Firing rate diversity lowers the dimension of population covariability in neuronal networks

Gengshuo John Tian[1-5], Ou Zhu[3,6], Vinay Shirhatti[3,6], Charles M. Greenspon[7], John E. Downey[7], David J. Freedman[2,3,6], Brent Doiron[1-6*]

[1]Department of Statistics, University of Chicago, Chicago, IL, USA.
[2]Grossman Center for Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA.
[3]Neuroscience Institute, University of Chicago, Chicago, IL, USA.
[4]National Institute for Theory and Mathematics in Biology, Chicago, IL, USA.
[5]Center for Living Systems, University of Chicago, Chicago, IL, USA.
[6]Department of Neurobiology, University of Chicago, Chicago, IL, USA.
[7]Department of Organismal Biology & Anatomy, University of Chicago, Chicago, IL, USA.

*Corresponding author(s). E-mail(s): bdoiron@uchicago.edu;

Abstract

Populations of neurons produce activity with two central features. First, neuronal responses are very diverse – specific stimuli or behaviors prompt some neurons to emit many action potentials, while other neurons remain relatively silent. Second, the trial-to-trial fluctuations of neuronal response occupy a low-dimensional space due to correlated activity across the population. We link these two aspects of population response using a randomly coupled recurrent circuit model and derive the following relation: the more diverse the firing rates of neurons in a population, the lower the effective dimension of population trial-to-trial covariability. We tested our prediction using simultaneously recorded neuronal populations from numerous brain areas in mice, non-human primates, and in the motor cortex of human participants. Surprisingly, when populations are restricted to a single brain area our result holds, but when a population is composed from neurons spanning multiple brain areas the relation breaks down. This reflects the fact that the macroscopic connectivity structure at the multi-regional level is significantly more pronounced than the local wiring within a brain area. Finally, using our result we present a theory where a more diverse neuronal code leads to better fine discrimination performance from population activity. In line with this theory, we show that neuronal populations across the brain exhibit both more diverse mean responses and lower-dimensional fluctuations when the brain is in more heightened states of information processing. In sum, we present a key organizational principle of neuronal population response that is widely observed across the nervous system and acts to synergistically improve population representation.

Biological networks are generally complex, with a plethora of components and factors to consider when building a mechanistic understanding of their dynamics [1]. External inputs to a network and the internal interactions of its components can both affect the output behavior. In some cases, complex activity is simply *inherited* from external inputs, and the network interactions are well described by a passive filter. Examples in nature include population dynamics correlated by common environmental fluctuations [2, 3], plankton distributions in oceanic flows [4], and cochlear processing by arrays of unconnected hair cells in the inner ear [5]. Alternatively, complex dynamics can *emerge* from interactions within the system. Canonical cases are the mechanics of visual hallucinations from cortical networks [6, 7], the dynamical patterns in bird flocking [8], and central pattern generators that drive movement [9]. While the organization principles of inheritance and emergence have enjoyed great success, most network activity results from a combination of complex external inputs and internal interactions and should be studied as such. Incorporating this complexity is especially crucial when studying systems like gene networks [10], social networks [11], and quorum sensing

bacterial colonies [12], where the networks function in response to changes in the environment through the interactions of its components. The central goal of this study is to put forth and validate a coherent theory for how the diversity of inputs and the complexity of interactions combine to determine system output. We focus our efforts on the nervous system, yet our approach can be applied in a wide array of disciplines.

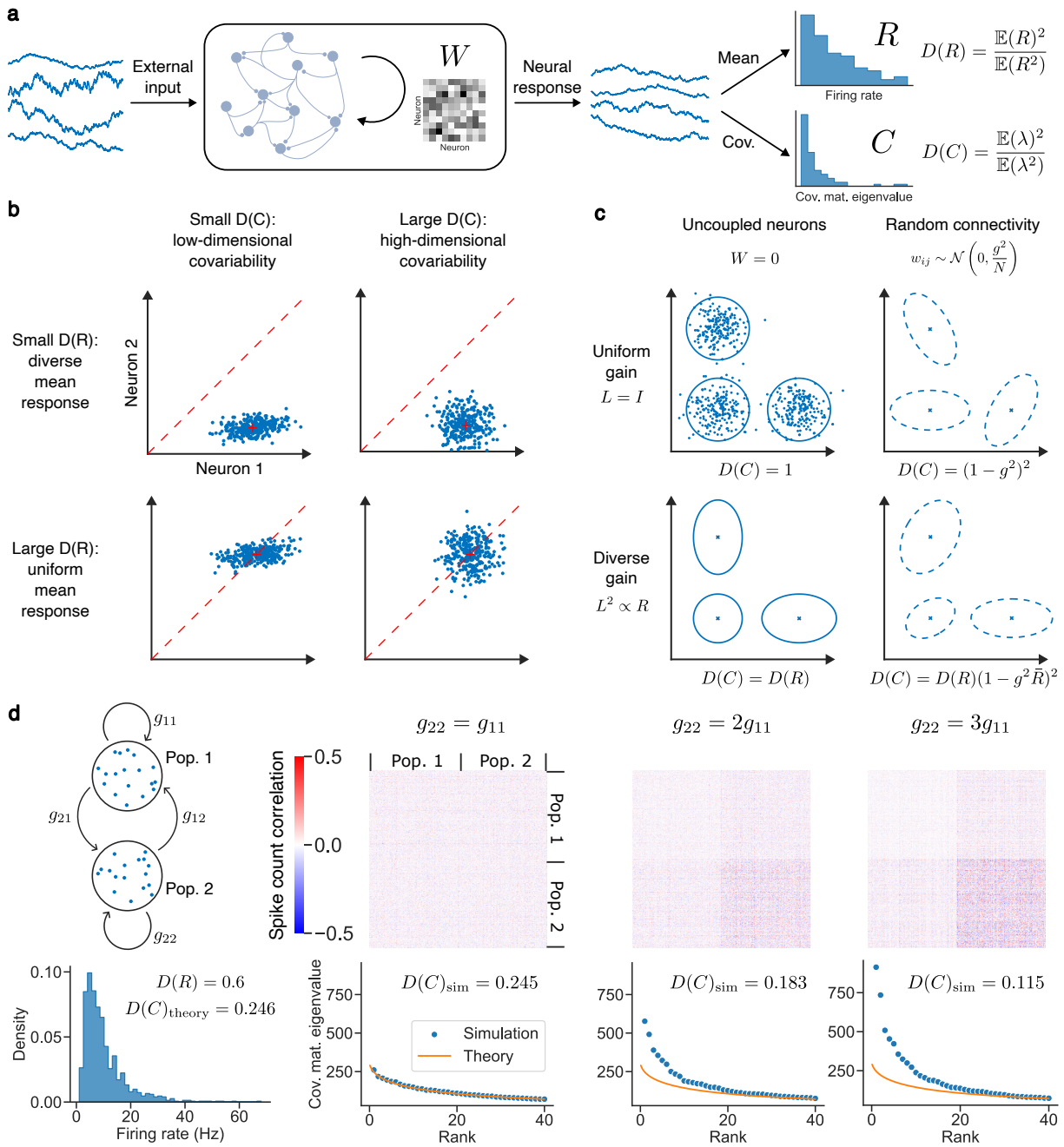

Fig. 1 (**a**) A neuronal circuit processes external input with its recurrent dynamics (where W denotes the recurrent connection weight matrix) to produce neuronal response, which can be characterized by its mean (i.e. the firing rate vector R) and covariance (i.e. the spike-count covariance matrix C, which has eigenvalues $\{\lambda_i\}$). $D(R)$ measures the uniformity of the firing rate distribution across neurons, and $D(C)$ measures the effective dimension of population covariability. (**b**) Illustration of the distribution of neuronal response in the spike count space with small/large $D(R)/D(C)$, respectively. Each dot is the response of a neuron pair for a given realization of input, and the location and shape of the cloud of responses determine $D(R)$ and $D(C)$, respectively. (**c**) Schematic of the expression for the effective dimension $D(C)$ under different model assumptions. The ellipses mark the shape of response cloud, and we consider three different locations to highlight any dependence on neuronal firing rates. (**d**) Block structure in the connectivity matrix W causes deviation from the theory (Eq. 3). Top-left: schematic of the two-population network. Each population has 500 neurons. Connection weights from population β to α is drawn from a Gaussian distribution with mean 0 and variance $g_{\alpha\beta}^2/N$. We set $g_{11} = g_{12} = g_{21}$ and vary g_{22}/g_{11} while keeping the overall variance of the elements of W constant. Top-right: example correlation matrices (with diagonal elements suppressed) under different degrees of cross-population asymmetry. Correlation matrices are plotted instead of covariance matrices for visualization purposes only. Bottom-left: distribution of firing rates across the population, which is the same for all simulations. Bottom-right: eigenvalues gradually deviate from theoretical prediction as the asymmetry between the two populations increases, leading to errors in $D(C)$.

