Predicting perturbation effects from resting state activity using functional causal flow

Amin Nejatbakhsh†1 , **Francesco Fumarola**†2 , **Saleh Esteki**3 , **Taro Toyoizumi**2 , **Roozbeh Kiani***3,4,5 , **Luca Mazzucato***6

[1] *Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA*

[2] *Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, Wako, Saitama 351-0198, Japan*

[3] *Center for Neural Science, New York University, New York, NY 10003, USA*

[4] *Neuroscience Institute, NYU Langone Medical Center, New York, NY 10016, USA*

[5] *Department of Psychology, New York University, New York, NY 10003, USA*

[6] *Institute of Neuroscience and Departments of Biology, Mathematics and Physics, University of Oregon, Eugene, OR 97403, USA*

† *co-first authors*

* *co-corresponding authors*

E-mail: roozbeh at nyu dot edu; lmazzuca at uoregon dot edu

Abstract: Targeted manipulation of neural activity will be greatly facilitated by understanding causal interactions within neural ensembles. Here, we introduce a novel statistical method to infer a network's "functional causal flow" (FCF) from ensemble neural recordings. Using ground truth data from models of cortical circuits, we show that FCF captures functional hierarchies in the ensemble and reliably predicts the effects of perturbing individual neurons or neural clusters. Critically, FCF is robust to noise and can be inferred from the activity of even a small fraction of neurons in the circuit. It thereby permits accurate prediction of circuit perturbation effects with existing recording technologies for the primate brain. We confirm this prediction by recording changes in the prefrontal ensemble spiking activity of alert monkeys in response to single-electrode microstimulation. Our results provide a foundation for using targeted circuit manipulations to develop new brain-machine interfaces or ameliorate cognitive dysfunctions in the human brain.

Contents

1 Introduction

Complex cognition in humans and other primates is an emergent property of the collective interactions of large networks of cortical and subcortical neurons. Targeted manipulation of the brain to alter cognitive behavior will be greatly facilitated by understanding the causal interactions within ensembles. Examples of such manipulations include altering the perceptual judgement of motion direction in area MT (Salzman et al., 1990, 1992), or biasing object classification towards faces in the inferior temporal cortex (Afraz et al., 2006, Moeller et al., 2017, Parvizi et al., 2012). Perturbations are typically achieved via targeted electrical stimulation, a widely used technique in both humans and monkeys, as well as opto- and chemogenetic manipulations, techniques that have become more widely available for monkeys only recently. Here, we focus on electrical stimulation as other alternatives remain infrequently used in humans. Perturbations of neural circuits represent a promising avenue for ameliorating cognitive dysfunction in the human brain, as well as development of future brain-machine interfaces.

A crucial challenge in targeted perturbation is to identify perturbation sites, satisfying at least two requirements. The first is selectivity: the local neural population around the site should exhibit specific selectivity properties for the desired perturbation effect, e.g., motion direction selectivity in area MT (Salzman et al., 1990, 1992), face selectivity in face patches of inferotemporal cortex (Afraz et al., 2006, Moeller et al., 2017, Parvizi et al., 2012), or the locus of seizures in epilepsy (Fisher and Velasco, 2014). The second is efficacy: stimulation of the local population should exert some significant effect on the activity of the rest of the brain, and consequently on behavior. While selectivity of sensory and motor neurons may be estimated by recording neural activity in simple and well-defined tasks, selectivity tends to be quite complex or variable across tasks in many regions of the association cortex. Further, discovering efficacy is currently achieved by trial-and-error: many perturbations are performed until a site whose stimulation leads to a significant change in activity is located. As a result, current methods for targeted perturbations are labor intensive, time consuming, and often unable to generalize beyond the limited task set they are optimized for.

A promising avenue for predicting the efficacy of a potential perturbation site is to examine its functional connectivity within a local neural circuit. Intuitively, one expects that perturbing an afferent node with strong functional connectivity to other nodes within a circuit may exert stronger effects than perturbing the nodes that are functionally isolated. Estimating the functional connectivity in cortical circuits is a central open problem in neuroscience (Marinescu et al., 2018). Existing methods for estimating functional interactions between multi-dimensional time series are challenged by the properties of neural activity in the cortex (Reid et al., 2019). Cortical circuits comprise highly recurrent neural networks (Binzegger et al., 2004, Braitenberg and Schüz, 2013, Lefort et al., 2009, Thomson and Lamy, 2007), where the notion of directed functional couplings is not obvious. Correlation-based and inverse methods (Cocco et al., 2009) lack sufficient power when correlations are weak, as in most cortical circuits (Cohen and Kohn, 2011). Entropy-based methods require large datasets hard to acquire in conventional experiments. Granger causality (Dhamala et al., 2008, Faes et al., 2011, Granger, 1969) is challenged when the circuit's dynamical properties are not well known. Consequently, commonly encountered confounding effects —- phase delay (Vakorin et al., 2013), self-predictability in deterministic dynamics, or common inputs (Sugihara et al., 2012) —- render these methods unreliable (Brinkman et al., 2018, Vidne et al., 2012). It is thus of paramount importance to develop new theoretical tools.

These new tools should be able to estimate functional connectivity in the presence of common inputs using extremely sparse recordings, typical of cortical recordings in humans and monkeys. A promising approach is offered by delay embedding methods, e.g., convergent cross-mapping, which are capable of reconstructing nonlinear dynamical systems from their time series data. These methods are devised to work precisely in the sparse recording regime (Sauer et al., 1991, Takens, 1981) and in the presence of common inputs. While this powerful framework, rigorously articulated in (Cummins et al., 2015), has been successfully applied in ecology (Sugihara et al., 2012), and on *in vitro* (Sugihara et al., 2012, Tajima et al., 2017) and EcoG neural activity (Tajima et al., 2015), it has never been adapted to spiking activity *in vivo.*

We build on the delay embedding methods to develop a novel statistical approach for inferring causal functional connectivity ("causal flow") based on spiking activity of a simultaneously recorded ensemble in the cortex of awake monkeys (Fig 1). We first demonstrate our method on ground truth data from simulated continuous dynamical systems and then validate it on a biologically plausible model of a spiking cortical circuit. We show that causal flow captures a network's functional structure even in the extremely sparse recording regime, solely based on short snippets of resting state data. We then demonstrate that our method infers the causal flow of ensemble neurons from sparse recordings of spiking activity, obtained from chronically implanted prefrontal multi-electrode arrays in awake, resting monkeys. Using the causal flow inferred during the resting state, we successfully predict the effect of electrical microstimulations of single electrodes on the rest of the circuit.

2 Results

2.1 Uncovering the functional causal flow with delay embedding

To illustrate the concept and methods of functional causal flow (FCF), we examined a deterministic network Z, comprising N units $z_i = x_i, y_i$ arranged in two subnetworks X and Y, each endowed with their own local recurrent connectivity and, crucially, directed projections from X to Y with coupling strength g; but no feedback couplings from Y to X. We aimed to capture the intuitive idea that the "upstream" subnetwork X drives the activity of the "downstream" subnetwork Y (Fig. 2). It is well known from the theory of deterministic dynamical systems that one can (at least partially) reconstruct the N-dimensional attractor topology of a network of coupled units, represented by the vector time series of the activity of all units $\{\vec{z}(t)\}_{t=1:T}$, by using only the information encoded in the temporal trajectory of a single unit $\{z_i(t)\}_{t=1:T}$. From the mapping between the activity of the full network and the activity of a single unit, one can derive a map between the activity of the units themselves and (at least partially) reconstruct the activity of

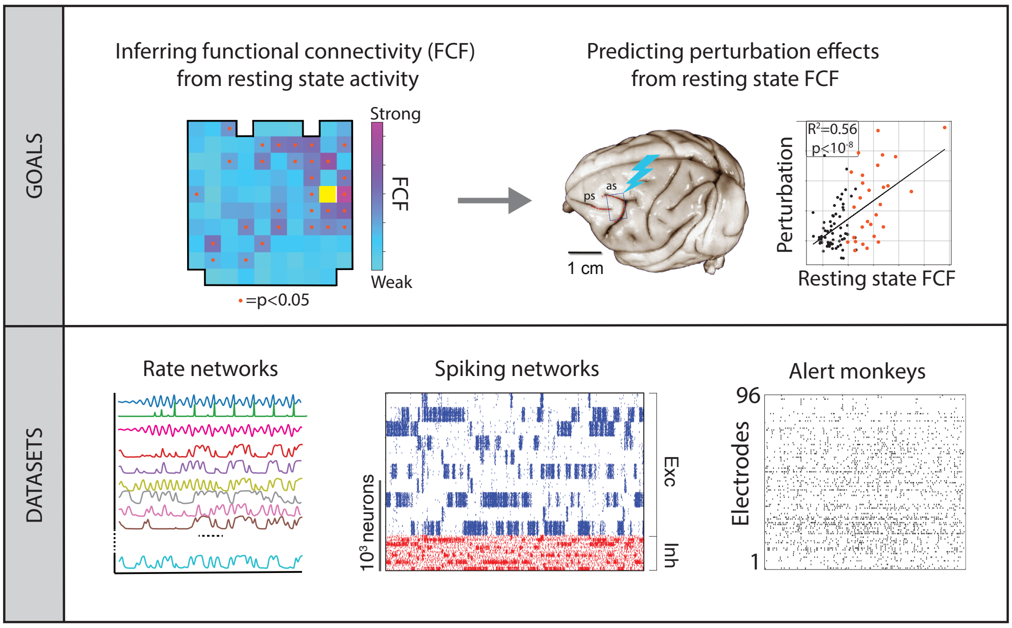

Figure 1. Conceptual summary. Top left: Functional causal flow (FCF) map inferred from alert monkey prefrontal cortex during resting state activity (yellow square: afferent electrode; orange circles: efferents with significant FCF to the afferent). Top right: Schematics of an electrical microstimulation experiment and prediction of stimulation effects from FCF (correlations of resting state FCF vs. perturbation effects on efferents with significant and non-significant FCF, orange and black dots, respectively). Bottom: We validated our method for predict perturbation effects from resting state FCF in three different datasets: a chaotic rate network, a spiking network with cell-type specific connectivity, and a prefrontal cortical circuit in alert monkeys.

one unit $\{z_i(t)\}_{t=1:T}$ from the activity of a different unit $\{z_j(t)\}_{t=1:T}$, for $i \neq j$. The reconstruction is possible whenever the two units are functionally coupled. This general property of dynamical systems is known as "delay embedding" (Sauer et al., 1991, Takens, 1981) and relies on a representation of network dynamics using "delay coordinates" (see Fig. 2A for details). This reconstruction was shown to be robust to noise in driven dynamical systems (Casdagli et al., 1991).

We used delay embedding to infer the FCF between all pairs of network units. We first considered the FCF between a unit y_i in the downstream subnetwork Y and a unit x_j in the upstream subnetwork X. The activity of unit x_j only depends on the other units in X, to which it is recurrently connected, but not on the units in Y, as there are no feedback couplings from Y to X. On the other hand, the activity of unit y_i depends both on the units in X, from which it receives direct projections, and on the other units in Y to which it is recurrently connected. In other words, $y_i(t)$ activity is influenced by units in both Y and X, whereas $x_j(t)$ activity depends only on other units in X. Thus, we expect that the reconstruction of $x_j(t)$ from $y_i(t)$ will be more accurate than the reconstruction of $y_i(t)$ from $x_j(t)$. For an intuitive explanation of this prediction consider that $y_i(t)$ reflects the

	FCF	Feature
Upstream	$F_{ij} > 0$ & sig; F_{ji} non-sig	j is causally upstream of i.
Downstream	F_{ij} non-sig; $F_{ji} > 0$ & sig	j is causally downstream from i.
Reciprocal	$F_{ij} \sim F_{ji}$ & both sig	i and j are reciprocally functionally connected.
Independent	F_{ij}, F_{ji} both non-sig	i and j are causally independent.

Table 1. Different cases of functional causal flow. Columns represent the units being reconstructed (afferents) given the activity of a row unit (efferent).

collective activity of X which determines the activity of $x_j(t)$. We tested our prediction by estimating the reconstruction accuracy $\rho(x_j|y_i)$ of the temporal series of unit $x_j(t)$ given $y_i(t)$. Reconstruction accuracy was quantified as the Fisher transform of the correlation between the *empirical* activity of unit x_j and its *predicted* activity obtained from unit y_i. The process was cross-validated to avoid overfitting (see Methods for details). Similarly, we estimated cross-validated reconstruction accuracy $\rho(y_i|x_j)$ of the temporal series of unit $y_i(t)$ given $x_j(t)$. As expected, reconstruction accuracy increased as a function of the dimensionality of the delay coordinate vector (i.e., how many time steps back we utilize for the reconstruction, Fig. 2A). The accuracy plateaued beyond a certain dimensionality (related to the complexity of the time series (Tajima et al., 2017)), whose value we fixed for our subsequent analyses.

We define the functional causal flow (FCF) from x_j to y_i as the Fisher z-transform of the reconstruction accuracy $F_{ij} = z[\rho(x_j|y_i)]$ (see Methods). In our conventions, a column of the FCF represents the *afferent* unit, whose activity is being reconstructed, given the activity of an *efferent* unit in a row-wise element. As explained above, the FCF was estimated using a cross-validation procedure to avoid overfitting. We established statistical significance by comparing the FCF estimated from the empirical data with that estimated from surrogate datasets carefully designed to preserve the temporal statistics of the network activity while destroying its causal structure (see Fig. S1 and Methods for details). Unlike the usual pairwise correlation r_{ij}, which is a symmetric quantity, the FCF is a directed measure of causality. By comparing the value and significance of F_{ij} with Fji, we can establish the directionality of the functional relationship between y_i and x_j, uncovering several qualitatively different cases which we proceed to illustrate (Table 1).

In the example above, the reconstruction accuracy of x_j given y_i was significant and large, while that of y_i given x_j was not significant. In other words, while one can significantly reconstruct x_j with high accuracy from y_i, because the latter receives information from the former, the opposite is not possible, matching predictions based on the simulated network. We refer to x_j as being *causally upstream* to y_i in the network functional causal flow.

