics*. (CAMBRIDGE university press, 2000).
66. Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A. & Tonegawa, S. Engram cells retain memory under retrograde amnesia. *Science* **348**, 1007–1013 (2015).

67. 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).
68. Fuchs, A., Jirsa, V. K. & Kelso, J. A. S. Theory of the relation between human brain activity (MEG) and hand movements. *Neuroimage* **11**, 359–369 (2000).
69. Jirsa, V. K. & Kelso, J. A. S. Spatiotemporal pattern formation in neural systems with heterogeneous connection topologies. *Physical Review E* **62**, 8462–8465 (2000).
70. Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience* **23**, 260–270 (2020).
71. Yu, B. M. *et al.* Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems* **21**, 1881–1888 (2008).
72. Tonegawa, S., Pignatelli, M., Roy, D. S. & Ryan, T. J. Memory engram storage and retrieval. *Current opinion in neurobiology* **35**, 101–109 (2015).
73. Pinotsis, D. A., Buschman, T. J. & Miller, E. K. Working Memory Load Modulates Neuronal Coupling. *Cerebral Cortex* (2018).
74. Domb, C. *Phase transitions and critical phenomena*. (Elsevier, 2000).

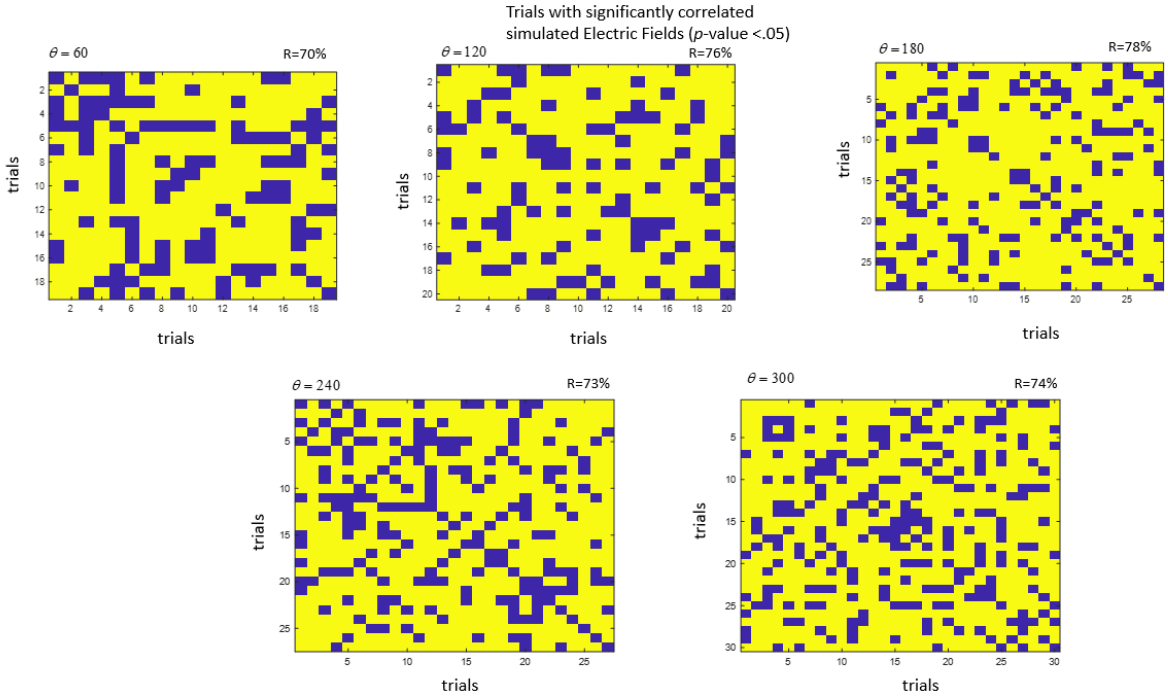
Supplementary Figures



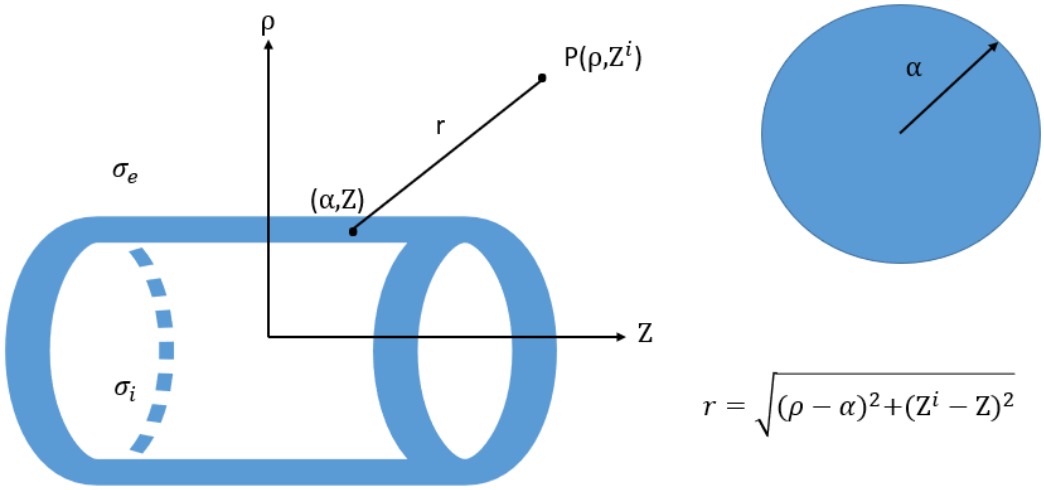
Supplementary Figure 1. A. Deep neural field model and connections. This is a biophysical rate model. It is obtained as a simplification of a neural network model of coupled excitatory—inhibitory populations. It provides a quantitative way to describe each ensemble's network interactions and patterns of activity across simultaneously recorded sites. The same model can describe different ensembles. Each electrode occupies a position on a cortical manifold *W* parameterized by the variable and is connected to all other electrodes with connections whose strength follows a Gaussian profile (coloured solid and dashed lines). **B.** Oculomotor spatial delayed response task. Monkeys hold the location of one of six randomly chosen visual targets in memory over a brief delay period and then saccade to the remembered location. If a saccade was made to the cued location, the target was presented with a green highlight and a water reward was delivered. Otherwise, the target was presented with a red highlight and reward was withheld. **C.** 32-electrode chronic arrays were implanted unilaterally in PFC, SEF and FEF in each monkey. Each array consisted of a 2 x 2 mm square grid, where the spacing between electrodes was 400um.



Supplementary Figure 2. Results of Canonical Decomposition. Panels show consistency and congruence for cued locations at θ=120, θ=240 (top) and θ=300, θ=180 (bottom) degrees. The format of each panel is the same as that of panels in Figure 1C. Consistency is shown using magenta bars, while congruence is shown using grey bars.



Supplementary Figure 3. *P*-values of correlations between single trial EF amplitudes for cued locations at θ=60, θ=120, θ=180 (top) and θ=240, θ=300 (bottom) degrees. The format of each panel is the same as Figure 3B. Yellow entries in the correlation matrix denote significant *p*-values, $p<.05$.



Supplementary Figure 4. Bidomain model for the electric field generated by an active fiber in a semi-conductor (see Methods for the meaning of various symbols).