While our goal is simply stated, it becomes especially difficult when the system consists of a large number of highly diverse and interacting components, making the problem irreducibly high-dimensional. This problem is particularly critical in the study of neuronal circuits, in which thousands or even millions of neurons integrate long-range synaptic inputs (feedforward and feedback) through rich recurrent dynamics to produce computationally meaningful responses. To date, while simulations and fitted models have incorporated high degrees of detail [13, 14], theoretical understanding has mostly been based on models that consider only one aspect of diversity in the network [15–19].

In this study we explore how the interplay between diverse external inputs and complex recurrent interactions in a neuronal circuit determine the structure of its outputs. In particular, we examine the geometry of trial-to-trial fluctuations of neuronal response, which plays an important role in neuronal coding [20–22]. We show that in randomly coupled networks that lack significant circuit structure, diversity in both synaptic wiring and firing rates combine multiplicatively to reduce the effective dimension of population covariability. Despite the simplicity of our model framework, our core prediction is verified with electrophysiology datasets from multiple brain areas in monkeys, mice, and humans. However, our theory fails to explain the activity of populations that include neurons spanning multiple brain regions, suggesting that brain-wide circuits have sufficiently rich circuit structure so as to break the assumptions of our theory. One consequence of our single-brain area result is that low-dimensional noise often accompanies a sparse population code, both of which are shown to benefit decoding. We present experimental evidence for this synergistic effect when brain states change due to processing needs, which can be understood using our theoretical framework.

The effective dimension of population covariability in a recurrent circuit model

Consider a network of N neurons with connectivity matrix W receiving noisy and heterogeneous external input (Supplementary Information A.1). Its output is characterized by the mean firing rates R and covariance matrix C of the fluctuating population response (Fig. 1a). While R is a function of the output, we take the firing rates as being determined primarily by their feedforward inputs [23], and as such are a proxy for the input to the network. By contrast, the population covariability is assumed to be dependent on the recurrent interactions among neurons, captured by W [24]. A mathematical statement of how the characteristics of R and W together determine the structure of C is the main goal of this paper.

We quantify the uniformity of the mean response R by the Treves-Rolls measure (denoted by $D(R)$; Fig 1a,b) [25, 26]. A $D(R)$ near 0 marks a diverse mean response distribution where a few neurons fire at much higher rates than the rest, whereas a $D(R)$ near 1 suggests neurons fire at rates similar to one another. We measure the effective dimension of population covariability by the participation ratio (denoted by $D(C)$; Fig 1a,b) [27, 28], where a $D(C)$ near 0 indicates the concentration of fluctuations around a small number of dimensions, whereas a $D(C)$ near 1 implies that fluctuations are distributed more or less uniformly across different directions and are thus high-dimensional. While the functional forms of $D(R)$ and $D(C)$ are identical, they nevertheless measure different aspects of the neuronal response (mean versus variability) and in principle have no immediate relation to one another (Fig. 1b; Methods).

It is a standard result [29, 30] that linearizing the network activity about an operating point gives us the following expression for the covariance matrix (Supplementary Information A.1):

$$C = (L^{-1} - W)^{-1} C_0 (L^{-1} - W)^{-T}, \tag{1}$$

where L is the diagonal matrix of neuronal gains and C_0 is the covariance matrix of the input noise. We assume for now that the input noise has the same magnitude across neurons and are uncorrelated ($C_0 = I$).

In the simplest case where neurons are uncoupled ($W = 0$) and have uniform gain ($L = I$), we have $C = I$ and $D(C) = 1$. This is the largest achievable effective dimension, where joint activity fluctuates equally in all directions (Fig. 1c, top-left). Previous works have studied the case where the coupling is independently drawn from a Gaussian distribution with mean 0 and variance g^2/N but the neurons are otherwise homogeneous ($L = I$). It was derived that $D(C) = (1 - g^2)^2$, suggesting the effective dimension is lower for more diverse coupling as quantified by a larger g [16, 18]. The intuition for the result is that a realization of the random wiring will promote joint activity along certain directions, while suppressing it along other directions (Fig. 1c, top-right). Any deformation from symmetric joint fluctuations acts to lower the dimension of C.

Setting the gain to be uniform ($L = I$) is a drastic oversimplification since neurons are known to exhibit extremely diverse responses, as illustrated by the roughly log-normal distribution of firing rates observed in many studies [31]. When the input-output transfer function of neurons is nonlinear, this diversity of firing rates would imply a diversity of the neuronal gain L. Experiments have shown that the input-output curve

of neurons can be well described by a threshold-power law function [32]:

$$r = f(x) = k[x - x_0]_+^p, \tag{2}$$

where x is the input, $[\cdot]_+ = \max(\cdot, 0)$, x_0 is the firing threshold, k is a constant scaling factor, and p is the power of the transfer function. Following previous work, we set the power p to be 2 [33], which would imply that the neuronal gain is associated with the firing rate by $L^2 \propto R$. If we consider again a population of uncoupled neurons ($W = 0$) but this time with a firing rate distribution of R and with a threshold-quadratic function, we get $D(C) = D(R)$. Thus, a more diverse mean response also results in a lower effective dimension of C (Fig. 1c, bottom-left). Interestingly, this also holds for a population of uncoupled Poisson neurons whose variances are equal to their means.

The above two scenarios each incorporate one aspect of diversity in the network (connectivity versus gain), with each having a critical impact on the effective dimension of response variability. The natural question to ask is then what happens when both of these factors are present in the system and interacting in a nontrivial manner (Fig. 1c, bottom-right), as is the case for real neuronal circuits. Relating the eigenvalue distribution of the covariance matrix C to that of W and R can be studied using techniques from random matrix theory. Critically, classical scalar approaches in probability theory fail when we consider two random matrices (W and R) because of the non-commutative nature of matrices (i.e, $WR \neq RW$). To deal with this issue we use operator-valued free probability theory [34] and derive an equation for the spectrum of C (Supplementary Information A.2, Eq. S8). This further allowed us to calculate the moments of C's eigenvalue distribution, which then yields our central analytic result (Supplementary Information A.3):

$$D(C) = D(R)(1 - g^2\bar{R})^2, \tag{3}$$

where $\bar{R}$ is the population mean firing rate. We see that diversity in the mean response (smaller $D(R)$) and the connectivity (larger g) combine multiplicatively to lower the effective dimension of population covariability (smaller $D(C)$). Intuitively, the heterogeneities in neuronal gains and in synaptic wiring are preserved in the recurrent dynamics and both contribute to the anisotropicity of the population fluctuations. Critically, $D(C)$ only depends on the first two moments of the firing rates R irrespective of its exact distribution. This was verified in simulations using Eq. (1) (Fig. S1).

The derivation of Eq. (3) depends critically on the assumption that the matrices R and W are freely independent (Supplementary Information A.2). This assumption can be broken when there is significant structure in the distribution of W beyond i.i.d. elements [34]. Indeed, real neuronal circuits almost never fit the description of an unstructured graph. For example, one particularly common circuit structure in the brain is neuronal ensembles [35], where neurons are grouped according to their functional coordination, with presumably different connectivity statistics within each ensemble. This produces block structures in the connectivity matrix W. To see how this affects our result, we simulated two interconnected populations of neurons with increasing asymmetry between them (Fig. 1d, Fig. S2a). As the within-ensemble connection statistics become more distinct, the eigenvalues deviate from the theoretical prediction, leading to large errors in the estimation of $D(C)$. Thus, it becomes unclear how well our unstructured circuit model approximates the complex dynamics from real neuronal circuits that likely have structured recurrent wiring. To answer this question, we next turn to neuronal data.