2.2 Hierarchical structure of functional causal flow

The notion of being causally upstream or downstream is an entirely *functional* relation and *a priori* different from the underlying structural/anatomical coupling between units. We illustrate here two more examples from the network in Fig. 2 to reveal the variety of the relationships encoded in the FCF. We considered the FCF between $x_1(t)$ and $x_3(t)$ within

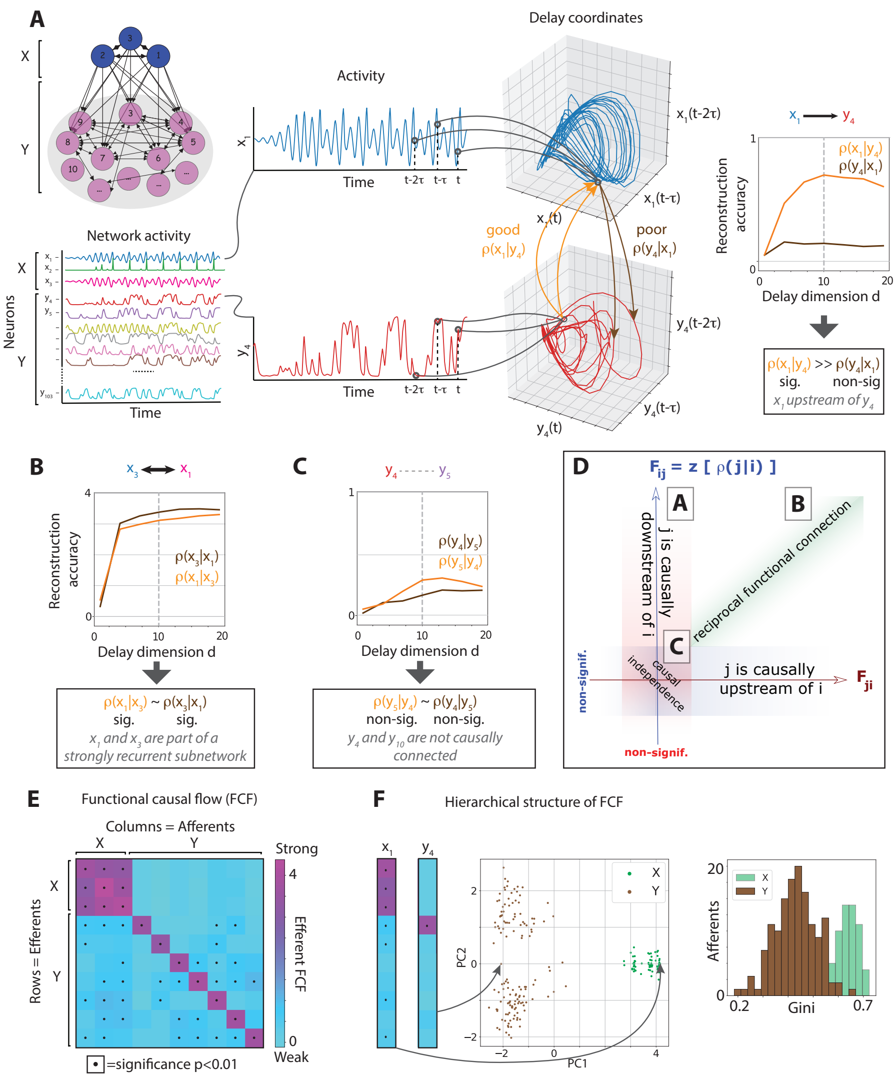

Figure 2. Functional causal flow. A) Left: Schematic of network architecture Z: two subnetworks X (blue nodes) and Y (pink nodes) comprising strong and weak recurrent couplings, respectively, are connected via feedforward couplings from X to Y (thickness of black arrows represents the strength of directed structural couplings). Center: Activity of units $y_4(t)$ (orange, bottom) and $x_1(t)$ (in blue, top) are mapped to the delay coordinate space $X_1 = [x_1(t), x_1(t-\tau), \ldots, x_1(t-(d-1)\tau)]$and Y_4 (right). Reconstruction accuracy increases with delay vector dimension d before plateauing. The reconstruction accuracy $\rho(x_1|y_4)$ of upstream unit x_1 given the downstream unit y_4 is significant and larger than the reconstruction accuracy $\rho(y_4|x_1)$ of y_4 given x_1 (non-significant). The FCF value F_{41} reveals a strong and significant functional connectivity from upstream node x_1 to downstream node y_4. B) The significant FCF between two units x_1 and x_3 within the strongly coupled subnetwork X reveal strong and significant causal flow between them, but no preferred directionality of causal flow. C) The non-significant FCF between two units y_4 and y_5 in the weakly coupled subnetwork Y suggests the absence of a causal relationship. D) Summary of the FCF cases in panels A, B, C (see Table 3). E) The FCF between 10 representative units sparsely sampled from the network (columns and rows represent afferents and efferent units, respectively; columns are sorted from functionally upstream to downstream units). F) The functional hierarchy in the network structure in encoded in the causal vectors (Left: PCA of columns of the FCF matrix F_{ij}, see Methods; Right: Gini coefficient of causal vectors).

the subnetwork X, whose units are part of a Rossler attractor, a well studied dynamical system (see Fig. 2B and Methods). Because the X subnetwork does not receive inputs from other network units in Y, it is causally isolated (i.e., its activity is conditionally independent from Y). Hence one can reconstruct the activity of one x_i unit from another with high accuracy, yielding large and significant values for both $\rho(x_1|x_3)$ and $\rho(x_3|x_1)$. This is a classic demonstration of the embedding theorem (Takens, 1981), ensuring accurate bidirectional reconstruction of variables mapping a chaotic attractor. The large and significant F_{13} and F_{31} reveal that the unit pair has a strong functional coupling, and the two units lie at the same level of the functional hierarchy. This is unlike the case of pairs x_i, y_j described above, where a significant F_{ij} but a non-significant F_{ji} showed a strong directional coupling and a functional hierarchy (Stark et al., 1997). As another qualitatively different pair, we considered two units y_4 and y_5 within the subnetwork Y, whose units are only sparsely recurrently coupled. The FCFs were not significant for this pair ((Fig. 2C), suggesting that the two units are functionally independent, namely, their activities do not influence each other significantly. The taxonomy of causal flows are summarized in Table 1 and Fig. 2D, and the FCF matrix F for sparsely sampled units from both X and Y is shown in Fig. 2E.

The variety of FCF features discussed so far suggests that, even if the FCF is a measure of pairwise causal interactions, it may reveal a network's global causal structure. We thus analyzed the N-dimensional *causal vectors* $\mathbf{f^{(i)}} = [F_{ki}]_{k=1}^{N}$, representing the (z-tranform of the) reconstruction accuracy of unit i given the activity of each one of the efferents k. The causal vector $\mathbf{f^{(i)}}$ encodes the FCF from unit i to the rest of the network. For example, a significant positive entry k of the causal vector implies that the afferent unit i has a strong functional coupling with efferent k. A Principal Component analysis of the causal vectors from a sparse subsample of the network units (10 out of 103) revealed a clear hierarchical structure present in the network dynamics showing two separate clusters corresponding to the subnetworks X and Y (Fig. 2F). Thus, causal vectors revealed the global network functional hierarchy from sparse recordings of the activity.

We further quantified the hierarchical functional structure of causal vectors, measured by their Gini coefficients (Fig. 3E). In the absence of hierarchies, one would expect all efferents from a given afferent unit to have comparable values, namely, yielding a low Gini coefficient. Alternatively, heterogeneity of FCFs across efferents for a given afferent would suggest a network hierarchy with a gradient of functional connectivities, yielding a large Gini coefficient. For our simulated network, we found a large heterogeneity in the distribution of causal vectors Gini coefficients, capturing the functional hierarchy in the network. For comparison, when restricting the causal vectors to afferents in either X or Y (green and brown bars in Fig. 2E, respectively), we found a clear separation with larger Gini coefficients for X afferents and lower Gini coefficients for Y afferents. This result shows that the feedforward structural couplings from X to Y introduce a hierarchy in the full network Z, encoded in the network causal vectors. Importantly, inferring this structure does not require observing the full network and can be achieved by recording from a small subset of the network units.

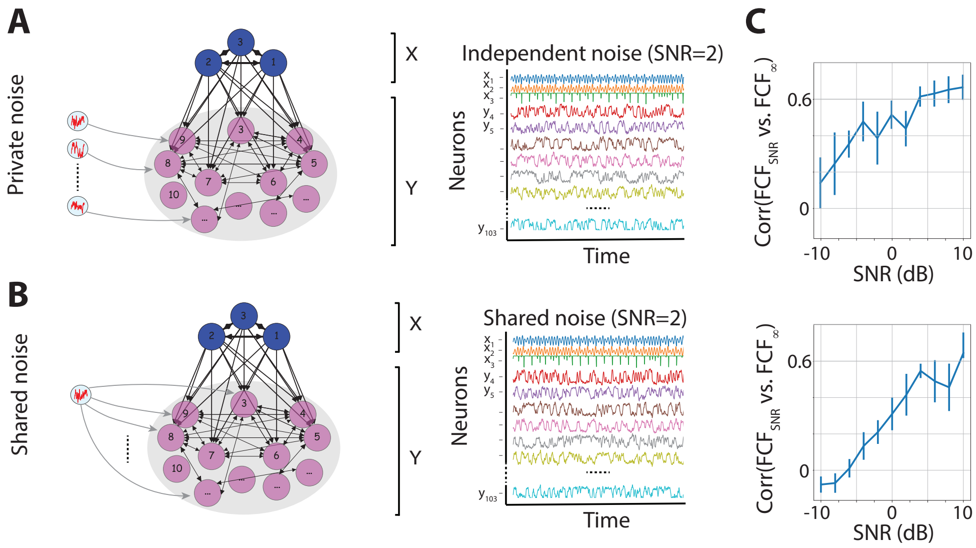

Figure 3. Robustness of functional causal flow inference (FCF). FCF is robust to both private noise (panel A, i.i.d. realizations of noise injected in each unit) or shared noise (panel B, a scalar noise source modulates all neuron equally). Same network as in Fig. 2. C) FCF inferred from the noiseless simulations (FCF_∞) is similar to the FCF inferred from simulations with varying SNR ratio (SNR, defined as $10\log_{10}[\sigma(signal)/\sigma(noise)]$, where σ =standard deviation, measure in dB) for private noise (top) and shared noise (bottom).

2.3 Robustness of functional causal flow estimation

Neural circuits in vivo are characterized by several sources of noise including both private (e.g., Poisson variability in spike times) and shared variability (e.g. low-rank co-fluctuations across the neural ensemble (Kanashiro et al., 2017, Rabinowitz et al., 2015)), where the latter may correlate to the animal's internal state such as attention or arousal (Dadarlat and Stryker, 2017, Engel et al., 2016, Huang et al., 2019, Ruff and Cohen, 2014). Based on previous theoretical work, we expected our delay embedding framework to be reasonably resilient against noise (Casdagli et al., 1991).

To quantify the robustness of inferred FCFs, we tested their changes as a function of the strength of a noise source injected in subnetwork Y. When driving the network with either private noise (i.i.d. for each neuron, Fig. 3A) or shared noise (same noise realizations across all neurons, Fig. 3B), we found that FCF inference degraded only when the signal-to-noise ratio (SNR) dropped below 0 dB (Fig. 3C, SNR is measured in logarithmic scale). The change was more precipitous for shared than private noise as expected. However, for a wide range of SNRs, the FCF inference maintained its accuracy. The degradation caused by private noise did not remove the informativeness of FCFs for the tested range of SNRs down to -10 dB, and shared noise became irrecoverably detrimental only for SNRs below -5.

We thus conclude that causal flow estimates are robust to both private and shared sources of neural variability.

We also explored another source of variability common in the experimental data: changes in the arousal state of an animal that can influence the dynamics of neural activity. We performed multiple simulations of the same network where each simulation varied in regard to initial conditions, and thereby the sequence of recorded neural activity. We found that FCF was indistinguishable across the simulated sessions with different initial conditions (not shown).

Our validation results thus demonstrate that the data-driven discovery of functional causal flow is robust to noise. Another source of variability is the presence of unobserved units in the network, which we will address below in Section 2.5

2.4 Inferred causal flow predicts the effects of perturbation

Can we predict the effects of perturbations on network activity from the causal flow inferred in the unperturbed system? We hypothesized that the effects of stimulating a specific node on the rest of the network can be predicted by the causal flow inferred during the resting state.

We simulated a perturbation protocol where we artificially imposed an external input on one afferent network unit for a brief duration, mimicking electrical or optical stimulation protocols to cortical circuits. We estimated the stimulation effect on each efferent unit, by comparing the distribution of binned activity in each efferent in intervals preceding the stimulation onset and following its offset (Fig. 4B). We found that stimulation exerted complex spatiotemporal patterns of response across efferent units, which we captured in the *perturbation vector*: $\mathbf{I^{(i)}} = \{I_{ki}\}_{k=1}^{N}$, where I_{ik} is the interventional connectivity matrix (Fig. 4B). Stimulation effects across efferents k strongly depended on the afferent unit i that was stimulated. Perturbation effects increased with stimulation strength for afferent-efferent pairs in $X \to X$, $X \to Y$ and $Y \to Y$, but did not depend on stimulation strengths for pairs $Y \to X$, consistent with the underlying structural connectivity lacking feedback couplings $Y \to X$ (Fig. 4C). Can one predict the complex spatiotemporal effects of stimulation solely based on the FCF inferred during resting state activity?