Validation in neuronal data across species and brain areas

It is generally difficult to experimentally measure the connection weights W (and thus g) and the activity of the same neuronal population, making it unfeasible to directly verify Eq. (3). However, Eq. (3) predicts a linear relationship between effective dimension $D(C)$ and firing rate uniformity $D(R)$. Fortunately, both $D(C)$ and $D(R)$ can be calculated from population recordings as long as enough trials are conducted, making it possible to test this novel prediction with neuronal data.

First, we demonstrate this relationship with simulations. We generated covariance matrices using Eq. (1) with the firing rates R drawn from a log-normal distribution $\mathcal{LN}(\mu, \sigma)$. We varied $D(R)$ by changing μ and σ while keeping g and $\bar{R}$ fixed and plotted the resulting $[D(R), D(C)]$ pairs (Fig. 2a; Methods). Indeed, the simulation result follows theoretical predictions closely. Interestingly, this correlation between $D(R)$ and $D(C)$ also breaks down in the two-population model (Fig. S2b), suggesting that the deviation of $D(C)$ from theoretical prediction (Fig. 1d and Fig. S2a) is not due to a simple translation or rescaling, but is more complex in nature.

Another source of structure in neuronal circuits is spatial organization, where neurons that are close to one another are more likely to connect to each other and receive correlated inputs [36, 37], again breaking the free independence of the matrices in Eq. (1). We explored its effect by simulating a network where neurons

are randomly placed in a two-dimensional space and where W and C_0 are generated according to the distance between neurons (Fig. S3; Methods). The correlation between $D(C)$ and $D(R)$ decreases as recurrent connections become more spatially concentrated or the input correlations become more widespread, i.e. as the system moves further away from our assumptions. Together, our analyses of block (Fig. S2b) and spatial connectivity structure (Fig. S3) show that the $[D(R), D(C)]$ relation may break down for the activity of a real neuronal circuit with sufficient structure in its wiring, and is hence not trivially guaranteed to hold in neuronal data.

Next, we trained monkeys to perform a visual categorization task (modified from previous experiments [2]; Fig. 2b; Methods) and recorded spike train data from the lateral intraparietal cortex (LIP). Different experimental conditions set the recorded neuronal populations at different operating points, producing many pairs of mean responses R and covariance matrices C (Fig. 2c; Methods), which mirrors what took place in the simulations (Fig. 2a). Within the time frame of one experimental session, it can be assumed that the connectivity between neurons (and thus the variance parameter of the synaptic weight distribution g) remains largely unchanged. Further, the population mean firing rate $\bar{R}$ does not widely fluctuate across conditions. We can then plot the effective dimension $D(C)$ against firing rate uniformity $D(R)$ across conditions (Fig. 2d) and see that they are indeed linearly related, as predicted by Eq. (3).

We note that the mathematical technique we used to derive the eigenvalue distribution of the covariance matrix C can only address the "bulk" of the distribution (meaning the portion that has positive measure as the system size grows to infinity). Finite-rank perturbations to either the connectivity matrix W or the input noise covariance matrix C_0 would result in a finite number of outliers in C's eigenvalues, which is beyond the scope of the theory [16]. Since these perturbations originate from factors that cannot be measured experimentally, we need to exclude them in order to test our theory. We achieve this by fitting our model to the data to obtain a reference model of C's eigenvalue distribution. We can then statistically test which eigenvalues cannot be explained by this reference model, and identify them as outliers that likely originate from low-rank perturbations to the system (Fig. 2d; Methods). Interestingly, as we shall see later, in many datasets the relationship between $D(C)$ and $D(R)$ persists even without removing the outliers (Fig. S7).

One might be tempted to use the slope of the linear fit to the $[D(R), D(C)]$ scatter plot to infer the variance parameter g of the recorded circuit's synaptic connections. We would caution against this because the recorded neurons constitute only a small sample of the underlying population. It can be shown (Supplementary Information B.2) that under this spatial sub-sampling, $D(C)$ and $D(R)$ are still linearly related (Eq. (S12), Fig. S4), but the coefficient of this proportionality now depends on the activity and connectivity of the unobserved population and can vary considerably across sessions that record from different subpopulations, making it hard to interpret the slope quantitatively. Previous work has also shown that the proportion of excitatory and inhibitory neurons sampled in the recording has significant impact on the population activity structure including its dimensionality [38].

Another confounding factor is that the actual input-output transfer functions in the recorded population is unknown, and may even be heterogeneous across neurons. To probe how it affects our theory, we reexamine different values of the power p in Eq. (2) (Eq. (3) was developed under $p = 2$). We have analytic formulae for general p's (Eq. (S9)), where $D(C)$ and $D(R)$ are again positively related, but the relation is no longer linear (Fig. S5a). Simulations show that $D(C)$ and $D(R)$ are positively correlated as long as the gain depends on the firing rate ($p \neq 1$) (Fig. S5b). We also explored the effect of having a distribution of p's across neurons and found that this additional source of diversity lowers $D(C)$ (Fig. S5c), but again preserves the positive correlation between $D(C)$ and $D(R)$ (Fig. S5d). Taken together, we see that while the $[D(R), D(C)]$ relationship is largely intact, the slope of a linear fit would clearly depend on the distribution of p in the population, again preventing us from estimating the connectivity parameter g from data.

Since our theory was developed for a generic recurrent circuit, it could in principle be applied more generally across diverse brain areas and species. We tested Eq. (3) in various cortical and subcortical brain areas across monkeys (Fig. 3a-c), mice (Fig. 3f-h), and human participants (Fig. 3k, Fig. S6) using novel data we collected (monkey areas LIP, FEF, and SC, and human M1; see Methods) as well as previously published datasets [39–43]. The positive correlation between $D(C)$ and $D(R)$ overwhelmingly holds across all the data we examined when compared to shuffled data (Fig. 3d, i). At the same time, the wide distribution of correlation values may reflect the moderate levels of structure present in local circuits (compare with Fig. S3, right).

As mentioned earlier, even when we do not remove the outlier eigenvalues when calculating $D(C)$, the correlation between $D(C)$ and $D(R)$ often (but not always) remains high (Fig. S7), depending on whether these outliers are large enough and are correlated with $D(R)$ in a way that negates the change in the bulk. Another notable observation is that, when we plot the $D(C)$ and $D(R)$ averaged across conditions for each brain area, they are still strongly correlated (Fig. 3e, j), even though each area presumably has very different connectivity structures.

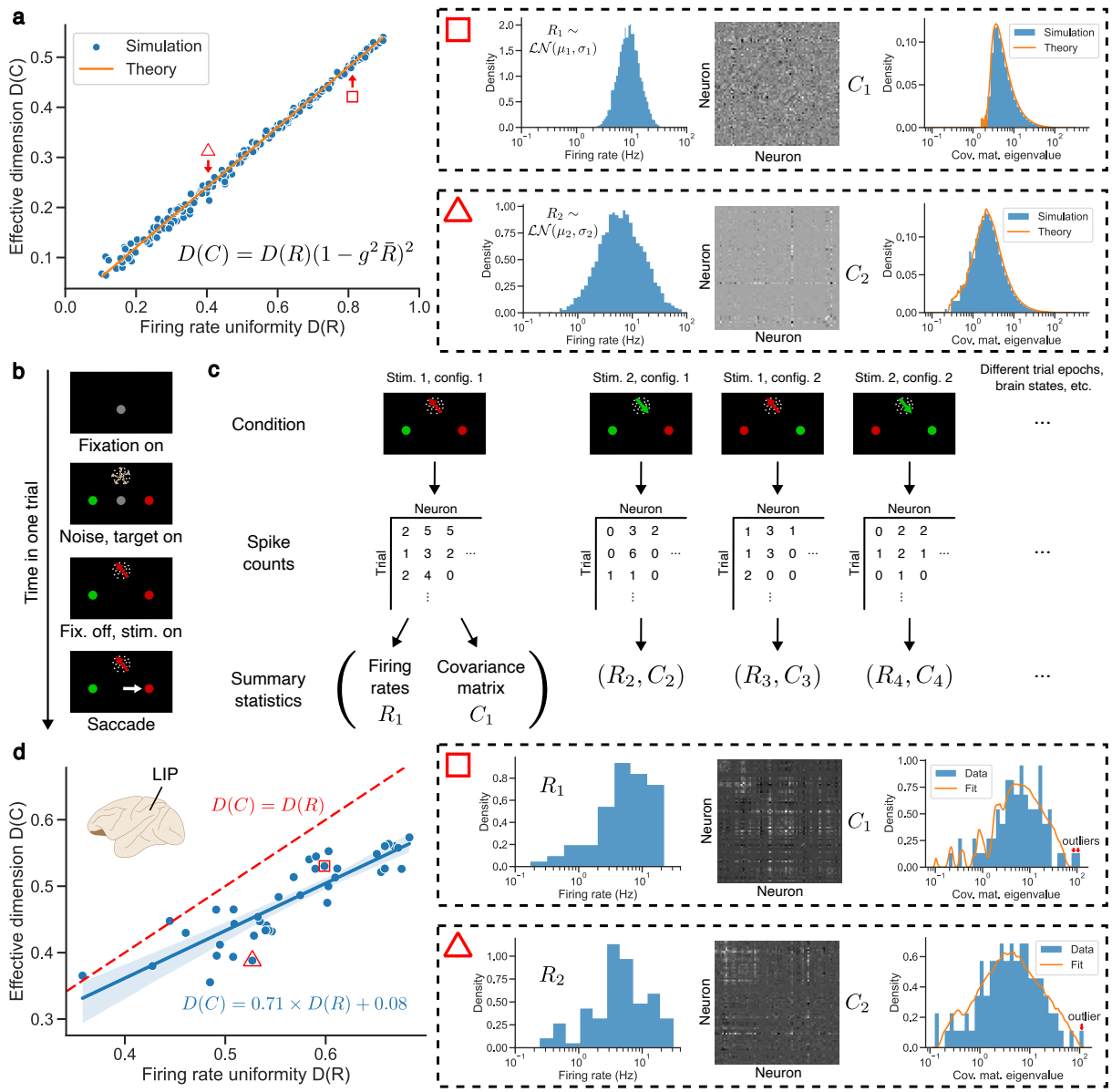

Fig. 2 (**a**) Theoretical prediction of the effective dimension $D(C)$ as firing rate uniformity $D(R)$ changes (Eq. (3)) matches simulation results. The population size is $N = 5000$. The boxes on the right demonstrate how a point in the plot is produced. Each point in the left plot has a different firing rate distribution R drawn from a log-normal distribution with parameters designed to keep the mean firing rate the same but covers a range of $D(R)$ values. The sampled R vector is used to generate the covariance matrix C according to Eq. (1) while keeping the connectivity W unchanged. It can be directly observed that the triangle point (bottom right) with smaller $D(R)$ and $D(C)$ has wider firing rate distribution and covariance matrix eigenvalue distribution than that from the square point (top right). The rightmost column additionally shows that the eigenvalue distribution predicted by our theory (Eq. (S8)) matches the simulation. (**b**) Structure of the visual categorization task (see Methods for details). (**c**) Different experimental conditions set the system at different operating points with different firing rates R and covariance matrices C, which are calculated from the spike count data matrix produced by each condition. Different conditions here could mean different experimental variables such as stimulus parameters (here the direction of moving dots) and response target configurations (here whether to saccade to the left or right target for the red category). They could also mean different epochs in the task as shown in (b) or different brain states such as spontaneous versus evoked. (**d**) The linear relationship between $D(C)$ and $D(R)$ verified with monkey LIP data. Linear regression: slope=0.71, intercept=0.08, r=0.87. The number of recorded neurons is $N = 88$. The red dashed line is the identity line $D(C) = D(R)$. Similar to (a), the boxes on the right illustrate how each point in the plot on the left corresponds to a different (R, C) pair as conditions vary. The rightmost column demonstrates how fitting our model to the data identifies outlier eigenvalues.

One might argue that the positive correlation between $D(R)$ and $D(C)$ is expected if we simply consider the network activity to be a perturbation from a population of uncoupled and uncorrelated Poisson-like neurons with a uniform Fano factor (where $D(C) = D(R)$ as in Fig. 1c, bottom left). To refute this argument we provide four lines of evidence. First, the distribution of pairwise noise correlations suggests neurons are far from uncorrelated (Fig. S8a). Second, if a network was a small perturbation from a set of uncoupled neurons, then its output covariance matrix C would be a perturbation of a diagonal matrix, and each eigenvector would be dominated by one neuron. This is not the case in our neuronal data (Fig. S8b). Third, uncoupled Poisson-like neurons would produce $D(C) = D(R)$, but we almost always observe $D(C)$

values that are smaller than their corresponding $D(R)$ values (Fig. 3l). In line with this, the variations of $D(C)$ across conditions are also smaller than those of $D(R)$ as indicated by the slope of the $[D(R), D(C)]$ plot being smaller than 1 (Fig. 3m). Finally, our theory requires that $0 \leq \hat{g} < 1$, where $\hat{g} = g\sqrt{\overline{R}}$ is the normalized connectivity diversity parameter (Supplementary Information A). A perturbation from an uncoupled network would have $\hat{g}$ near zero. However, the $\hat{g}$ fitted from data is typically larger than 0.5 (Fig. 3n), suggesting that recurrent connections play an important role in generating neuronal variability. Indeed, our model always explains the eigenvalue distribution in the data much better than an uncoupled model (Fig. S8c-d).

In sum, these results show that the $[D(R), D(C)]$ relationship predicted by Eq. (3) largely holds for neuronal populations across the brain, suggesting that the structure in their recurrent wiring W is typically not prominent enough to break our theoretical prediction. However, these analyses were all restricted to data from single brain regions. It is unclear whether this relationship still holds when neurons from multiple brain regions are put together, where the (block) structure in W is presumably more pronounced. We examine this in the next section.

Combining multiple brain regions breaks the relationship between $D(R)$ and $D(C)$

Multi-regional brain connectomes show a clear block structure, with distinct within-area and between-area connectivity [45]. To investigate how such diverse interareal connectivity structures affect the $[D(R), D(C)]$ relation, we analyzed data recorded simultaneously from monkey V4 and PFC [40] as well as from mouse VISp and VISl [42]. When we pool neurons from two brain regions into one large population, the $[D(R), D(C)]$ correlation that is present in each area disappears (Fig. 4, compare orange and green violin plots), consistent with what was seen in simulations (Fig. S2b, green asterisk). One potential cause of this result is the single-neuron differences (in firing rates and variances) between the two areas overshadowing the changes of activity across experimental conditions and obscuring the correlation between $D(R)$ and $D(C)$. To test this, we recalculated $D(C)$ using only the variances (or equivalently, setting the off-diagonal elements of the covariance matrices to 0). In this case, the $[D(R), D(C)]$ relation persists (Fig. 4, red violin plots), suggesting that the observed breakdown of our theory is at least partly due to the correlation structure in the multi-area population. This structure likely stems from the interaction between brain regions significantly violating the free independence assumption we used to derive Eq. (3) (as in Fig. 1d). Together, the above results demonstrate that our model is most readily applicable at the local circuit level, but additional considerations are needed when there are significant macroscopic structures present in the network such as connectivity differences between brain areas.

Diverse mean response works synergistically with low-dimensional variability to improve neuronal coding

The distribution of trial-to-trial fluctuations is critical to consider when measuring the performance of a population code for an input stimulus [20–22]. Consider the neuronal responses to two stimuli with means R_1, R_2 and noise covariances C_1, C_2. To impose stability on the network dynamics and rule out runaway activity, we consider the mean responses to have a fixed population average. This restricts R_1 and R_2 to lie on a simplex in the neuronal response space (Fig. 5a, yellow region). We would like to understand how mean response uniformity $D(R)$ affects the neuronal code. Within the stability simplex, the mean responses with the same $D(R)$ are defined by the intersection of the simplex with a sphere with a certain radius (Fig. 5a, magenta region), forming a codimension 2 (i.e. $(N-2)$-dimensional) manifold S in neuronal space (Fig. 5a, blue region). As $D(R)$ gets smaller, the size of S gets larger so that there is more space to distribute the responses to different stimuli. At the same time, Eq. (3) suggests that the effective dimension of the population variability in both response distributions decreases, making it less likely for the trial-to-trial fluctuations to interfere with the coding dimension when the overall amount of variability is constant (compare Fig. 5b1 and b2). In sum, a diverse mean response (as measured by a small $D(R)$) works synergistically with low-dimensional variability (as measured by a small $D(C)$) through Eq. (3) to improve the population code.