We hypothesized that, when manipulating afferent unit i, its effect on efferent unit k could be predicted by the FCF estimated in the absence of perturbation (Fig. 4D). Specifically, we tested whether stimulation of afferent unit i would exert effects only on those efferent units k that have significant FCFs, F_{ki}; but no effects on units whose FCFs were not significant. We found a strong correlation between FCF and perturbation effects (Fig. 4D). More specifically, we confirmed the three hypotheses above, namely, we found that the perturbation effects on the efferent units were localized on units with significant FCFs (red dots in Fig. 4D and Fig. 4E)); no effects were detected on upstream or unengaged units. In particular, we found that pairs where the stimulated afferent was in Y and the efferent in X did not show any significant effects of perturbations (black dots in Fig. 4D); this was expected given the absence of feedback couplings $Y \to X$. Two crucial features of the FCF, underlying its predictive power, were its directed structure and its causal properties, which are not present in alternative measures of functional connectivity such as

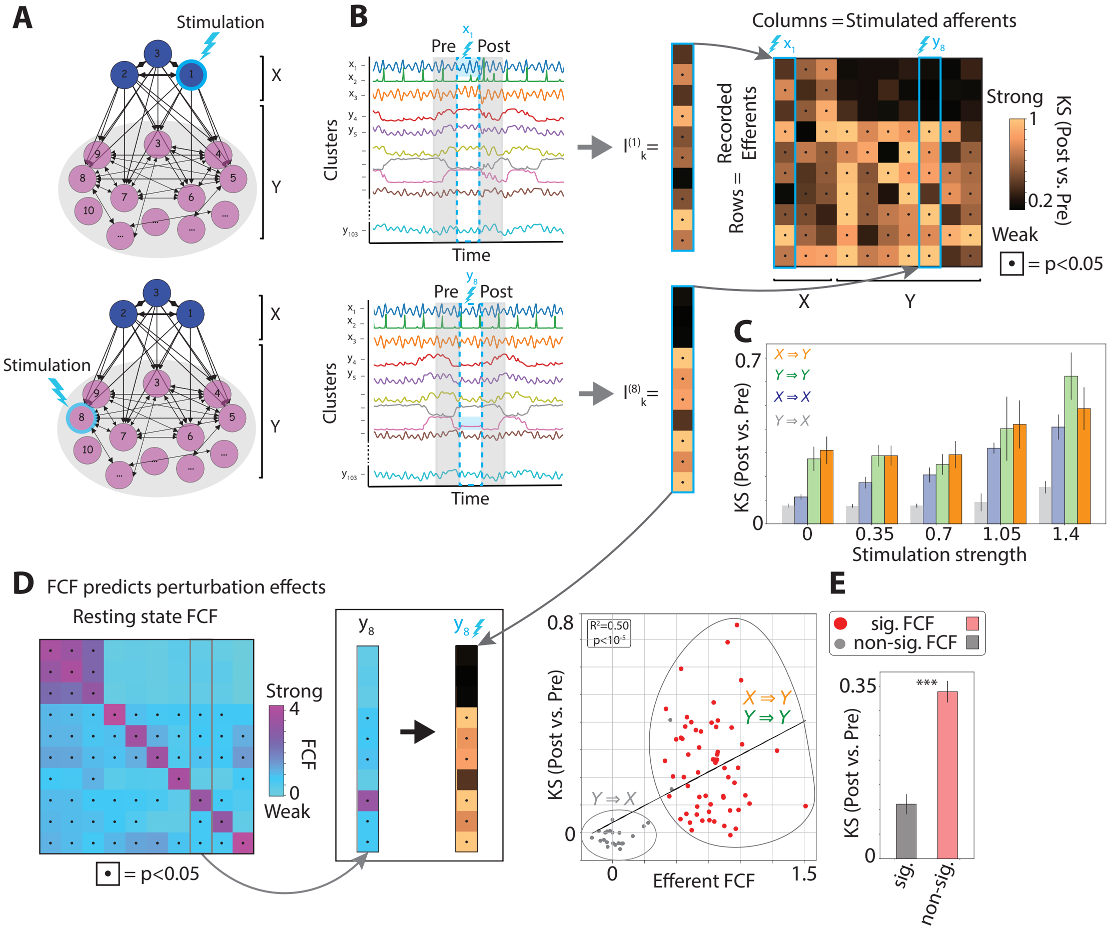

Figure 4. Causal flow predicts perturbation effects. A) Perturbation protocol: single nodes are stimulated with a pulse of strength S lasting for 100ms (representative trials with stimulation of units x_3 and y_8, top and bottom, respectively). B) Perturbation effects on efferent units are estimated by comparing the activity immediately preceding onset and following offset of the perturbation (Kolmogorov-Smirnov test statistics, black dot represents significant effect, $p < 0.05$). The effects of stimulating one afferent i on all efferents k is encoded in the perturbation vector $\mathbf{I^{(i)}}$. C) Perturbation effects increase with the stimulation strength S for afferent-efferent pairs in populations $X \to X$, $Y \to Y$, and $X \to Y$, but not $Y \to X$, reflecting the absence of feedback structural couplings from Y to X. D) For each afferent, its causal vector (column of the resting state FCF matrix representing unit y_4) is compared with the perturbation vector (columns of the interventional connectivity matrix), revealing that FCF predicts perturbation effects (linear regression of causal vectors vs. interventional vectors, $R^2 = 0.50$, $p < 10^{-5}$; red and gray dots represent significant and non-significant FCF pairs, respectively). E) Efferent units with significant resting state FCF (red and gray dots in the scatterplot of panel C) had a larger response to perturbation, compared to pairs with non-significant FCF (t-test, $*** = p < 10^{-6}$).

pairwise correlations, which failed at predicting the effects of stimulation in our data (not shown).

We thus conclude that the causal effect of perturbations on network units can be reliably and robustly predicted by the FCF inferred during the resting state (i.e., in the absence of the perturbation). Moreover, the specific features of the FCF predict the effects of perturbation at the fine grained level of pairs of units.

2.5 Causal flow from sparse recordings in spiking circuits

To apply our predictive framework to cortical circuits in behaving animals, we aimed at extending the method outlined above to encompass the following additional issues. First, while the network of Figs. 2-4 comprised real-value continuous rate units, neurons in cortical circuits exhibit spiking activity. Second, cortical circuits are characterized by strong recurrent connectivity (Binzegger et al., 2004, Braitenberg and Schüz, 2013, Lefort et al., 2009, Thomson and Lamy, 2007) obeying Dale's law, with cell-type specific connectivity of excitatory (E) and inhibitory (I) neurons. Third, spike trains recorded with commonly-used multi-channel electrode arrays typically yield extremely sparse recordings of the underlying circuit activity: the number of active contacts in these electrodes (tens to hundreds) captures activity from only a small fraction of neurons in a circuit ($<$1%). Crucially, each electrode records the aggregate spiking activity of a neural cluster, namely, a small number of neurons in a cortical column surrounding the electrode, some of which can be isolated as single units. We thus sought to extend our methods to address these critical issues, by performing a series of simulated experiments on a spiking neural network. Can we reliably infer functional causal flow between recorded neurons using sparse recordings of spiking activity at the level of neural clusters?

We inferred the FCF from sparse recordings of spiking activity in a large simulated cortical circuit (Fig. 5). In this model E and I spiking neurons were arranged in clusters, consistent with experimental evidence supporting the existence of functional clusters in cortex (Kiani et al., 2015, Lee et al., 2016, Perin et al., 2011, Song et al., 2005). In the model, E/I pairs of neurons belonging to the same cluster have potentiated synaptic couplings, compared to weaker couplings between pairs of neurons belonging to different clusters. Resting state activity in the clustered network displayed rich spatiotemporal dynamics in the absence of external stimulation, whereby different subsets of clusters activated at different times (with a typical cluster activation lifetime of a few hundred ms, Fig. 5A-B). These metastable dynamics were previously shown to capture physiological properties of resting state activity in cortical circuits (Litwin-Kumar and Doiron, 2012, Mazzucato et al., 2015, 2016, 2019, Rostami et al., 2020, Wyrick and Mazzucato, 2020).

We estimated the FCF from short periods of resting state activity (8 simulated seconds) from sparse recordings of E neurons (Fig. 5C, 10 neurons per cluster, 3% of total neurons in the network, only neurons with firing rates above 5 spks/s on average were retained for further analyses). Visual inspection of the sparse FCF in Fig. 5C suggested the presence of two causal functional hierarchies in the circuit: the first hierarchy between strong intra-cluster functional couplings and weaker inter-cluster couplings; and the second hierarchy between different clusters. We confirmed the first, large hierarchy and found that the distributions of FCFs for pairs of neurons belonging to the same cluster, was significantly larger than the distribution of FCFs for the pairs belonging to different clusters (Fig. 5C, t-test, $p < 10^{-20}$).

In primate recordings from multi-electrode arrays, each electrode typically records the aggregated spiking activity of a neural cluster surrounding the electrode. To model this scenario, we then examined the properties of the FCF between neural clusters in our simulated network. We investigated whether any hierarchy was present in the causal functional connectivity between different clusters. Close inspection of the FCF for pairs of neurons belonging to different clusters revealed the existence of clear off-diagonal blocks, suggesting the presence of a structure in the FCF between different clusters (Fig. 5C) at the *mesoscopic* level, namely, at the level of neural populations rather than single neurons.

We sought to quantify this mesoscopic structure by testing whether FCF F_{ij} between pairs of neurons $i \in A$ and $j \in B$ encoded the identity of the efferent cluster A and afferent cluster B that the neurons were sampled from. We thus sampled small "ensemble FCF" matrices, obtained from subgroups of neurons, each consisting of one randomly sampled excitatory neuron per cluster (Fig. 5D, six neural clusters were considered). We hypothesized that, if the ensemble FCF captured the identity of the clusters from which neurons were recorded, then the causal vectors $\mathbf{f^{(i)}} = \{F_{ki}\}_{k=1}^{N}$ (i.e., the columns of the ensemble FCF matrix, $N = 6$ in Fig. 5D) corresponding to afferent neurons i in the same cluster would be highly correlated, thereby allowing us to infer the network connectivity from sparsely recorded neurons. However, a naive dimensionality reduction on the columns of the FCF matrix would have trivially led to the emergence of groups simply due to the fact that the diagonal entries of the FCF (self-reconstructability of the afferents) were much larger than the off-diagonal ones $F_{ii} >> F_{ij}|_{j \neq i}$. This fact reflected the strong hierarchy in the intra- vs. inter-cluster FCF discussed above because we had sampled one neuron per cluster. In order to control for this diagonal effect and examine the smaller inter-cluster structure effects, we removed the diagonal from the $N \times N$ ensemble FCF matrices, and considered $N-1$-dimensional "between-cluster" causal vectors $\mathbf{f^{(i)}_{betw}} = \{F_{ki}\}_{k \neq i}$ (Fig. 5E). We found that the between-cluster causal vectors grouped in Principal Component space according to the cluster membership of their afferents (Fig. 5E). We further confirmed the existence of this mesoscopic structure by showing that the correlation between between-cluster causal vectors from afferents belonging to the same cluster was much larger than the one obtained from afferents in different clusters (from Fig. 5E). These results show that FCF is a property shared by all neurons within the same neural cluster.

Building on this insight, we thus introduced the *mesoscopic* FCF as the coarse-grained causal flow between neural clusters, defined as the average FCF of the neurons in that cluster (block-average of FCF, see Fig. 5F). Between-cluster causal vectors from the mesoscopic FCF stand at the center of each group of causal vectors from the between-cluster ensemble FCF (Fig. 5E), thus recapitulating the causal properties of each neural cluster. This mesoscopic causal flow is an emergent property of the clustered network dynamics and arises from the only source of quenched heterogeneity in the network, namely, the Erdos-Renyi sparse connectivity in the structural couplings.

Together, these results uncovered a nested hierarchy of causal flow in a biologically plausible model of recurrent cortical circuits based on functional clusters. The first hierarchy separates large within- from small between-cluster FCF; the second hierarchy reveals the mesoscopic functional organization between neural clusters. Crucially, the mesoscopic FCF suggests that our theory can be reliably applied to aggregate spiking activity of neural clusters from micro-electrode arrays in primate.

2.6 Predicting perturbation effects from causal flow in spiking circuits

Is resting state FCF predictive of perturbation effects in the case of a sparsely recorded spiking network? To investigate this question, we devised a stimulation protocol whereby we briefly stimulated single neural clusters and examined the effect of stimulation on the activity of efferent ensemble neurons (Fig. 6). This stimulation protocol was designed to model perturbation experiments in alert monkeys (see below), where a brief electrical microstimulation of a single electrode on a multi-electrode array directly perturbs the cortical column surrounding the electrode, represented in our model by a neural cluster. Perturbation effects in the model were estimated by comparing the network activity immediately following the offset and preceding the onset of the stimulation (see Methods and Fig. 6A, representative clusters 2 and 3 were stimulated). The effects of stimulating a specific afferent neural cluster were encoded in the perturbation vector $\mathbf{I_k^{(i)}}$, estimated for each afferent neuron i in the stimulated cluster, and for the sparsely recorded efferent neurons k. The entries in the perturbation vector represent the Kolmogorov-Smirnov test statistics between pre- and post-stimulation spiking activity aggregated over several stimulation trials of the same neural cluster (Fig. 6B).

We found that perturbations exerted an afferent-specific effect on network activity, whereby stimulating neurons in different clusters led to differential patterns of responses across the clusters (Fig. 6B). We tested whether FCF was predictive of these perturbation effects. We had found above that the FCF between pairs of neurons belonging to the same cluster was much stronger than between pairs belonging to different clusters (Fig. 5C). We found the same hierarchy in the perturbation vectors, whereby the perturbation effects for efferents belonging to the same stimulated cluster were much larger than for efferents in different clusters (Fig. 6B). This hierarchy of perturbation effects closely matched the hierarchy present in the structural connectivity (Fig. 5A) and found in the resting state FCF (Fig. 5C). Overall, we found a strong correlation between resting state FCF and perturbation effects (Fig. 6D).

A crucial feature of the FCF was that it captured the fine-grained mesoscopic structure of causal flow between different clusters, encoded in the between-cluster FCF (Fig. 5). We thus set out to test whether the FCF could predict the effect of perturbations on efferent neurons belonging to different clusters from the stimulated one. For this analysis, we consider "between-cluster" causal vectors and perturbation vectors, obtained by masking all efferents belonging to the stimulated cluster, thus retaining only the efferents belonging to the other, non-stimulated clusters (Fig. 6C). We found a strong correlation for "between-cluster" causal vectors and perturbation vectors (Fig. 6D). We then performed a more fine-grained analysis of the predictive relation between resting state FCF and perturbation vectors, by separating the efferent neurons into those with significant FCF (“functionally connected”) and non-significant FCF (“unconnected") to each stimulated afferent. Our