The performance of a downstream decoder can be quantified by the (linear) discriminability index (d'). We calculated d' as a function of $D(R)$ when the covariance matrix of trial-to-trial fluctuations is given by Eq. (1) (Supplementary Information C). Indeed, d' decreases substantially as $D(R)$ becomes larger (Eq. (S14), Fig. 5c). To verify this, we estimated the d' of neuronal activity in mouse primary visual area (VISp) in response to pairs of drifting grating stimuli (Methods) and examined how the value of d' for each stimulus

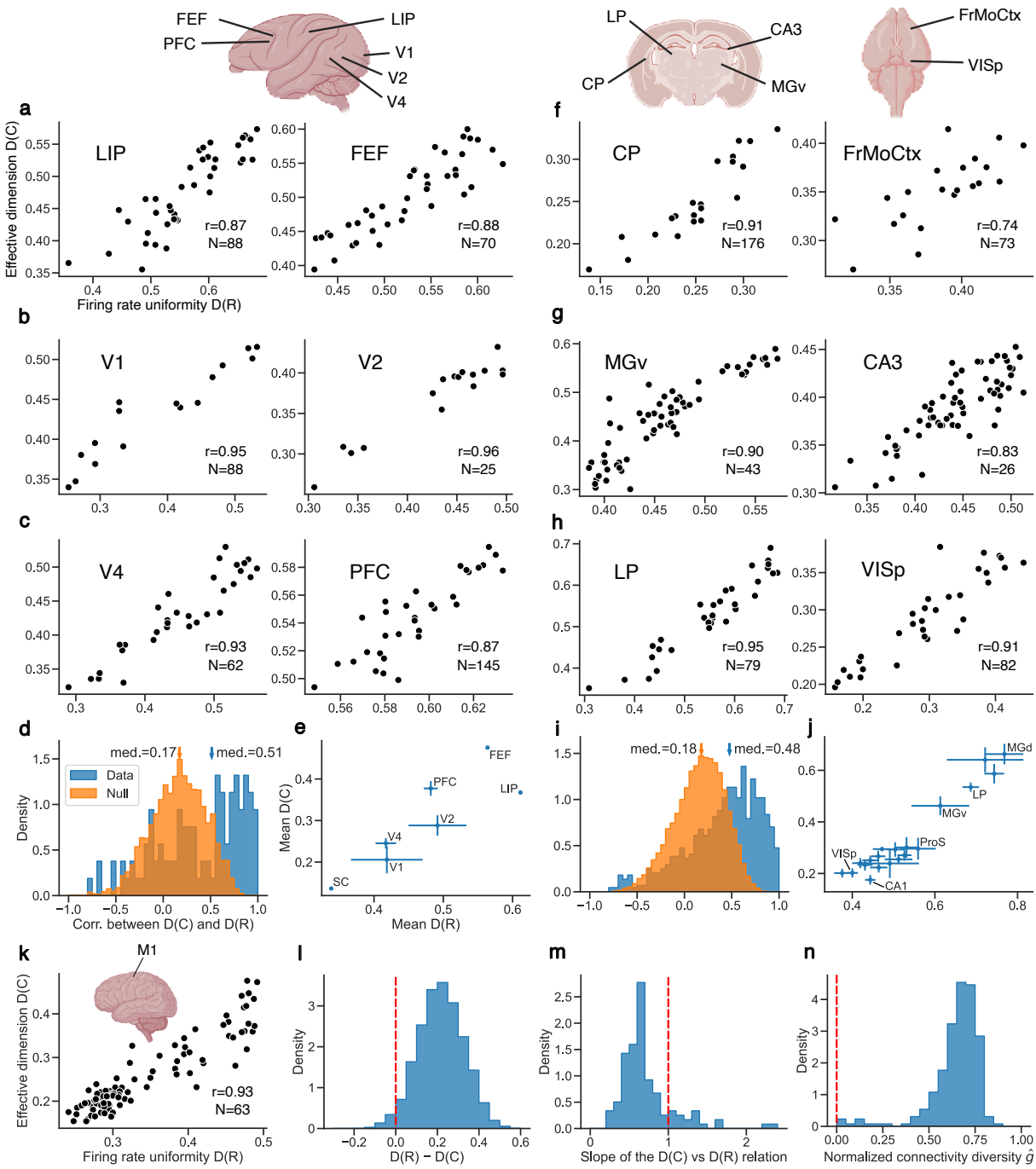

Fig. 3 The relationship between $D(C)$ and $D(R)$ verified in electrophysiology datasets from multiple brain areas in monkey (a-e), mice (f-j), and humans (k). Each scatter plot (a-c, f-h, k) shows data from one recorded population in a single session across a range of conditions as in Fig. 2d. r is the Pearson correlation coefficient between $D(C)$ and $D(R)$, and N is the number of recorded neurons. (**a**) Data recorded from monkey LIP (lateral intraparietal) and FEF (frontal eye field) cortices (see Methods). (**b**) Monkey visual areas V1 and V2 data taken from [39]. (**c**) Monkey visual area V4 and prefrontal cortex (PFC) data taken from [40]. (**d**) Aggregate r values from all monkey datasets in (a-c) compared to shuffled data. 7 brain areas and 32 sessions in total. The data distribution is significantly larger than the null (median test, $p < 0.0001$). (**e**) Mean $D(C)$ and $D(R)$ across conditions for each brain area in the monkey datasets. Error bars are standard errors of the mean across sessions. To show the original geometry of the neuronal response, outlier eigenvalues were not removed here. (**f**) Mouse caudoputamen (CP) and frontal motor cortex (FrMoCtx) data taken from [41]. (**g**) Mouse medial geniculate nucleus, ventral part (MGv) and CA3 data taken from [42, 44]. (**h**) Mouse lateral posterior nucleus of the thalamus (LP) and primary visual area (VISp) data taken from [43]. (**i**) Same as (d) but for mouse datasets in (f-h). 40 brain areas and 79 sessions in total. The data distribution is significantly larger than the null (median test, $p < 0.0001$). (**j**) Same as (e), but for the mouse dataset [42]. (**k**) Data recorded from human M1. (**l-n**) Recurrent connectivity has a large impact on $D(C)$. The histograms are over all brain areas and all sessions in dataset [42]. The red dashed lines are what would be expected for independent Poisson neurons, which are far away from the data distributions in blue. (**l**) Distribution of $D(R) - D(C)$. (**m**) Distribution of the slope obtained by performing linear regression on $D(C)$ against $D(R)$. Only data where the correlation between $D(C)$ and $D(R)$ are greater than 0.5 is included because the slope is not very informative otherwise. (**n**) The distribution of the normalized connectivity diversity parameter $\hat{g}$ obtained from fitting our model to covariance matrix eigenvalue distributions in data.

pair relates to the average $D(R)$ of the pair (Fig. 5d). Satisfyingly, d' is negatively correlated with firing rate uniformity $D(R)$ (Fig. 5e), agreeing with our prediction.

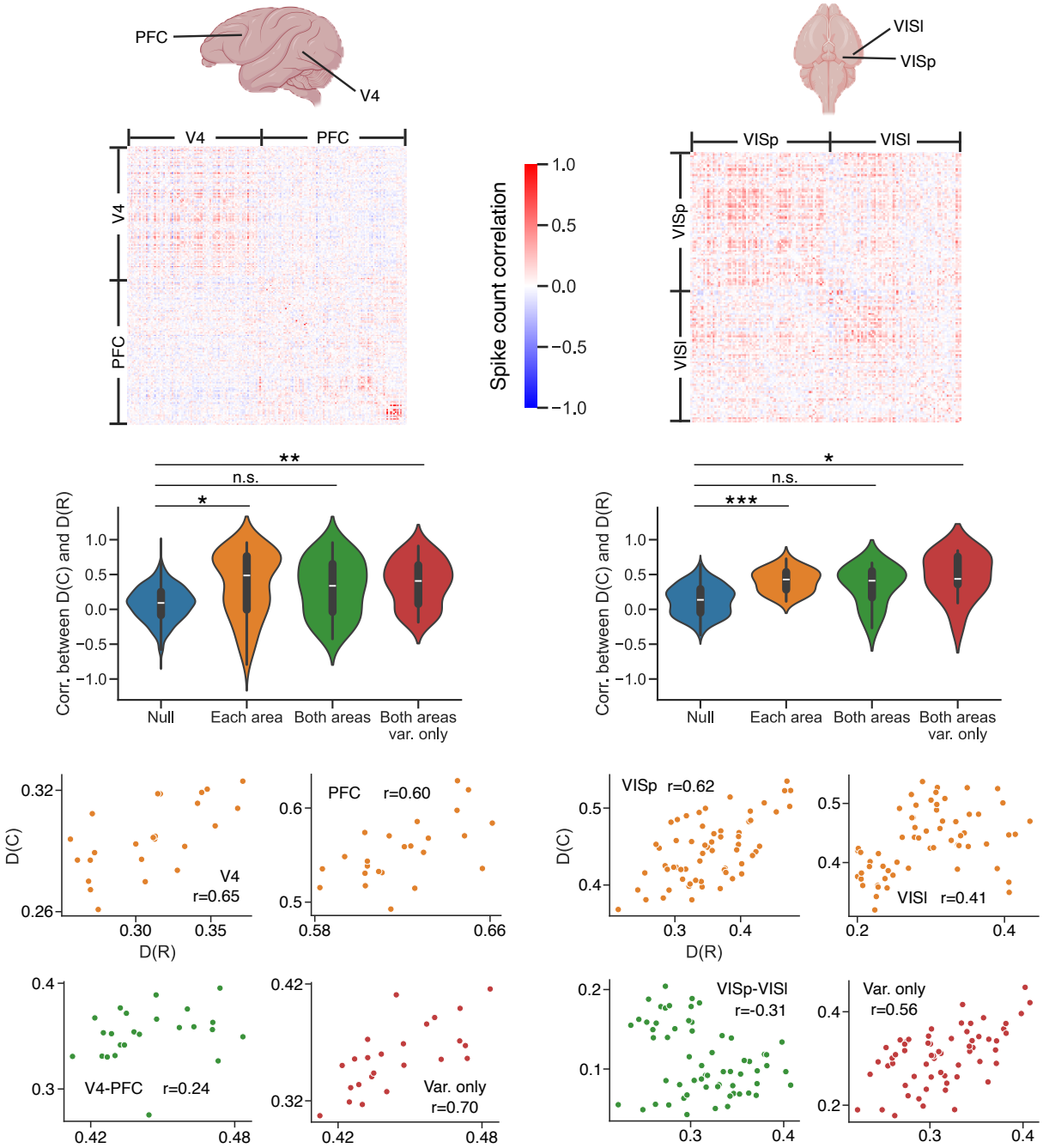

Fig. 4 Neuronal populations that span multiple brain areas breaks the correlation between $D(R)$ and $D(C)$. The left column shows data from monkey PFC and V4 [40]. The right column shows data from mouse VISp and VISl [42]. Top: example correlation matrices (with diagonal elements suppressed). Correlation matrices are used over covariance matrices for visualization purposes only. Middle: distributions of the $[D(R), D(C)]$ correlation across sessions for shuffled data, for each area separately (i.e. calculating $D(C)$ using each of the two diagonal blocks of the full covariance matrix), for combining the two areas (i.e. calculating $D(C)$ using the full covariance matrix), and for only taking into account only the variance but not covariances of the two areas (i.e. calculating $D(C)$ using the diagonal elements of the full covariance matrix). Combining the two areas renders the $[D(R), D(C)]$ correlation indistinguishable from the null, but not if only variances are considered, suggesting the effect cannot be explained solely by single-neuron heterogeneity between areas. Bottom: example $[D(R), D(C)]$ scatter plots from the distributions above taken from the same session for monkey and mouse respectively.

Diverse mean response and lower-dimensional variability in states of enhanced information processing

The evidence presented above (Fig. 5e) can be viewed as describing the geometry of the neuronal code and could depend on which neurons are sampled in the experimental recording. To further investigate $D(R)$'s effect on the entire computational process, we turn to neuronal data spanning different behavioral states where the caliber of information processing is expected to change.

In addition to data collected during the visual categorization task (as described in Fig. 2b), which we call the "evoked state", we also collected resting state data for an extended period of time ($\sim$1 hour) when no task was given, which we refer to as the "spontaneous state". Our theory would predict that neuronal activity has a more diverse mean and a lower-dimensional variability in the evoked state in order to better code for task related sensory stimuli, motor responses and other task variables. Indeed, there is a clear clustering of the data where both $D(R)$ and $D(C)$ are large for the spontaneous state and small for the evoked state (Fig. 5f, top), consistent with the prediction of our theory. The same state-dependent clustering is also present in a dataset from visual areas of mice [42] where the spontaneous state is the same as before and the mice are shown drifting grating stimuli in the evoked state (Fig. 5f, bottom). A reduction of effective dimension by stimulus presentation has also been reported in previous work [46].

In another dataset [43], the same sequence of visual stimuli was shown twice to the same mouse in two blocks, where it gets a reward for reacting to stimulus change in the first block (which we call the "active state") but not the second block (which we call the "passive state"). In the active state, in order to perform the change detection task accurately within the given time frame, the brain needs to not only encode the sensory signal more sharply, but also be more task-engaged to quickly make a decision and transform it into motor command. Consistent with this, we see a smaller $D(R)$ and $D(C)$ across the brain in the active state compared to the passive state (Fig. 5g).

In mice, locomotion is associated with both better performance in visual detection and enhanced neuronal response in the visual cortex of mice [47]. Our theory would then predict more diverse mean response and lower-dimensional noise when the mouse is running to support the behavioral improvement. We indeed observe smaller $D(R)$ for the locomotive state in mouse VISp [42] (Fig. 5h). Interestingly, $D(C)$ is only smaller during locomotion when we remove the outlier eigenvalues, but the relation is flipped when we consider the full eigenvalue distribution (Fig. 5h, compare the bottom plot with the small box). This is not the case for the previous two types of states we considered, where the results are consistent with or without the outliers removed (Fig. 5f-g, compare the bottom plots with the small boxes). This suggests the presence of some low-rank modulation of population covariability associated with locomotion signals that affects the effective dimension in the opposite direction as what is caused by changes in the firing rate distribution, which is distinct from what happens for the evoked/spontaneous and active/passive states.

Discussion

In this paper, we derived a relationship between the diversity of mean response and recurrent connectivity of a neuronal circuit to the effective dimension of its population covariability (Eq. (3)), which was then tested with a wide range of data. The key tool in our analysis is random matrix theory, in particular free probability theory applied to random matrices [34]. It enabled us to combine high-dimensional random variables in a nontrivial way without having to reduce the dimensionality of the system with approximation methods like mean-field theory which inevitably loses information in the process. Random matrix theory has already seen extensive applications both in neuroscience [16, 49–53] and in biology at large [54, 55]. However, most often the only random structure considered in the model is the interaction between the units in the system. Our result suggests that free probability can be very useful when trying to integrate multiple sources of randomness including complex external inputs and internal interactions.

Random matrix theory often produces results that have broad universality properties, meaning that they apply to a large class of matrix distributions satisfying some loose constraints. We developed our theory using i.i.d. Gaussian connection matrices W, but our results should hold as long as W is asymptotically a circular element and is freely independent from R. One sufficient condition is for the elements of W to be i.i.d. with all moments finite, and our result can be applied immediately by matching the variance. While real neuronal networks never satisfy these assumptions in the strict sense, the extent to which the $[D(R), D(C)]$ relationship predicted by our theory holds can be an indirect indicator of the structures in the population's connectivity (or lack thereof). In the datasets we analyzed, we observe a clear positive correlation between $D(R)$ and $D(C)$ when neurons are localized to a single brain region (Fig. 3) but not when neurons span multiple areas (Fig. 4), reflecting the strong patterns in interareal connections. This set of contrasting results highlight the free independence assumption as a key mathematical structure of interest in understanding population covariability in a recurrent circuit.