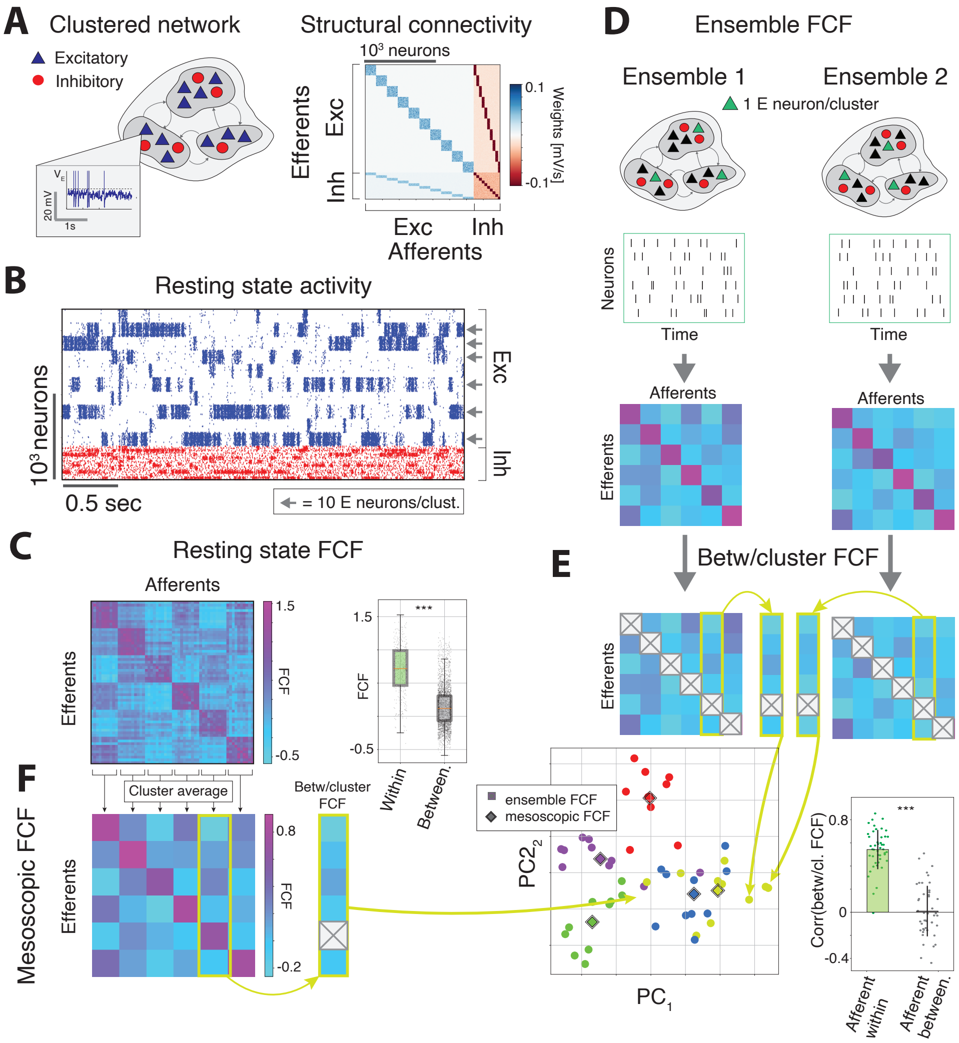

Figure 5. Mesoscopic causal flow in a sparsely recorded spiking circuit. A) Left: Schematics of the spiking network architecture with E and I neural clusters (inset: membrane potential trace from representative E cell), defined by potentiated within-assembly synaptic weights (right: blue and red represent positive and negative weights, respectively). B) Raster plots from a representative trial showing the entire network activity during the resting state (blue and red marks represent E and I cells action potentials, respectively; neurons arranged according to cluster membership). C) Left: Functional causal flow (FCF) between sparsely recorded activity (10 E cells per cluster were recorded, only six clusters firing above 5 spks/s on average were recorded; grey arrows in panel B). Right: A large hierarchy of FCF values reveals the separation between pairs of neurons belonging to the same cluster (green) and different clusters (orange), reflecting the underlying anatomical connectivity (mean±SD, t-test, ***$= p < 10^{-20}$). D) Ensemble FCFs inferred from two different recorded ensembles (one E cell were recorded from each of six clusters). E) Top: The between-cluster causal vectors (i.e., the columns of the off-diagonal ensemble FCF matrices) reveal the existence of a cluster-wise structure and group according to the cluster membership of their afferent neurons (left: Principal Component Analysis of between-cluster causal vectors: circles and rhomboids represent, respectively, between-cluster ensemble and mesoscopic vectors; right: Pearson correlations between vectors whose afferents belong to the same cluster, left, or different clusters, right; t-test $*** = p < 10^{-20}$).

theory predicted that perturbation effects would be much stronger for functionally connected efferents, compared to the other efferents. We confirmed this prediction for the

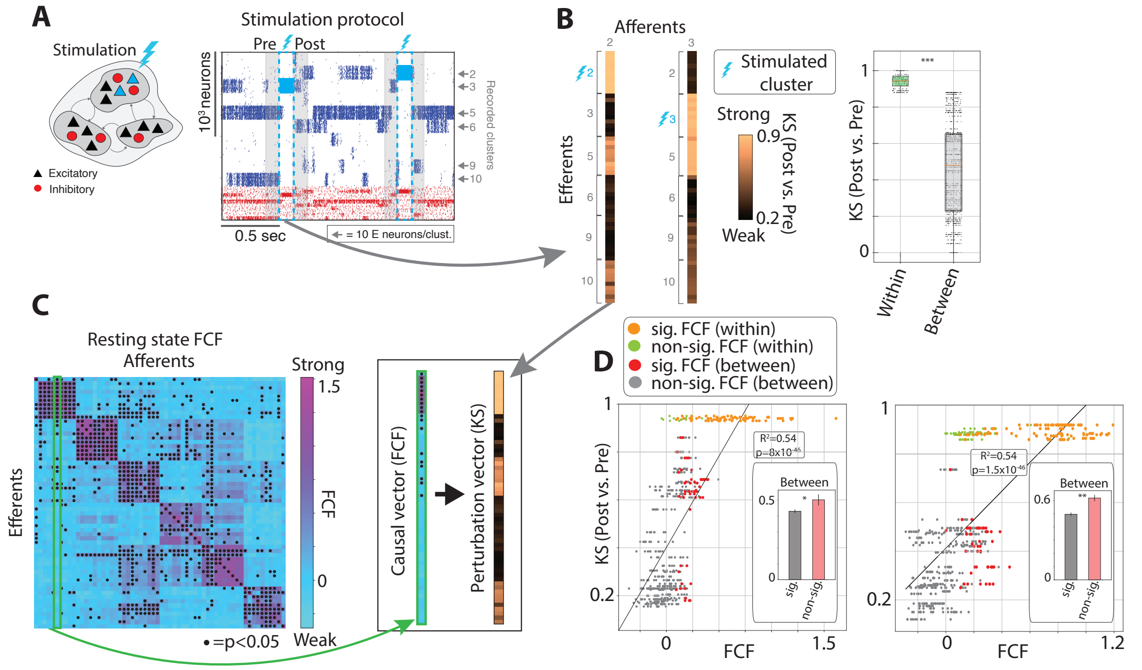

Figure 6. Causal flow predicts perturbation effects in spiking circuits. A) Perturbation protocol in the network model: Clusters 2 and 3 were stimulated for 100ms (blue interval in raster plot, two representative stimulation trials). B) Perturbation effects for a subset of efferent cells in clusters 2, 3, 5, 6, 9, 10 were estimated by comparing the spike count distribution during post- vs. pre-stimulation intervals (shaded gray areas; only clusters with average firing rate above 5 spks/s were considered; 10 E neurons per recorded cluster) and captured by the Kolmogorov-Smirnov test statistics encoded in the perturbation vectors. Right: Perturbation effects were larger for efferent neurons in the same cluster as the stimulated one (Within), compared to efferents in different clusters (Between). T-test, ***$= p < 10^{-20}$. C) Resting state FCF causal vectors were used to predict between-cluster perturbation vectors (black dots: significant FCF pair, $p < 0.05$). D) Resting state FCF is strongly correlated to perturbation vectors for afferent neurons in clusters 2 and 3 (same stimulated channels as panels A and B; pairs of neurons within the same cluster: orange and green for significant and non significant FCF; pairs in different different clusters: red and gray for significant and non-significant FCF). Pairs of neurons in different clusters were grouped into subsets with significant and non-significant FCF (red and gray bars and dots), based on their FCF inferred during the resting state. Between-cluster pairs with significant FCF exhibited stronger perturbation effects when the afferent was stimulated, compared to pairs with non-significant FCF (insets in scatterplots; t-test, *,**$= p < 0.05, 0.01$).

between-cluster afferent-efferent pairs, thus demonstrating that causal flow was predictive of perturbation effects even in the sparse recordings regime (Fig. 6D).

These combined set of results show that in a cortical circuit model based on functional clusters, the effects of perturbations can be robustly captured at the mesoscopic level of neural clusters. The causal flow between clusters inferred during the resting state reliably predicted the effect of stimulation in a biologically plausible model of a cortical circuit. These results were obtained in the sparse recording regime, typical of cortical recordings in experimental protocols, which we investigate next.

2.7 Inferring the causal flow from resting state activity in alert monkeys

To test our theory, we performed an experiment comprising simultaneous recording and stimulation of spiking activity in alert monkey prefrontal cortex (pre-arcuate gyrus, area 8Ar) during a period of quiet wakefulness (resting state) while the animal was sitting awake in the dark. The experiment had two phases (Fig. 1 and 7). In the first phase, we recorded population neural activity from a multi-electrode array (96-channel Utah array, with roughly one electrode in each cortical column in a $4 \times 4\text{mm}^2$ area of the cortex) during the resting state, and we estimated the FCF between pairs of neural clusters (multiunit activities collected by each recording electrode). In the second phase, we perturbed cortical responses by a train of biphasic microstimulating pulses (15 μA, 200 Hz) to a cluster for a brief period (120ms) and the recorded population neural activity across the array before and following each pulse train.

We first examined whether causal flow could be estimated reliably for the recorded population, which constituted a small fraction of neurons in the circuit. Our modeling study suggested that FCF can be defined at the mesoscopic level as a property of functional assemblies or neural clusters: FCF inferred from different neurons recorded from the same cluster yielded consistent results (Fig. 5). Following previous experimental evidence supporting the existence of assemblies in monkey pre-arcuate gyrus (Kiani et al., 2015), we reasoned that the activity of neural clusters around each electrode may represent sparse samples from a local cortical assembly. This experimental setup thus provided a similar scenario to the one validated in our modeling study (Figs. 5).

We found that FCF was characterized by a complex set of spatiotemporal features (Fig. 7A, for the full 96×96-dimensional FCF matrix see Fig. S2). In Fig. 7A we show four representative 96-dimensional causal vectors representing the FCF for each of four different afferents clusters recorded in two different sessions (channels 14 and 56 from session 1 and channels 42 and 29 from session 2). We overlaid the causal vectors onto the array geometry (location of recording electrodes in the array) for illustration (each array-geometry causal vectors in Fig. 7A corresponds to a specific column of the full FCF matrix in Fig. S2).

Comparison of the causal vectors across afferents revealed remarkable features about the structure of the functional connectivity. First, FCF is channel-specific, namely, it depends on the afferent clusters whose activity is being reconstructed. Second, each causal vector shows a hierarchical structure, with significant FCF in a subset of downstream efferents, while most efferents cannot reconstruct the afferent activity (Fig. 7A). This result is qualitatively consistent with the FCF obtained from our models (Fig. 3F and 6), supporting the hypothesis of functional hierarchies embedded within prefrontal cortical circuits (Kiani et al., 2015).

Given an afferent cluster, is FCF of efferent clusters uniformly distributed across the array, or is there a preferential spatial footprint of FCF? We found a spatial gradient whereby FCF was largest in the efferent clusters immediately surrounding the afferent cluster, while FCF for distant efferents typically plateaued at low but nonzero values (Fig. 7C). We thus concluded that FCF inferred during the resting state was cluster-specific and revealed a hierarchy of functional connectivity where functionally downstream neural clusters are spatially localized around the afferent cluster. These results extend previous correlation analyses of spatial clusters in alert monkeys (Kiani et al., 2015) highlighting a spatial gradient of directed functional couplings at the mesoscale level.

2.8 Perturbation effects on cortical circuits in alert monkeys

We next proceeded to examine the effect of microstimulation on the cortical activity in alert monkeys. We estimated perturbation effects by comparing the activity of neural clusters in the intervals preceding the onset and following the offset of the stimulation of the afferent, for each pair of stimulated afferent and recorded efferent (see Fig. 7B). Perturbation effects were quantified via a Kolmogorov-Smirnov test statistics aggregated over all stimulations of a specific neural cluster (comparison between the pre- and post-perturbation distributions of activity, see Methods).

We first examined the spatiotemporal features of stimulation effects. We found that perturbations exerted a strong effect on ensemble activity, and that these effects where specific to which afferent channel was stimulated (Fig. 7B; perturbation effects for each stimulated afferent i are visualized as a perturbation vector $\mathbf{I^{(i)}}$ overlaid on the array geometry). By comparing the effects of perturbing a single cluster across all efferents, we found a hierarchical structure with strong effects elicited in specific subsets of efferent clusters. The identity of strongly modulated efferents was specific to the stimulated channel. Remarkably, we found that the hierarchical structure of perturbation effects had a clear spatial gradient, where strongly perturbed efferents were most likely located close to the stimulated cluster, while distant efferents were less affected by perturbation, though the effects were nonzero even far away from the stimulated afferent (Fig. 7D). Strikingly, this spatial gradient closely aligned to the spatial gradient we found in the hierarchy of resting state FCF (Fig. 7C).

2.9 Predicting perturbation effects from resting state activity in alert monkeys

Our theory posits that the effects of stimulation of afferent cluster i on the other efferent neural clusters can be predicted by the corresponding causal vector $\mathbf{f^{(i)}}$ inferred at rest (i.e., a column of the FCF matrix; four representative causal vectors are overlaid on the array geometry in Fig. 7A). Specifically, our theory predicts that perturbing an afferent cluster exerts a strong effect on those efferent clusters which have a strong functional connectivity to the afferent, identified by a significant resting state FCF as read out from the afferent causal vector. Moreover, perturbation effects on efferents with significant FCF should be stronger compared to efferents with non-significant FCF. Visual inspection of the resting state FCF causal vectors (Fig. 7A) and comparison to the map of perturbation effects (Fig. 7B, perturbation vectors) suggest that the FCF and perturbations fo are strikingly similar for a given afferent. We confirmed this intuition quantitatively and found that the FCF inferred at rest was indeed predictive of perturbation effects at the level of single stimulated afferent (Fig. 7E, Pearson correlations between causal vectors and perturbation vectors). In particular, we found that for all stimulated clusters, the effect of a perturbation was significantly stronger on efferents with strong functional connectivity to the stimulated cluster compared to efferents with weak functional connectivity, as predicted by our theory

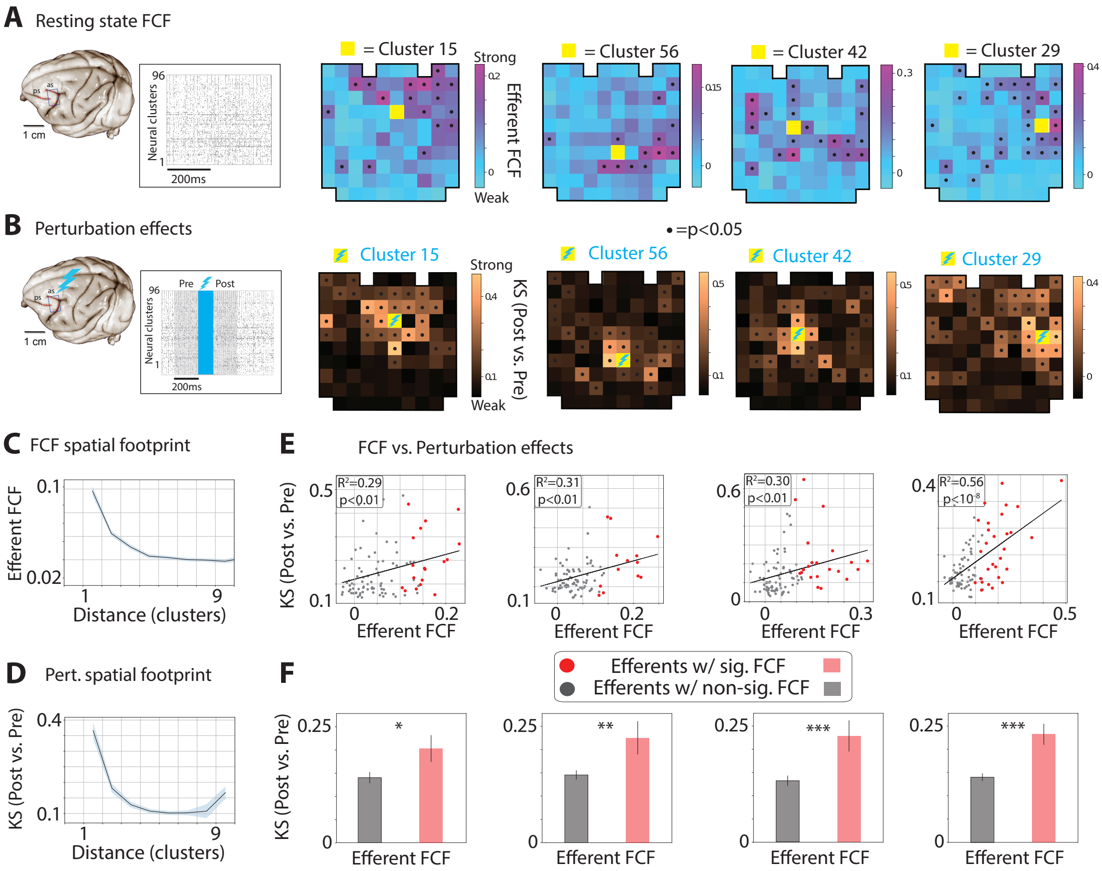

Figure 7. Causal flow predicts perturbation effects in alert monkeys. A) Left: Ensemble spiking activity in representative session from multi-electrode array activity in the pre-arcuate gyrus during quiet wakefulness (black tick marks are spikes from each neural cluster, defined as the aggregated spiking activity around each recording electrode). Right: FCF inferred from resting state activity for four representative afferent clusters (clusters 15 and 56 from session 1 and clusters 42 and 29 from session 2; yellow squares represent the reconstructed afferent cluster for each causal vector; full FCF matrix in Fig. S2). FCF causal vectors for each afferent are overlaid to the array geometry (black dots represent significant FCF values, established by comparison with surrogate datasets, $p < 0.05$, see Fig. S1 and Methods). B) Left: The perturbation effect from electrical microstimulation of cluster 15 (120ms stimulation train, blue shaded area) was estimated by comparing the activity in the 200ms intervals immediately preceding and following the perturbation (grey shaded areas). Right: Perturbation effects from four stimulated clusters (same clusters as in A) overlaid on the array geometry (Kolmogorov-Smirnov test statistics between post- vs. pre-perturbation activity distribution; black dots represent a significant difference, $p < 0.05$). C) The spatial footprint of resting state FCF decays with increasing distance of the efferent from the afferent cluster (mean±s.e.m. across 96 clusters in two sessions). D) Spatial footprint of perturbation effects for the four stimulated clusters decays with increasing distance from the stimulated cluster (mean±s.e.m. across four stimulated afferents). E) Resting state FCF predicts perturbation effects. For each stimulated afferent, the perturbation effects on all efferent clusters are shown (Kolmogorov-Smirnov test statistics between post- and pre-stimulation activity) as functions of the corresponding resting state FCF (gray and red dots represent efferents with non-significant and significant FCF, respectively, $p < 0.05$; black line: linear regression, R^2 and p-value reported). F) For each stimulated afferent in panel E, aggregated perturbation effects are larger over efferents with significant resting state FCF vs. efferents with non-significant FCF (mean±s.e.m. across gray and red-circled dots from panel E; t-test, $*, **, *** = p < 0.05, 0.01, 0.001$).

(Fig. 7F). The predictive power of FCF held at the level of single stimulated afferents, thus achieving a high level of granularity in prediction. Because both the FCF and perturbations displayed a characteristic decay proportional to the distance from the afferent electrode, we tested whether the predictive relation between them still held after controlling for this potential confound. Remarkably, after removing the spatial dependence the predictive relation between FCF and perturbation still held for the residuals (Fig. S3).

These results demonstrate that the causal flow estimated from sparse recordings during the resting state accurately predicts the effects of perturbation on the neural ensemble with extreme precision, at the single channel level, thus establishing the validity of our theory in cortical circuits of alert primates.

3 Discussion

Predicting the effect of targeted manipulations on the activity of cortical circuits is a daunting task but it could be achieved by capturing the causal functional interactions in the circuit. A central challenge using common multi-electrode arrays in monkeys and humans is the extremely sparse recording regime, where the activity of only a small fraction of neurons in a circuit is observed. In this regime, traditional methods fail due to unobserved neurons and common inputs to the circuit.

Here, we demonstrated a new methods for inferring causal functional interactions within a circuit from sparse recordings of spiking activity: the functional causal flow (FCF). We validated the method on ground truth data showing that FCF captures the structural, functional, and perturbational connectivity in a biologically plausible model of a cortical circuit, even when recording only a small fraction of the network's neurons. Using FCF inferred during the resting state in the network, we predicted the effect of perturbing a neural cluster on the rest of the recorded neurons, revealing the set of efferent neurons with directed functional couplings to a given afferent. Remarkably, when applying FCF to spiking activity from ensemble recordings in alert monkeys, we were able to reconstruct the causal functional connectivity between the recording electrodes using resting state activity. The resting state FCF predicted the effect of single-electrode microstimulation on efferent electrodes which were classified as functionally coupled to the afferent. Our results establish a new avenue for predicting the effect of stimulation on a neural circuit solely based on sparse recordings of its resting state activity. They also provide a new framework for discovering the rules that enable generalization of resting state causal interactions to more complex behavioral states, paving the way toward targeted circuit manipulations in future brain-machine interfaces.

3.1 The role of resting state activity

In traditional neurophysiological studies, resting state activity is defined as a pre-stimulus background activity, immediately preceding stimulus presentation, and it is regarded as random noise or baseline, devoid of useful information (Heggelund and Albus, 1978, Vogels et al., 1989, Werner and Mountcastle, 1963). However, recent results have challenged this widely held picture, producing evidence that resting state activity may encode fundamental information regarding the functional architecture of neural circuits.

Studies investigating the dependence of neural responses on the background activity, quantified with local field potentials (Fontanini and Katz, 2008, Gervasoni et al., 2004), single neuron membrane potentials (McGinley et al., 2015), or population spiking activity (Engel et al., 2016), found that it encodes information about the animal's behavioral state, including even fine grained movements (Musall et al., 2019, Salkoff et al., 2020, Stringer et al., 2019). Moreover, resting state activity immediately preceding stimulus onset predicts the trial-to-trial variability in stimulus evoked response, potentially explaining the observed dependence of sensory responses on the underlying state of the network (Arieli et al., 1996, Super et al., 2003).

Recent studies have suggested that resting state activity is finely structured, containing information on the functional architecture of the neural circuits (Kenet et al., 2003, Kiani et al., 2015, Tsodyks et al., 1999) and providing a repertoire of network patterns of activation (Luczak et al., 2007, 2009, Mazzucato et al., 2015), potentially linked to developmental plasticity (Berkes et al., 2011, Fiser et al., 2004). The population coupling of single neurons estimated during the resting state *in vivo* is correlated with the synaptic input connection probability measured *in vitro* from the same cortical circuit (Okun et al., 2015). Also, in neuronal cultures, the causal functional connectivity inferred from ongoing activity is predictive of the structural connectivity estimated from electrical stimulation: functionally downstream neurons have faster response latency to stimulation compared to functionally upstream neurons (Tajima et al., 2017).

3.2 Circuit models of resting state activity

We validated our theoretical framework for causal inference using two classes of models: a continuous rate network and a network model of resting state activity in a cortical circuit. The latter is a biologically plausible model based on a recurrent spiking network where excitatory and inhibitory neurons were arranged in functional assemblies (Litwin-Kumar and Doiron, 2012, Mazzucato et al., 2015, Rostami et al., 2020, Wyrick and Mazzucato, 2020). Experimental evidence including multielectrode recordings in behaving monkeys (Kiani et al., 2015) strongly supports the existence of functional assemblies in cortex (Lee et al., 2016, Perin et al., 2011, Song et al., 2005, Wong et al., 2016). Clustered spiking networks capture complex physiological properties of cortical dynamics during resting and stimulus-evoked activity due to the metastable dynamics of cluster activations. Such physiological properties include context- and state-dependent changes in neural activity and variability (Deco and Hugues, 2012, Litwin-Kumar and Doiron, 2012, Mazzucato et al., 2015, 2016, Rostami et al., 2020, Schaub et al., 2015); as well as neural correlates of behavior and cognitive function such as expectation, arousal, locomotion, and attention (Mazzucato et al., 2019, Rostami et al., 2020, Wyrick and Mazzucato, 2020).

3.3 Estimating functional connectivity

Estimating the underlying connectivity is a formidable task, especially susceptible to errors in the presence of strong recurrent couplings and unobserved common inputs ubiquitous in cortical circuits. Even when unlimited data are available, sophisticated methods typically fail in the presence of strong correlations between unconnected neurons (Das and Fiete, 2020).

We define functional interactions as the causal interaction of cortical neurons or task-relevant regions. Existing methods for estimating functional interactions between multi-dimensional time series include linear regression (Semedo et al., 2019), Granger causality (Bressler and Seth, 2011), and inter-areal coherence (Bastos and Schoffelen, 2015, Sun et al., 2004). While correlation-based methods are problematic for weak correlations, entropy-based methods such as transfer entropy (Schreiber, 2000) are extremely data hungry. Detecting a clear causal relationship by transfer entropy (Schreiber, 2000) or Granger causality (Dhamala et al., 2008, Faes et al., 2011, Geweke, 1982, Granger, 1969) is not straightforward unless the system's dynamical properties are well known, due to the confounding effects of phase delay (Vakorin et al., 2013), self-predictability in deterministic dynamics (Sugihara et al., 2012) or common inputs (Brinkman et al., 2018, Vidne et al., 2012).

Alternatives such as inverse methods based on Ising models utilize time-consuming learning schemes (Tkacik et al., 2006) though recently faster algorithms have been proposed (Cocco et al., 2009, Maoz et al., 2020). Other approaches applicable to spike trains include generalized linear models (Pillow et al., 2008) or spike train cross-correlograms (English et al., 2017). Remarkably, the latter method was successfully validated using optogenetic perturbations *in vivo*.

Inferring causal functional connectivity from *extremely sparse recordings* of neural activity is a long standing problem. Here, we proposed a new method to estimate causal functional connectivity ("functional causal flow" or FCF) from observing spiking activity in the absence of perturbations. Our method relies on delay embedding techniques used for reconstructing nonlinear dynamical systems from their time series data with convergent cross-mapping (Sugihara et al., 2012). Crucially, convergent cross-mapping was designed to work precisely in the sparse recording regime (Sauer et al., 1991, Takens, 1981), where other methods fail. While this powerful framework has been successfully applied in ecology (Sugihara et al., 2012), and previously applied to EcoG data (Tajima et al., 2015) and in vitro spiking data (Tajima et al., 2017), here we pioneered its use for estimating causal functional connectivity from spiking activity of a neural population in awake monkeys. Using synthetic ground truth data from recurrent spiking networks, we showed that FCF can be reliably estimated using extremely sparse recordings and very short samples of neural activity (tens of seconds, Fig. 5); and that FCF is robust to private and shared sources of noise typically encountered in cortical circuits (Fig. 3).

3.4 Hierarchical structures in cortical circuits

Previous studies strongly support the existence of a hierarchical structure in brain architecture both at the whole-brain level (Felleman and Van, 1991, Harris et al., 2019) as well as locally within single cortical areas (Arieli et al., 1996, Okun et al., 2015). In the latter case, ensemble neurons were ranked based on the degree to which their activity correlated with the average population activity (i.e., soloist and choristers). Here, we took a step beyond correlational analysis and revealed the hierarchical structure of causal interactions using resting state data. For each afferent channels, we were able to classify its efferents as functionally "upstream" or "downstream" within the network's causal flow. When the ground truth structural connectivity is known, as in our simulated networks, we showed that the causal hierarchy may reflect structural couplings and the existence of neural assemblies. Surprisingly, we found that causal hierarchies within a local circuit may naturally emerge from heterogeneities in recurrent couplings between neural assemblies, even in the absence of a feedforward structure.

In cases where the structural connectivity is not accessible, as in alert primate recordings, we showed that the causal hierarchy inferred from resting state predicts the effect of perturbations via electrical microstimulation. Specifically, perturbation of an afferent electrode affects more strongly efferent electrodes which are functionally coupled to the afferent compared to those that are not functionally connected. Remarkably, we uncovered a strong spatial gradient in the structure of directed functional interactions, consistent with previous results in alert monkeys (Kiani et al., 2015).

3.5 Microstimulation effects on neural activity in primates

Microstimulation experiments have played a crucial role for our understanding of the organization and function of neural circuits in the primate brain. Among many successful examples are microstimulation of motion-selective middle temporal (MT) neurons to alter choice (Salzman et al., 1990), reaction time (Ditterich et al., 2003), or confidence (Fetsch et al., 2014) of monkeys performing a direction discrimination task. However, outside of sensory or motor bottlenecks of the brain, the use of microstimulation (or other perturbation techniques) is fraught with challenges. In many regions of the primate associate cortex, neurons have complex and task-dependent selectivities. Identifying these selectivities is often time-consuming and may not be always possible. Further, perturbation of the activity of these neurons does not necessarily lead to behavioral changes commensurate with their empirically defined selectivities, partly because selectivities in one task condition may not generalize to others and partly due to network effects of perturbations beyond the directly manipulated neurons.

3.6 Model-based approach to perturbation experiments in primates

Our approach to quantify FCF based on the activity of a large neural population spread out in multiple neighboring cortical columns provides an easily implementable solution with many advantages. First, we directly assess the network effects of the activity of each neuron, and thereby generate predictions about the impact that perturbing the activity of one cluster of neurons will have on the rest of the population. Second, population activity has proven quite powerful in revealing the neural computations that underlie behavior, with features that are robust to the exact identity of the recorded neurons and their complex selectivities (Kiani et al., 2014, Mante et al., 2013, Pandarinath et al., 2018, Trautmann et al., 2019). We suggest that characterizing FCF during a task and using our model-based approach to predict the impact of a variety of perturbations on the population level representations offer an attractive alternative to the traditional trial-and-error approaches where different neural clusters are manipulated in search for a desirable behavioral effect.

We speculate that our model-based approach may lead to crucial advances in brain-machine interfaces if one can use FCF inferred from resting state activity to predict perturbation effects during a task – a direction we will actively pursue in the future.

Two key challenges in interpretation of typical microstimulation experiments are: (i) indirect activation of distant neurons through the activation of the neural cluster around the stimulating electrode, and (ii) effects on fibers of passage that could cause direct activation of neurons distant to the stimulating electrode (Histed et al., 2013). Our approach directly addresses the first challenge by mapping the FCF based on the ensemble activity. The second effect acts as noise in our approach because the FCF is quantified based only on the activity of the neurons recorded by the electrodes. The success of our approach (Fig. 7) suggests that this noise is not overwhelming. The robustness of our approach likely stems from the synergy of our delay embedding methods with our focus on the population neural responses and the large number of simultaneously recorded neural clusters in our experiments, which effectively capture key features of the intrinsic connectivity in the circuit (Fig. 2) (Kiani et al., 2015, Litwin-Kumar and Doiron, 2012, Mazzucato et al., 2015, 2019, Rostami et al., 2020).