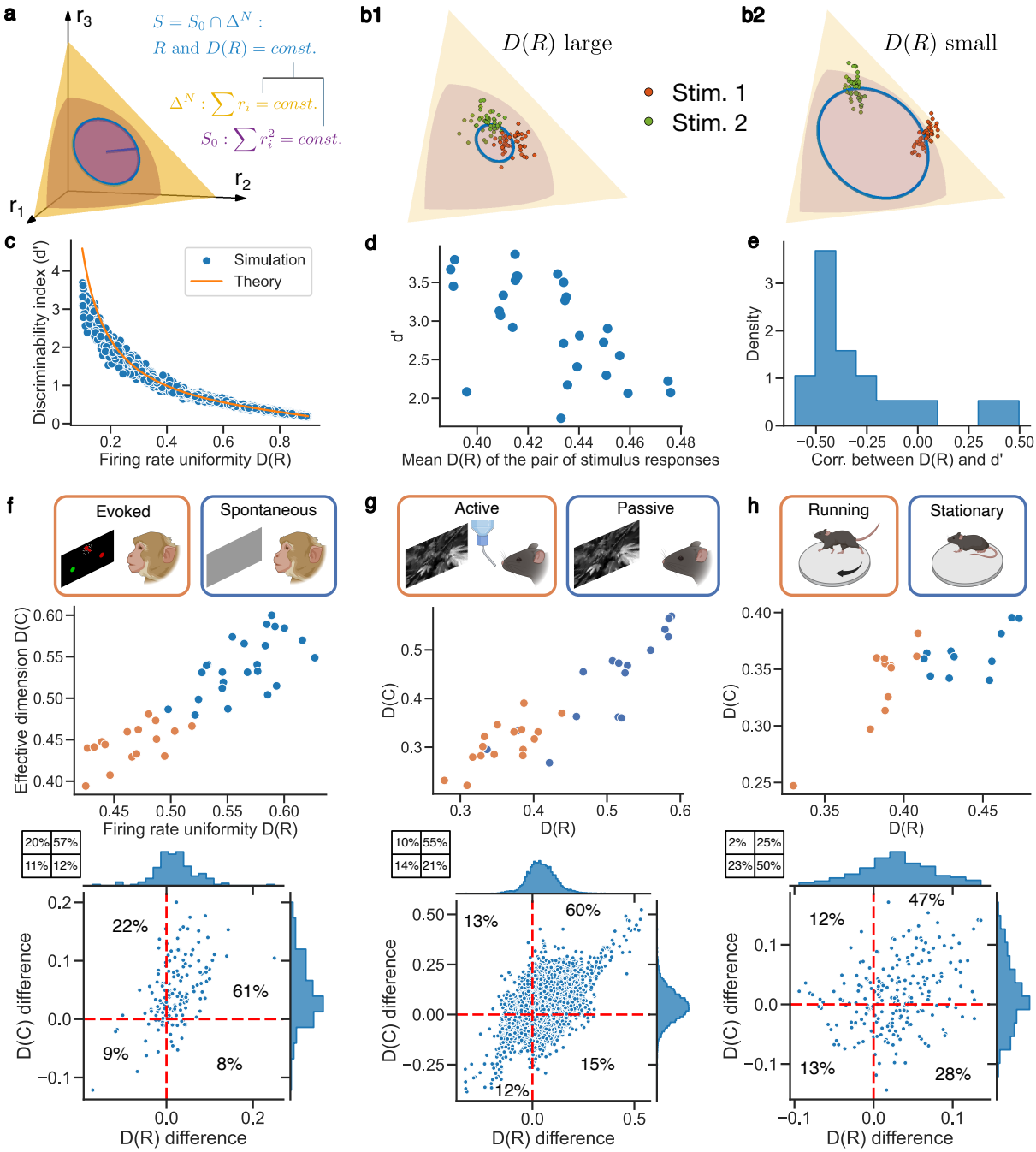

Fig. 5 Computational implications of our result. (**a**) Schematic of the region in neuronal space with constant firing rate uniformity $D(R)$, which is the blue circle defined by the intersection of the sphere of constant second moment and the simplex of constant first moment. The radius of the blue circle is proportional to the standard deviation of the firing rates. (**b**) Schematic of how smaller $D(R)$ improves the population code by enlarging the coding space (given by the size of the blue circle) and reducing the effective dimension of population variability. The red and green point clouds correspond to the neuronal response distributions to two stimuli, which the downstream decoder needs to distinguish between. (**c**) Simulation of how population codes ($N = 100$) with more diverse mean response facilitates decoding as measured by the discriminability index (d'). The theory is given in Eq. (S14). (**d**) An example session of the mean $D(R)$ of the neuronal responses in mouse VISp [42] to two visual stimuli plotted against the estimated d' between them. (**e**) The histogram across sessions of the correlation between the mean $D(R)$ and d' as in (d). It is overwhelmingly negative ($p < 0.001$). (**f**) Data contrasting the evoked state where the animal receives visual stimulus and the spontaneous state where the animal rests freely. The top plot was recorded in monkey FEF. The bottom plot is a summary of mouse visual area data from [42]. The areas included are VISp, VISl, VISrl, VISal, VISam, VISpm, VISmma, VISli, LP, LGd, and SGN. It shows the mean of $D(R)$ and $D(C)$ in the spontaneous state subtracted by those of the evoked state. The small box on the left side between the two large plots shows the percentages in the bottom plot when the outlier eigenvalues are not removed. (**g**) Similar to (f) but for the active state where the mice are performing a change detection task versus the passive state where the same sequence of stimuli is shown but no task is present. All data comes from [43]. The top plot is from VISal. The bottom plot shows the mean of $D(R)$ and $D(C)$ in the passive state subtracted by those of the active state for all recorded areas. (**h**) Similar to (f) but for the comparison between the state where the mice are running versus the state where the mice are stationary. Trials where the mean running speed of the mouse is greater than 1 cm/s are classified as "running" while others are classified as "stationary" [48]. All data is from mouse VISp in [42]. The bottom plot shows the mean of $D(R)$ and $D(C)$ when the mice are stationary subtracted by those when the mice are running.

While the participation ratio $D(C)$ has been widely employed in neuroscience research [16, 18, 28, 56–61], there are a wide range of other measures for the effective dimension of population activity [62, 63]. The key is to use the most appropriate definition for each scenario. At least in the context of fine discrimination tasks, we have shown that $D(C)$ is a reasonable choice as it is directly related to the discriminability index d' (Supplementary Information C). The relevance of $D(C)$ is further supported by its reliable change across brain states in multiple neuronal datasets (Fig. 5f-h). In particular, one should not interpret say $D(C) = 0.4$ in a population of 10,000 neurons as stating there being 4,000 dimensions in its activity each corresponding to some behaviorally meaningful variable. Rather, it is a measure of how concentrated the variability is around a lower-dimensional subspace, which is a statement about its geometry.

Previous works have sometimes produced seemingly different predictions on how the effective dimension depends on network diversities. Susman et al. [64] and Clark et al. [65] predict that the effective dimension of the network dynamics increases with connection weight variance (as measured by larger g), and Gast et al. predicts that the more heterogeneous the neurons are in terms of their firing thresholds, the larger the effective dimension [17], both of which disagree with our prediction in Eq. (3). The main difference between their models and ours is that in their networks population covariability is internally generated from network dynamics and is measured by evolving the system for some length of time. By contrast, in our model variability is fundamentally inherited from (uncorrelated) input noise and is measured by repeating the same experimental trial multiple times. In their models, the network is expanding a temporally one-dimensional input, as is typical in reservoir computing, while our model is responding to a maximal-dimensional input (i.e. $D(C_0) = D(I) = 1$), and can only reduce the effective dimension with network heterogeneities. It is possible that population covariability in the dynamical setting (such as what would be obtained by calculating a covariance matrix across time in the production of a motor movement) may behave differently. Interestingly, we observed similar trends from spontaneous data as well (Fig. 3a, f, g, k), where we calculated the covariance matrix across time (since there was no notion of trials). This suggests that the neuronal circuits being recorded are not in a regime described by Gast et al. in the spontaneous state.

The diverse mean response discussed here is closely related to sparse coding [66]. In fact, the Treves-Rolls measure $D(R)$ used here was originally developed to characterize the sparseness of neuronal codes [25, 26]. There are many benefits to a sparse code [66], and our result provides another perspective on this classical idea. A natural question to ask is why neuronal responses do not always occupy the small $D(R)$ regime, at least in sensory areas. One reason might be to keep the population code smooth with respect to stimulus parameters [67]. As $D(R)$ gets smaller, the codimension 2 manifold S (Fig. 5a) becomes topologically more complex. In particular, it acquires "holes" and becomes "less and less connected" in the sense of homotopical connectivity, eventually becoming disconnected (Eq. (S13)). Representation of stimulus features amounts to embedding the feature space into the neuronal space. A smoother embedding may inevitably need to include regions of larger $D(R)$, and the full neuronal manifold that encodes all stimuli of interest likely spans a range of $D(R)$ values. How different stimuli are mapped to codes with different levels of sparseness is an interesting question. Our results here would predict that stimulus sets that require higher precision in discerning the finer details would correspond to sparser codes. This prediction is supported by a recent study in which training mice to perform a task sparsened the population responses to task-relevant orientations to a much greater extent than non-task orientations [68]. More generally, sparsification of neuronal response due to active learning or repeated passive exposure to sensory stimuli has been observed in the visual cortex [68, 69], somatosensory cortex [70], olfactory bulb [71], auditory cortex [72] and hippocampus [73], and has also been produced in a model with synaptic plasticity rules inferred from data [74]. Thus, a sparsification of neuronal response may be reserved for computations that require a fine discrimination of stimulus features.