Current methods for establishing site efficacy for perturbation experiments are labor intensive, time consuming, and often unable to generalize beyond the limited task set they are optimized for. Here, we demonstrated a new statistical method capable of predicting the impacts and efficacy of a targeted microstimulation site using only the resting state activity. Crucially, our method can directly be applied to monkeys and humans, where commonly used "large-scale" recording technologies often permit sampling from only a small fraction of neurons in a circuit (typically $< 1\%$). Our method is thus likely to improve the safety and duration of the procedure, a key step toward targeted circuit manipulations for ameliorating cognitive dysfunction in the human brain, as well as development of future brain-machine interfaces.

4 Methods and Materials

4.1 Network models

4.1.1 Rate network

The network consists of 100+3 nodes, 3 of which belong to the subnetwork X following Rossler dynamics described by the equations below:

$$\begin{cases} \frac{dx_1}{dt} = -x_2 - x_3 \\ \frac{dx_2}{dt} = x_1 + \alpha x_2 \\ \frac{dx_3}{dt} = \beta + x_3(x_1 - \gamma) \end{cases}$$

Other nodes in subnetwork Y evolve according the following dynamics:

$$\frac{dy}{dt} = -\lambda y + 10\tanh(J_{YX}x + J_{YY}y) + I$$

As observed from the above equations nodes in X are uni-directionally projecting to nodes in Y. The weight matrix J_{YX} connecting X to Y is the product of a scalar g_i (connection

Rate network simulations		
Parameter	Description	Value
g_r	Strength of the recurrent weights	4
g_i	Strength of X to Y weights	0.1
p	Connection probability from X to Y	1
α	Rossler parameter	0.2
β	Rossler parameter	0.2
γ	Rossler parameter	5.7
λ	Recurrent time constant	1

Table 2. Parameters for the rate network.

strength) and a binary matrix where the elements are sampled from $Bernoulli(p)$. The recurrent weight matrix J_{YY} is drawn from $\mathcal{N}(0, g_r)$. It is shown that increasing g_r will transition the network into a chaotic regime. See Table 2 for a description of the model parameters and their values.

4.1.2 Spiking network

We modeled the cortical circuit as a network of $N = 2000$ excitatory (E) and inhibitory (I) spiking neurons ($n_E = 80\%$ and $n_I = 20\%$ relative fractions). Connectivity was Erdos-Renyi with connection probabilities given by $p_{EE} = 0.2$ and $p_{EI} = p_{IE} = p_{II} = 0.5$. When a synaptic weight from pre-synaptic neuron j to post-synaptic neuron i was nonzero, its value was set to $J_{ij} = j_{ij}/\sqrt{N}$, with j_{ij} sampled from a gaussian distribution with mean $j_{\alpha\beta}$, for $\alpha, \beta = E, I$, and variance δ^2. E and I neurons were arranged in $p = 10$ clusters of equal size Pairs of neurons belonging to the same cluster had potentiated synaptic weights by a ratio factor $J^+_{\alpha\beta}$, for $\alpha, \beta = E, I$. Network parameters were chosen to generate spontaneous metastable dynamics with a physiologically realistic cluster activation lifetime, consistent with previous studies (Jones et al., 2007, Mazzucato et al., 2015, 2019, Wyrick and Mazzucato, 2020). Parameter values are in Table 3.

We used leaky-integrate-and-fire (LIF) neurons whose membrane potential V evolved according to the dynamical equation

$$\frac{dV}{dt} = -\frac{V}{\tau_m} + I_{rec} + I_{ext} \ ,$$

where τ_m is the membrane time. When V hits threshold V^{thr}_α (for $\alpha = E, I$), the neuron emits a spike and V is held at reset V^{reset} for a refractory period τ_{refr}. Thresholds were chosen so that the homogeneous network (i.e.,where all $J^{\pm}_{\alpha\beta} = 1$) was in a balanced state with rates $(r_E, r_I) = (2, 5)$ spks/s (Amit and Brunel, 1997, Mazzucato et al., 2019, Wyrick and Mazzucato, 2020). Input currents contained a contribution I_{rec} from the recurrent connections and an external current $I_{ext} = I_0 + I_{pert}(t)$ (units of mV s^{-1}). The first term I_0 is a constant term representing input to the E or I neuron from other brain areas. For each neuron, I_0 it is drawn from a uniform distribution in the interval $[I_{0\alpha}(1-a_0), I_{0\alpha}(1+a_0)]$, where $I_{0\alpha} = N_{ext} J_{\alpha 0} r_{ext}$ (for $\alpha = E, I$), $N_{ext} = n_E N p_{EE}$, and $a_0 = 2.5\%$. $I_{pert}(t)$

Spiking network simulations		
Parameter	Description	Value
j_{EE}	mean E-to-E synaptic weights $\times\sqrt{N}$	0.24
j_{IE}	mean E-to-I synaptic weights $\times\sqrt{N}$	0.45
j_{EI}	mean I-to-E synaptic weights $\times\sqrt{N}$	3.07
j_{II}	mean I-to-I synaptic weights $\times\sqrt{N}$	-.15
J_{E0}	mean external input weights to E neurons $\times\sqrt{N}$	0.058
J_{I0}	mean external input weights to I neurons $\times\sqrt{N}$	0.052
J_{EE}^{+}	E-to-E within-cluster weight potentiation factor	31.5
J_{IE}^{+}	mean E-to-I synaptic weights between clusters	9.3
J_{EI}^{+}	mean I-to-E synaptic weights between clusters	8.6
J_{II}^{+}	mean I-to-I synaptic weights between clusters	6.2
p_{EE}	mean E-to-E connection probability within clusters	.2
p_{IE}	mean E-to-I connection probability within clusters	.5
p_{EI}	mean I-to-E connection probability within clusters	.5
p_{II}	mean E-to-E connection probability within clusters	.5
r_{ext}	Average baseline afferent rate to E and I neurons	5 spks/s
V_E^{thr}	E threshold potential	1.43 mV
V_I^{thr}	I threshold potential	0.74 mV
V^{reset}	E and I reset potential	0 mV
τ_m	E and I membrane time constant	20 ms
τ_{refr}	E and I absolute refractory period	5 ms
τ_s	E and I synaptic time constant	5 ms

Table 3. Parameters for the spiking network in Fig. 5.

represents the time-varying afferent perturbation (see below). The recurrent term evolved according to

$$\tau_{syn}\frac{dI_{rec}}{dt} = -I_{rec} + \sum_{j=1}^{N} J_{ij} \sum_{k} \delta(t - t_k) \ ,$$

where τ_s is the synaptic time, J_{ij} are the appropriate recurrent couplings and t_k represents the time of the k-th spike from the j-th presynaptic neuron. Parameter values are in Table 3.

We modeled the effect of perturbations as a 100ms-long constant step input increase to the external current I_{pert}. Simulations were performed using custom software written in Python. Simulations in the resting state comprised 80s. Each network was initialized with random synaptic weights and simulated with random initial conditions in each trial. Python code to simulate the model during resting state and perturbations is located at https://github.com/amin-nejat/CCM.

4.2 Functional causal flow estimation

The algorithm used for functional causal flow estimation was based on the convergence cross-mapping (CCM) method, proposed in ecosystem analysis (Sugihara et al., 2012) and only recently tested in a *in vitro* electrophysiology (Tajima et al., 2017).

In contrast to more traditional causality-detection algorithms based on information transfer, which test the prediction of a downstream time series through information from an upstream one, CCM operates through "nowdiction" (reconstruction of simultaneous time segments) of an upstream series through information from a downstream one. The functional causality relation in between any pair of units is then determined by comparing the accuracy of nowdiction in the two directions. As ensured by a powerful theorem (Sauer et al., 1991), nowdiction of the upstream channels tends toward zero error in the asymptotic limit of infinite data size.

Each time segment of data is encoded as a so-called delay vector, by choosing an embedding dimension d and a delay time τ and constructing higher dimensional vectors

$$X^{(d)}(t) = [x(t), x(t-\tau), \ldots, x(t-d\tau+\tau)]$$

from the given time series $x(t)$ (Fig. 1). The high-dimensional time series constructed through embedding lives on a manifold diffeomorphic to the full attractor only as long as the embedding dimension d is larger than twice the dimensionality of the attractor (Takens, 1981), a statement that generalizes to the box-counting dimension for fractal attractors (Sauer et al., 1991). This can be far smaller than the number of variables involved in processing, as routinely happens in brain activity during any given task. Two parameters are thus involved in the embedding procedure, the embedding dimension d and the delay time tau. The condition that d be large enough is sufficient for an ideal setting, but how to select them for a given real dataset has been the topic of a vast literature (see (Thiel et al., 2006)). The choice of τ and d depend on the specific dataset. For the spiking data from alert monkeys and network simulations, we estimated spike counts in $b = 60ms$ bins, used a delay time $\tau = b$ and an embedding dimension $d = 7$, although we confirmed that reconstruction results held robustly for a wide range of d (not shown). We split the full multi-dimensional time series into two segments – a training period and a test period. Specifically we choose the latter from the end of the sample being studied, and took it to be 1/10 of the full. Given that cross-validation serves as a guarantee against overfitting, it allows us to rely on a simple nearest neighbor algorithm to concretely perform reconstructions.

Given two time series $x(t)$ and $y(t)$, d-dimensional time series of delay vectors $X(t)$ and $Y(t)$ are constructed. To test the accuracy of reconstruction, the data are segmented into a library period and a test period. For each putative downstream vector $X(t)$ in the test sample, a reconstruction $\hat{Y}(t)$ is obtained by listing the k time points $t_j[t]$ ($j = 1, \ldots, k$) corresponding to the delay vectors that are nearest neighbor to $X(t)$ according to the euclidean distance $\Delta_j(t) = ||X(t) - X(t_j[t])||$. For each neighbor, the corresponding weight is computed as a positive, decreasing function of its distance from $X(t)$, namely

Hyperparameters for spiking activity		
Parameter	Description	Value
b	bin size for spike counting	60 ms
α	ratio of recording used as test sample	0.1
τ	delay time (in units of b)	1
d	embedding dimension	7
k	number of nearest neighbors for reconstruction	30

Table 4. Hyperparameters applied to estimate causal flow from resting state spiking activity in awake monkeys and network simulations.

Metadata and settings for stimulation response analysis		
Parameter	Description	Value
$T_{resting}$	duration of resting state recording	10 min
T_{pulse}	duration of stimulus	$\sim$ 120 ms
dT	time step used for response detection	7 ms
Δ_{pre}	time cushion before onset for extracting pre-pulse distributions	10 ms
Δ_{post}	time cushion after onset for extracting post-pulse distributions	4 ms
T_{max}	maximal time lapse considered after stimulus	500 ms

Table 5. Metadata and settings for response analysis.

$w_j(t) = f_j(\Delta_1(t), \ldots, \Delta_k(t))$, normalized to yield the reconstruction

$$\hat{Y}(t) = \sum_{j=1}^{k} w_j(t) Y_{t_j[t]} \Big/ \sum_{j=1}^{k} w_j(t)$$

In the limit of infinitely long datasets, any finite k and any function f will yield asymptotically the same reconstruction. For a finite dataset, a common choice which we adopted is a uniform weight associated to the "simplex dimension" $k = d+1$ (Sugihara and May, 1990). The major computational bottleneck lies in the extraction of the nearest neighbors. We used a ball tree data structure, which partitions data in a series of nesting hyper-spheres as suitable to the structure of the training data. Reconstruction for a single recording took a matter of minutes on a regular laptop.

Once reconstructions are obtained, their accuracy is estimated through the linear correlation coefficient between the test-period time series $Y(t)$ and its reconstruction $\hat{Y}(t)$. In the noiseless infinite-data limit such correlation saturates to one if a causal flow exists in the direction opposite to reconstruction. With finite sample size, we adopted significance of the correlation coefficient as a sufficient condition for various causal scenarios – unidirectional flow in either direction, recurrence, and independence (Fig 2). Notice that the chosen measurement of reconstruction accuracy is a random variable over the space of system trajectories but not necessarily a normal one – being a correlation coefficient it is in fact bounded between 0 and 1. This poses statistical problems to comparing different reconstruction coefficients, and in particular coefficients in two converse directions ((Deyle et al., 2013, Wang et al., 2014)). Crucially, for this reason why define the FCF as the Fisher z-transform of the reconstruction coefficient, which can be relied upon to have near-normal distribution (Fisher, 1925).

4.2.1 Data preprocessing

While the creation of delay vectors directly from spikes has been explored in the literature (Sauer, 1995), we found it convenient to use spike count as continuous variables from whose time series to build delay vectors. Smoothing the spike counts is a delicate step that can introduce extraneous interference even if the linear filter is of a causal type. Specific denoising techniques (Hamilton et al., 2017) have been proposed to circumvent this problem for the preprocessing pipeline of delay-embedding analyses. We found that the least invasive approach was to rely entirely on a nearest neighbor method to perform the denoising.

Since the binning already performs a temporal coarse graining on the information available from recordings, we chose to pick the delay time step equal to one in units of the bin width $b = 60$ms. The bin width and all the other hyperparameters are listed in table 4.

Properties studied by delay embedding are invariant under any differentiable change of variables, and in particular under linear transformations. However, because differences in scale between the activities of individual neurons can affect the calculation of nearest-neighbor weights, the firing rate time series of all channels were shifted and normalized to have zero mean and unit variance, thereby removing concerns about the scale dependence of the exponent in the nearest neighbor weights.

4.3 Significance of causal flow

To establish significance of estimated values of FCF, we adopted an approach to hypothesis-testing now widely used in nonlinear science that consists in generating "surrogate" data, i.e. artificially constructed time series that match the original dataset according to some statistical benchmarks but where the property being tested has been scrambled. The ranking of a discriminating statistic over the distribution of the same quantity calculated on the surrogate allows a significance test on the hypothesis. We chose a surrogate generation method, first proposed in (Thiel et al., 2006), that was designed to preserve all large-scale nonlinear properties of the system. Surrogate time series are produced in three stages. Firstly, we evaluated phase-space distances among Takens states constructed from each time series: nearest neighbors were defined as states within a certain maximum radius from each others. Secondly, an equivalence relationships is defined between states possessing the same set of neighbors (known as "twins"). Finally, surrogate trajectories are initialized randomly and generated by allowing each subsequent step to start with equal probability from the state it just reached or from one of its twins. Each set of twins is thus replaced by a probability superposition of them – macrostates that coarse-grain phase space by making trajectories stochastic. Each surrogate time series emerges thus as the instantiation of a Markov process whose transition matrix has diagonal elements $p_{ii} = (n_i - 1)/n_i$, for n_i equal to the number of twins. This method has the advantage of preserving the temporal statistics of the full system, at the price of introducing a hyperparameter, the neighborhood radius, choosen as the 10% quantile in the nearest-neighbor distances distributions of Takens states as the threshold for neighborhood to base the twinning upon.

The rational for choosing the twin surrogate method described above as opposed to surrogates based on random reshuffling of time points (which destroys all but the amplitude distribution) or isospectral surrogates (based on reshuffling the phases of the Fourier transform and anti-transforming with certain precautions about boundaries, thus preserving autocorrelation) is the following. The latter two methods destroy not just the causal links but rather any nonlinear property of the system (e.g. its density distribution in phase space and its entropy) and therefore pose a considerable risk for false positive. we thus adopted the choice of twin surrogates as a vastly more conservative one which preserves the nonlinear properties of the system while destroying the causal links (for details see (Thiel et al., 2006)).

4.3.1 Estimating perturbation effects

We estimated perturbation effects by comparing network activity in intervals of length T_{max} ending Δ_{pre} before the onset and beginning Δ_{post} after the offset of each perturbation. Such cushion periods reflected the transient pause in recording in those short periods adjacent to each microstimulation (see below). The spiking activity of each efferent neural cluster was estimated in dT bins covering the lapse interval T_{max} and the distribution of spike counts was aggregated across all stimulation trials of the same afferent. The results were largely consistent for a variety of bin sizes and T_{max}; we chose values that maximized our analysis power without obscuring the dynamics of the effects. A Kolmogorov-Smirnov (KS) test between the pre- and post-stimulation aggregated spike count distributions was performed to assess a significant effect of a perturbation, whose effects were reported in the form of KS statistics and p-value (Fig. 6 and Fig. 7). All parameters are reported in Table 5.

4.4 Experimental data

We recorded and perturbed the activity of neurons in the pre-arcuate gyrus (area 8Ar) of macaque monkeys (macaca mulatta) using chronically implanted multi-electrode Utah arrays (96 electrodes; Blackrock Microsystems). All experimental procedures conformed to the National Institutes of Health *Guide for the Care and Use of Laboratory Animals* and were approved by the New York University Animal Welfare Committee.

During the experiments, monkeys sat in a primate chair, with their heads fixed using a titanium head post. The room was mildly lit and quiet. Monkeys did not perform any task nor receive reward. We monitored the monkey's eye and limb movements using infra-red camera systems (Eyelink for eye tracking, 1 KHz sampling rate). The monkey remained awake (open eyes) and rarely moved limbs during these resting blocks, which were typically 10-20 min long. We started a session with a resting state recording block that we used to quantify functional causal flows and predict the perturbation effects. This block was followed with a recording and microstimulation block that we used for testing the predictions.

The electrodes of the Utah array were 1mm long with 400 μm spacing between adjacent electrodes, permitting simultaneous recordings from neighboring columns in a 4 mm × 4 mm region of cortex. Raw voltage signals were filtered and thresholded in real time to identify spikes. Spike waveforms and raw voltage were saved at 30 KHz sampling frequency for offline processing.

Electrical microstimulation was delivered through individual electrodes of the Utah array. Microstimulation pulse trains consisted of low current (15 μA) biphasic pulses (Afraz et al., 2006, Fetsch et al., 2014, Salzman et al., 1992), each 0.2 ms long, delivered at 200 Hz. Pulse trains were 120 ms long and occurred once in any 5 s period. The exact time of the microstimulation with the 5s periods varied randomly. Electrophysiological recording was done in between the microstimulation trains; it resumed with a short latency (<5 ms) at the end of each pulse train and continued until the beginning of the subsequent train. Microstimulation of the prearcuate gyrus with currents >50 μA could trigger saccadic eye movements (Bruce et al., 1985). Our low current microstimulation was chosen well below this motor threshold and never triggered saccades in our experiments.

4.5 Spatial dependence of FCF and perturbation

Both the FCF and the perturbation effects decayed with the distance from the afferent electrode. In order to control for this effect, we performed a partial correlation analysis by detrending the spatial dependence of FCF and the perturbation effects (KS) using a linear regression, and then reevaluating the Pearson correlation between the their residuals (Fig. S3). In particular, fixing an afferent j and given three measurements for each efferent i, namely FCF_i (functional causal flow), KS_i (interventional connectivity, and d_i (physical distance between electrodes i and j) we fit two linear regression models:

$$FCF_i = \beta_0^{FCF} + \beta_1^{FCF} d_i + \epsilon_i^{FCF}$$
$$KS_i = \beta_0^{KS} + \beta_1^{KS} d_i + \epsilon_i^{KS}$$

The partial correlation between FCF and KS controlling for physical distance is then given by: $\rho(\epsilon^{FCF}, \epsilon^{KS})$. This quantity allows us to measure the unique contribution of FCF in predicting KS by removing their linear spatial dependence. We further repeated this experiment nonparametrically by subtracting the median curve shown in Fig. 7 and computing the correlation between the residuals. These results confirm our observation that the relationship between FCF and KS is not simply a confound of the physical distance between the electrodes.

5 Acknowledgments

The authors would like to thank Rainer Engelken and Ramin Khajeh for advice on the simulations and Tim Gardner for discussions. AN was supported by NSF DBI-1707398 and by the Gatsby Charitable Foundation grant GAT3708. RK was supported by Simons Collaboration on the Global Brain (grant 542997), McKnight Scholar Award, Pew Scholarship in the Biomedical Sciences, and National Institute of Mental Health (R01 MH109180).

LM was supported by National Institute of Neurological Disorders and Stroke grant R01-NS118461 (BRAIN Initiative) and by National Institute of Deafness and Communication Disorders grant K25-DC013557.

References

S.-R. Afraz, R. Kiani, and H. Esteky. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–695, 2006.

D. J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex*, 7(3):237–52, 1997. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9143444.

A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283):1868–71, 1996. ISSN 0036-8075 (Print) 0036-8075 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/8791593.

A. M. Bastos and J.-M. Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.*, 9:175, 2015.

P. Berkes, G. Orban, M. Lengyel, and J. Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–7, 2011.

T. Binzegger, R. J. Douglas, and K. A. Martin. A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24(39):8441–8453, 2004.

V. Braitenberg and A. Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media, 2013.

S. L. Bressler and A. K. Seth. Wiener-Granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, Sept. 2011.

B. A. W. Brinkman, F. Rieke, E. Shea-Brown, and M. A. Buice. Predicting how and when hidden neurons skew measured synaptic interactions. *PLoS Comput. Biol.*, 14(10):e1006490, Oct. 2018.

C. J. Bruce, M. E. Goldberg, M. C. Bushnell, and G. B. Stanton. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J Neurophysiol*, 54(3):714–734, Sep 1985.

M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98, 1991.

S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods, 2009.

M. R. Cohen and A. Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811, 2011.

B. Cummins, T. Gedeon, and K. Spendlove. On the efficacy of state space reconstruction methods in determining causality. *SIAM Journal on Applied Dynamical Systems*, 14(1):335–381, 2015.

M. C. Dadarlat and M. P. Stryker. Locomotion enhances neural encoding of visual stimuli in mouse v1. *Journal of Neuroscience*, 37(14):3764–3775, 2017.

A. Das and I. R. Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, pages 1–11, 2020.

G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS Comput. Biol.*, 8(3):e1002395, Mar. 2012.

E. R. Deyle, M. Fogarty, C.-h. Hsieh, L. Kaufman, A. D. MacCall, S. B. Munch, C. T. Perretti, H. Ye, and G. Sugihara. Predicting climate effects on pacific sardine. *Proceedings of the National Academy of Sciences*, 110(16):6430–6435, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1215506110. URL https://www.pnas.org/content/110/16/6430.

M. Dhamala, G. Rangarajan, and M. Ding. Estimating granger causality from fourier and wavelet transforms of time series data, 2008.

J. Ditterich, M. E. Mazurek, and M. N. Shadlen. Microstimulation of visual cortex affects the speed of perceptual decisions. *Nat Neurosci*, 6(8):891–898, Aug 2003.

T. A. Engel, N. A. Steinmetz, M. A. Gieselmann, A. Thiele, T. Moore, and K. Boahen. Selective modulation of cortical state during spatial attention. *Science*, 354(6316):1140–1144, 2016.

D. F. English, S. McKenzie, T. Evans, K. Kim, E. Yoon, and G. Buzsáki. Pyramidal cell-interneuron circuit architecture and dynamics in hippocampal networks. *Neuron*, 96(2): 505–520, 2017.

L. Faes, G. Nollo, and A. Porta. Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique, 2011.

D. J. Felleman and D. E. Van. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

C. R. Fetsch, R. Kiani, W. T. Newsome, and M. N. Shadlen. Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron*, 84(1):239, Oct 2014.

J. Fiser, C. Chiu, and M. Weliky. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, 431(7008):573–8, 2004. ISSN 0028-0836. doi: 10.1038/nature02907.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

R. S. Fisher and A. L. Velasco. Electrical brain stimulation for epilepsy. *Nature Reviews Neurology*, 10(5):261–270, 2014.

A. Fontanini and D. B. Katz. Behavioral states, network states, and sensory response variability. *Journal of neurophysiology*, 100(3):1160–1168, 2008.

D. Gervasoni, S.-C. Lin, S. Ribeiro, E. S. Soares, J. Pantoja, and M. A. Nicolelis. Global forebrain dynamics predict rat behavioral states and their transitions. *Journal of Neuroscience*, 24(49): 11137–11147, 2004.

J. Geweke. Measurement of linear dependence and feedback between multiple time series, 1982.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods, 1969.

F. Hamilton, T. Berry, and T. Sauer. Kalman-takens filtering in the presence of dynamical noise. *The European Physical Journal Special Topics*, 226(15):3239–3250, 2017.

J. A. Harris, S. Mihalas, K. E. Hirokawa, J. D. Whitesell, H. Choi, A. Bernard, P. Bohn, S. Caldejon, L. Casal, A. Cho, A. Feiner, D. Feng, N. Gaudreault, C. R. Gerfen, N. Graddis, P. A. Groblewski, A. M. Henry, A. Ho, R. Howard, J. E. Knox, L. Kuan, X. Kuang, J. Lecoq, P. Lesnar, Y. Li, J. Luviano, S. McConoughey, M. T. Mortrud, M. Naeemi, L. Ng, S. W. Oh, B. Ouellette, E. Shen, S. A. Sorensen, W. Wakeman, Q. Wang, Y. Wang, A. Williford, J. W.

Phillips, A. R. Jones, C. Koch, and H. Zeng. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202, Nov. 2019.

P. Heggelund and K. Albus. Response variability and orientation discrimination of single cells in striate cortex of cat. *Experimental Brain Research*, 32(2):197–211, 1978.

M. H. Histed, A. M. Ni, and J. H. Maunsell. Insights into cortical mechanisms of behavior from microstimulation experiments. *Prog Neurobiol*, 103:115–130, Apr 2013.

C. Huang, D. A. Ruff, R. Pyle, R. Rosenbaum, M. R. Cohen, and B. Doiron. Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348, 2019.

L. M. Jones, A. Fontanini, B. F. Sadacca, P. Miller, and D. B. Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc Natl Acad Sci U S A*, 104(47): 18772–7, 2007.

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

T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–6, 2003.

R. Kiani, C. J. Cueva, J. B. Reppas, and W. T. Newsome. Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Curr Biol*, 24(13): 1542–1547, Jul 2014.

R. Kiani, C. J. Cueva, J. B. Reppas, D. Peixoto, S. I. Ryu, and W. T. Newsome. Natural grouping of neural responses reveals spatially segregated clusters in prearcuate cortex. *Neuron*, 85(6):1359–1373, 2015.

W.-C. A. Lee, V. Bonin, M. Reed, B. J. Graham, G. Hood, K. Glattfelder, and R. Clay Reid. Anatomy and function of an excitatory network in the visual cortex, 2016.

S. Lefort, C. Tomm, J.-C. F. Sarria, and C. C. Petersen. The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron*, 61(2):301–316, 2009.

A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat Neurosci*, 15(11):1498–505, 2012. ISSN 1546-1726 (Electronic) 1097-6256 (Linking). doi: 10.1038/nn.3220. URL http://www.ncbi.nlm.nih.gov/pubmed/23001062.

A. Luczak, P. Bartho, S. L. Marguet, G. Buzsaki, and K. D. Harris. Sequential structure of neocortical spontaneous activity in vivo. *Proc Natl Acad Sci U S A*, 104(1):347–52, 2007. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.0605643104. URL http://www.ncbi.nlm.nih.gov/pubmed/17185420.

A. Luczak, P. Bartho, and K. D. Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62(3):413–25, 2009.

V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, Nov 2013.

O. Maoz, G. Tkačik, M. S. Esteki, R. Kiani, and E. Schneidman. Learning probabilistic neural representations with randomly connected circuits. *Proceedings of the National Academy of Sciences*, 117(40):25066–25073, 2020.

I. E. Marinescu, P. N. Lawlor, and K. P. Kording. Quasi-experimental causality in neuroscience and behavioural research. *Nature human behaviour*, 2(12):891–898, 2018.

L. Mazzucato, A. Fontanini, and G. La Camera. Dynamics of multistable states during ongoing and evoked cortical activity. *The Journal of Neuroscience*, 35(21):8214–8231, 2015.

L. Mazzucato, A. Fontanini, and G. La Camera. Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience*, 10:11, 2016.

L. Mazzucato, G. La Camera, and A. Fontanini. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nat Neurosci*, 22(5):787–796, 05 2019.

M. J. McGinley, M. Vinck, J. Reimer, R. Batista-Brito, E. Zagha, C. R. Cadwell, A. S. Tolias, J. A. Cardin, and D. A. McCormick. Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015.

S. Moeller, T. Crapse, L. Chang, and D. Y. Tsao. The effect of face patch microstimulation on perception of faces and objects. *Nat Neurosci*, 20(5):743–752, May 2017.

S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019.

M. Okun, N. A. Steinmetz, L. Cossell, M. F. Iacaruso, H. Ko, P. Barthó, T. Moore, S. B. Hofer, T. D. Mrsic-Flogel, M. Carandini, et al. Diverse coupling of neurons to populations in sensory cortex. *Nature*, 521(7553):511–515, 2015.

C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, and D. Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat Methods*, 15(10):805–815, 10 2018.

J. Parvizi, C. Jacques, B. L. Foster, N. Withoft, V. Rangarajan, K. S. Weiner, and K. Grill-Spector. Electrical stimulation of human fusiform face-selective regions distorts face perception. *Journal of Neuroscience*, 32(43):14915–14920, 2012.

R. Perin, T. K. Berger, and H. Markram. A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci. U. S. A.*, 108(13):5419–5424, Mar. 2011.

J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

N. C. Rabinowitz, R. L. Goris, M. Cohen, and E. P. Simoncelli. Attention stabilizes the shared gain of V4 populations. *Elife*, 4:e08998, Nov. 2015.

A. T. Reid, D. B. Headley, R. D. Mill, R. Sanchez-Romero, L. Q. Uddin, D. Marinazzo, D. J. Lurie, P. A. Valdés-Sosa, S. J. Hanson, B. B. Biswal, et al. Advancing functional connectivity research from association to causation. *Nature neuroscience*, 22(11):1751–1760, 2019.

V. Rostami, T. Rost, A. Riehle, S. J. van Albada, and M. P. Nawrot. Spiking neural network model of motor cortex with joint excitatory and inhibitory clusters reflects task uncertainty, reaction times, and variability dynamics. *bioRxiv*, 2020.

D. A. Ruff and M. R. Cohen. Attention can either increase or decrease spike count correlations in visual cortex. *Nature neuroscience*, 17(11):1591–1597, 2014.

D. B. Salkoff, E. Zagha, E. McCarthy, and D. A. McCormick. Movement and performance explain widespread cortical activity in a visual detection task. *Cerebral Cortex*, 30(1):421–437, 2020.

C. D. Salzman, K. H. Britten, and W. T. Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346(6280):174–177, 1990.

C. D. Salzman, C. M. Murasugi, K. H. Britten, and W. T. Newsome. Microstimulation in visual area mt: effects on direction discrimination performance. *Journal of Neuroscience*, 12(6): 2331–2355, 1992.

T. Sauer. Interspike interval embedding of chaotic signals. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):127–132, 1995.

T. Sauer, J. A. Yorke, and M. Casdagli. Embedology, 1991.

M. T. Schaub, Y. N. Billeh, C. A. Anastassiou, C. Koch, and M. Barahona. Emergence of Slow-Switching assemblies in structured neuronal networks. *PLoS Comput. Biol.*, 11(7): e1004196, July 2015.

T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461–464, July 2000.

J. D. Semedo, A. Zandvakili, C. K. Machens, B. M. Yu, and A. Kohn. Cortical areas interact through a communication subspace, 2019.

S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.*, 3(3):e68, Mar. 2005.

J. Stark, D. Broomhead, M. Davies, and J. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods & Applications*, 30(8):5303–5314, 1997.

C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437), 2019.

G. Sugihara and R. M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734–741, 1990.

G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems, 2012.

F. T. Sun, L. M. Miller, and M. D'Esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage*, 21(2):647–658, Feb. 2004.

H. Super, C. van der Togt, H. Spekreijse, and V. A. Lamme. Internal state of monkey primary visual cortex (v1) predicts figure–ground perception. *Journal of Neuroscience*, 23(8):3407–3414, 2003.

S. Tajima, T. Yanagawa, N. Fujii, and T. Toyoizumi. Untangling Brain-Wide dynamics in consciousness by Cross-Embedding. *PLoS Comput. Biol.*, 11(11):e1004537, Nov. 2015.

S. Tajima, T. Mita, D. J. Bakkum, H. Takahashi, and T. Toyoizumi. Locally embedded presages of global network bursts. *Proc. Natl. Acad. Sci. U. S. A.*, 114(36):9517–9522, Sept. 2017.

F. Takens. Detecting strange attractors in turbulence, 1981.

M. Thiel, M. C. Romano, J. Kurths, M. Rolfs, and R. Kliegl. Twin surrogates to test for complex synchronisation. *EPL (Europhysics Letters)*, 75(4):535, 2006.

A. M. Thomson and C. Lamy. Functional maps of neocortical local circuitry. *Frontiers in neuroscience*, 1:2, 2007.

G. Tkacik, E. Schneidman, M. J. Berry II, and W. Bialek. Ising models for networks of real neurons. *arXiv preprint q-bio/0611072*, 2006.

E. M. Trautmann, S. D. Stavisky, S. Lahiri, K. C. Ames, M. T. Kaufman, D. J. O'Shea, S. Vyas, X. Sun, S. I. Ryu, S. Ganguli, and K. V. Shenoy. Accurate Estimation of Neural Population Dynamics without Spike Sorting. *Neuron*, 103(2):292–308, 07 2019.

M. Tsodyks, T. Kenet, A. Grinvald, and A. Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–6, 1999. ISSN 0036-8075 (Print) 0036-8075 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/10583955.

V. A. Vakorin, B. Mišić, O. Krakovska, G. Bezgin, and A. R. McIntosh. Confounding effects of phase delays on causality estimation. *PLoS One*, 8(1):e53588, Jan. 2013.

M. Vidne, Y. Ahmadian, J. Shlens, J. W. Pillow, J. Kulkarni, A. M. Litke, E. J. Chichilnisky, E. Simoncelli, and L. Paninski. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells, 2012.

R. Vogels, W. Spileers, and G. A. Orban. The response variability of striate cortical neurons in the behaving monkey. *Experimental brain research*, 77(2):432–436, 1989.

X. Wang, S. Piao, P. Ciais, P. Friedlingstein, R. B. Myneni, P. Cox, M. Heimann, J. Miller, S. Peng, T. Wang, et al. A two-fold increase of carbon cycle sensitivity to tropical temperature variations. *Nature*, 506(7487):212–215, 2014.

G. Werner and V. B. Mountcastle. The variability of central neural activity in a sensory system, and its implications for the central reflection of sensory events. *Journal of Neurophysiology*, 26 (6):958–977, 1963.

Y. T. Wong, M. M. Fabiszak, Y. Novikov, N. D. Daw, and B. Pesaran. Coherent neuronal ensembles are rapidly recruited when making a look-reach decision, 2016.

D. Wyrick and L. Mazzucato. State-dependent control of cortical processing speed via gain modulation. *bioRxiv*, 2020. doi: 10.1101/2020.04.07.030700.

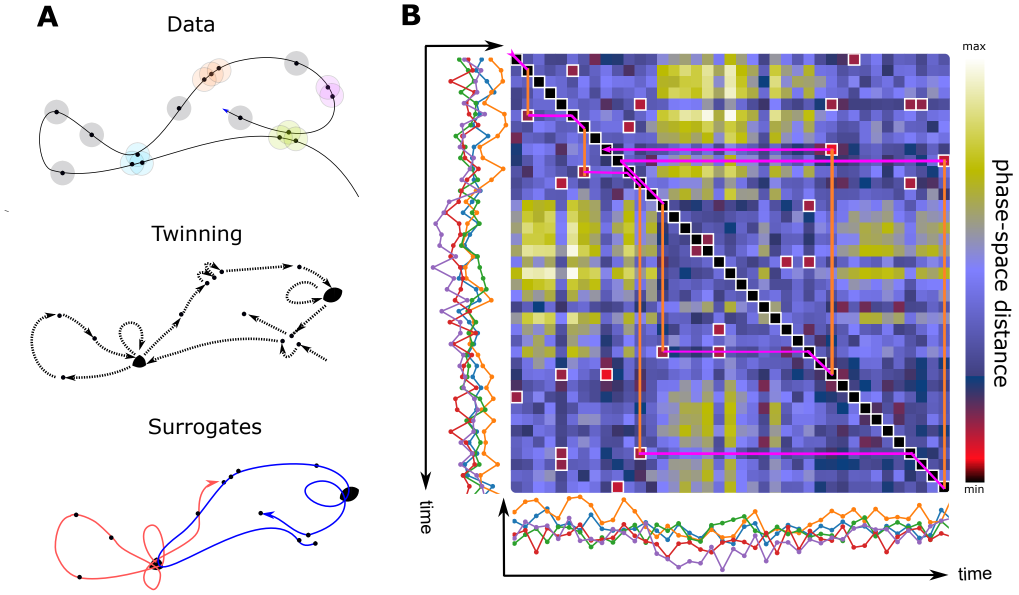

Figure S1. Surrogate data generation. A) Significance of functional causal flow (FCF) is established by comparison to surrogate datasets designed to preserve all large-scale nonlinear properties of the system. Surrogates are produced in three stages: top, phase-space distance is evaluated among Takens states constructed from each time series, and nearest neighbors are identified; center, states in the trajectory are coarse-grained by collapsing states with the same set of neighbors (in the example, the blue and purple clusters merge but the orange and green ones do not); bottom, surrogate trajectories are generated from random initial conditions by regarding twin-sets as retentive states in a Markov process (retention is represented by self-loop and has probability $p = (n-1)/n$ where n is the numebr of twins. B) Example trajectory depicted over the matrix of phase-space distances for the multi dimensional time series shown along both axes. Whereas the main diagonal corresponds to the flow of the recorded time series, at each step the surrogate time series can either move forward as in the recording or depart from the main diagonal (vertical orange lines) to pick one of the low-distance states in its own twin set, and perform from there a forward step mimicking the recording (purple broken lines). Notice that, as in the example, motion can be both forward and backward in time.

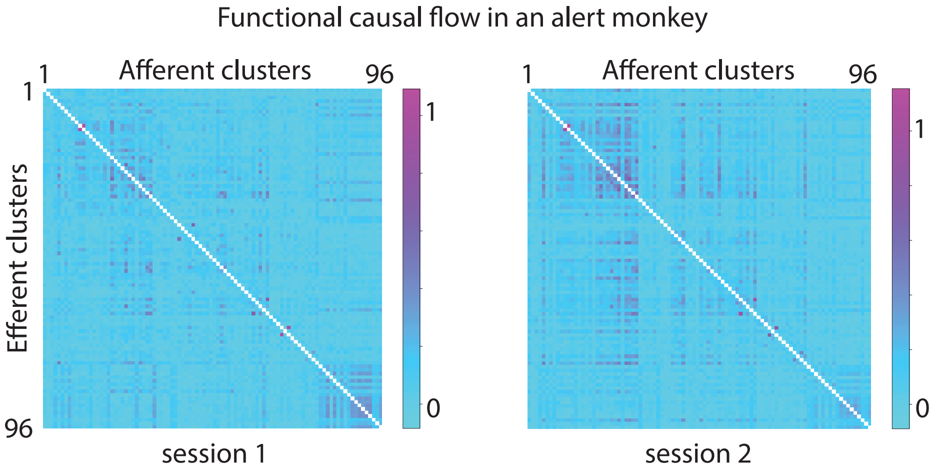

Figure S2. Full resting state FCF. Full resting state FCF matrix inferred from ensemble spiking activity in two sessions from multi-electrode array recordings in the pre-arcuate gyrus during quiet wakefulness (same sessions as in Fig. 7).

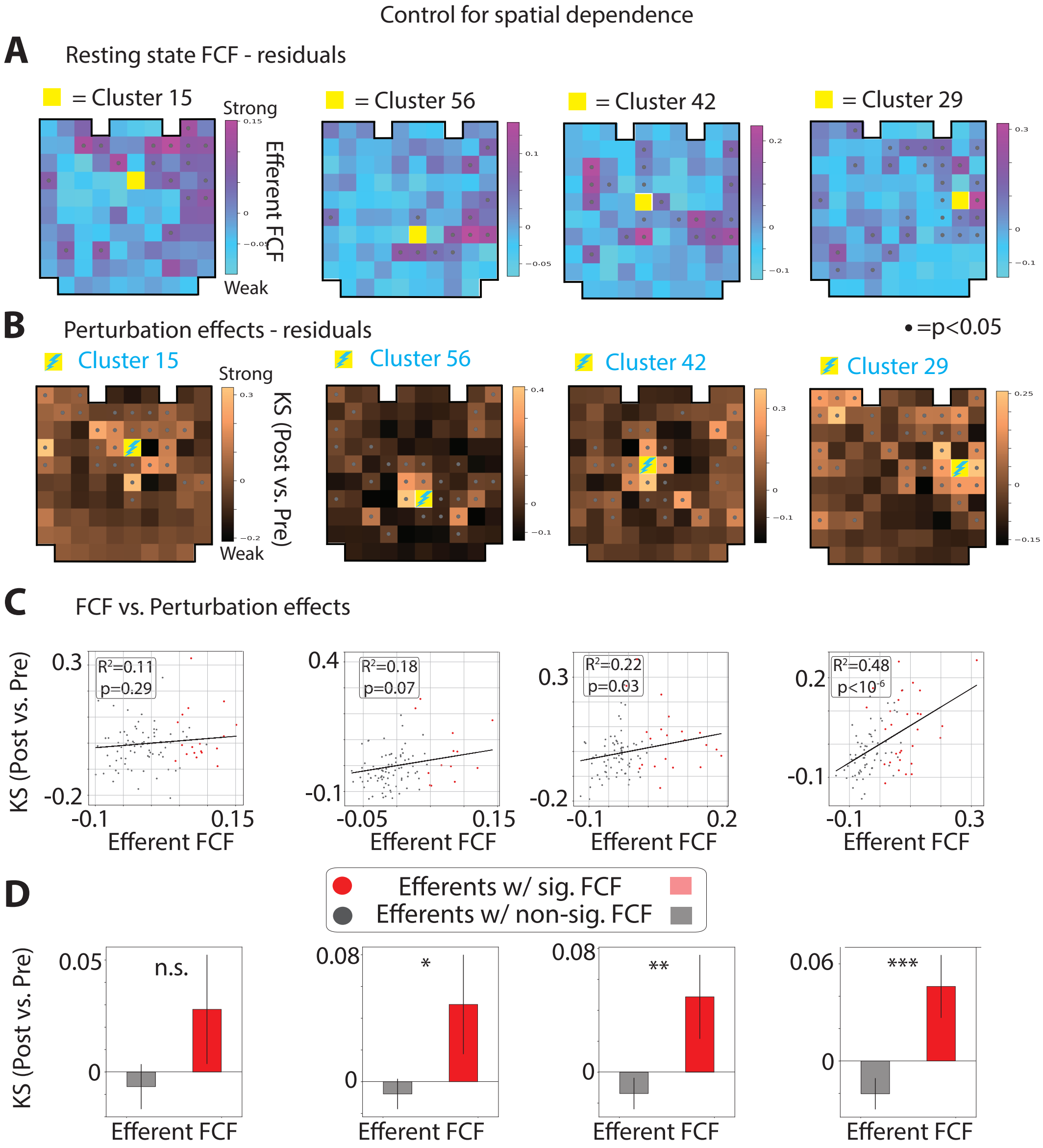

Figure S3. Controlling for spatial dependence of FCF and perturbation effects. The residuals FCF (A) and perturbation (B) effects from Fig. 7A-B after removing the dependence on distance from the afferent electrode. C) Partial correlation between FCF and perturbation effects (KS) after controlling for spatial dependence. D) For each stimulated afferent in panel B, after removing the spatial distance, the residual aggregated perturbation effects are larger over efferents with significant residual resting state FCF vs. efferents with non-significant residual FCF (mean±s.e.m. across gray and red-circled dots from panel E; t-test, $*, **, *** = p < 0.05, 0.01, 0.001$