In our model W and C_0 are independent of R, so the coding dimensions are independent of the directions of variability in neuronal space. In this setting, the effective dimension is a very good summary of the geometry of neuronal variability. However, it is possible that the modes with the largest eigenvalues are the most important for coding, and the bulk portion of the eigenvalue distribution studied here is not as critical. However, the two largest eigenmodes of mouse primary visual cortex activity have been shown to be orthogonal to the coding dimensions [75], suggesting that they are likely not relevant to visual scene representation. We would further argue that the bulk portion of the eigenspectrum is almost always important. These eigenmodes can be interpreted as originating from the component of the recurrent connectivity that is unrelated to the task at hand but may serve other functions in other contexts, and is high-rank in nature. Low-rank noise can be eliminated downstream if it is shared across areas [76] with adjustments in only a few subspaces, but the bulk of the noise, even with a magnitude not as large, cannot be easily removed due to it being distributed across many dimensions. Reducing the effective dimension by increasing the diversity of the mean response then presents a global strategy of improving the population code without any fine-tuning in the connectivity. To be sure, the correlation between W, C_0 and R definitely

has an enormous impact on the neuronal code, as has been extensively studied [77, 78], and it would be interesting to investigate how to incorporate this aspect into our framework.

Much attention has been devoted to understanding how the structure of population covariability and its effective dimension $D(C)$ in particular can be controlled in the brain [79]. In our model, gain modulation (i.e a change in L) is the root cause of how C changes across brain states. This can be achieved by changing the mean input to the circuit, either feedforward or feedback, which in turn changes the operating point of the system. This amounts to changing the "effective connectivity" of the network, whose effect on neuronal variability has also been studied in previous works [33, 80, 81]. The dominant factor behind a change in effective dimension may be different depending on the brain area and the task setting, and needs to be examined on a case-by-case basis. The mechanism proposed by our model is perhaps the easiest to check, since $D(R)$ can be immediately calculated with the same data that was used to estimate $D(C)$ without any additional procedure. Our result thus sets a basic expectation for changes in $D(C)$ based on changes in $D(R)$. When the effective dimension $D(C)$ changes across conditions or states, one can first check if the changes in $D(R)$ accounts for it (as we show in Fig. 5f-h). If not, one can then look closer at the structure of the covariance matrix C (such as large outlying eigenvalues which could correspond to low-rank feedback modulation) or perform further experiments to reveal other factors responsible for controlling the effective dimension of the variability.

In summary, we derived in a recurrent circuit model that the more diverse the mean response and the recurrent connectivity are in the network, the lower the effective dimension of population covariability. We verified this result in neuronal data from multiple brain areas from monkeys, mice, and humans. We showed that a diverse mean response works synergistically with low-dimensional noise to improve the population code, which is again supported by experimental evidence. Our work provides a mechanistic framework for understanding how different sources of diversity in a network interact to determine the geometry of neuronal variability.

Methods

Non-human primate experiments

Subjects and behavioral tasks. Two male rhesus monkeys (Macaca mulatta) participated in this experiment: Monkey M, 21-24 years old, ∼10 kg; Monkey S, 9-12 years old, ∼13 kg. Both monkeys were implanted with a headpost and trained to perform the memory-guided saccade (MGS) task and the rapid categorization (RC) task (described in detail in the next paragraph). For all tasks, the monkeys were head restrained and seated in a primate chair inside an isolation box (Crist Instruments) or in a dedicated experimental room. Visual stimuli were displayed on a 21-inch color CRT monitor (1280×1024 resolution, 75 Hz refresh rate, 57cm viewing distance). Gaze positions were tracked using an Eyelink 1000 optical eye tracker (SR Research, Ottawa, Canada) at a sample rate of 1 kHz. Task events, stimulus presentation, behavioral signals, and reward delivery were controlled by the MATLAB-based toolbox NIMH MonkeyLogic [82, 83] on a Windows 10 machine. After training, both monkeys were implanted with two custom titanium recording chambers in the right hemisphere (Rogue Research): one chamber was positioned over the frontal eye fields (FEF) and the other was positioned over both the lateral intraparietal area (LIP) and the superior colliculus (SC) based on stereotaxic coordinates from pre-implantation MRI scans. All surgical and experimental procedures adhered to the guidelines of the University of Chicago's Institutional Animal Care and Use Committee and the National Institutes of Health.

Memory-guided saccade task. To identify visual and motor receptive fields of LIP, FEF, and SC, we used the MGS task [84–86]. Monkeys initiated trials by acquiring and maintaining fixation on a central white fixation point (0.2° radius) within a 2.5° fixation window throughout the trial. After 500 ms of fixation, a 0.6° white square target was briefly presented at one of eight peripheral locations (7° eccentricity) for 300 ms. After a 1000 ms delay period, the fixation cue disappeared, and monkeys made a saccade to the remembered location of the target for a juice reward delivered through a custom-made juice-spout. Recording coordinates in all areas were selected if they contained neurons with visual or motor receptive fields to the left of the fixation point (contralateral to the recorded hemisphere).

Rapid categorization task. Similar to the one-interval categorization task reported in previous experiments [86], monkeys categorized random-dot motion stimuli into one of two categories with a saccadic eye movement. The motion stimuli consisted of 5° diameter circular patches of ∼100 high-contrast dots moving at 12°/s with 100% coherence. A set of six directions spanning 360° range of motion was arbitrarily divided into two categories, and monkeys must report the learned motion category with a saccade to a red or green color target (Red: 75°, 135°, 195°; Green: 15°, 255°, 315°). Monkeys initiated trials by acquiring and maintaining fixation on a central fixation point (0.2° radius) within a 2.5° fixation window throughout the trial. After 500 ms of fixation, a place-holder random dot motion stimulus with 0% coherence (5° diameter) was positioned 7° to the left of the fixation point. This place-holder stimulus is used to avoid the bottom-up attention capture of the full coherence motion stimulus and any transients of neuronal activity corresponding to the appearance of any stimulus in the receptive field of the neuron. After 300 ms of continued fixation, red and green targets (1° square) appeared at positions 7° above and below the fixation point. Importantly, the relative positions of the color targets were randomized on every trial to decouple specific saccade directions from motion categories (e.g. the red target can either be above or below the fixation point on different trials). After 350 ms of continued fixation, the 0% coherence noise stimulus is replaced by a 100% coherence motion stimulus and simultaneously, the fixation point disappears, signaling to the monkeys that they must make a saccade within 500 ms. If they saccade to and fixate on the correct color target for 300 ms, they receive a juice reward delivered through a juice spout.

Passive viewing (spontaneous state). To measure the spontaneous patterns of activity in each area, recordings continued after performing the previous tasks while monkeys passively sat in a dark environment for 30-60 minutes without any visual stimuli on the monitor, sounds, or behavioral requirements. Gaze positions were continuously tracked to monitor for spontaneous eye movements or for periods of sleep.

Electrophysiological recordings. Electrophysiological procedures are similar to previous studies from our group [86]. Recordings were conducted in separate sessions using combinations of 24-channel or 32-channel linear arrays (Plexon V-probes and Plexus S-probes) and a NAN microdrive system (NAN Instruments). Up to six probes were used to simultaneously record LIP, FEF, and SC (two per area, up to four in one recording chamber). Neurophysiological signals were amplified, digitized, and stored for offline spike sorting (Plexon) using the MATLAB-based Kilosort 2 toolbox, which ran on a Windows 10 PC with a NVIDIA RTX 3070 graphics card [87]. Putative neuron clusters identified via Kilosort 2 were curated using a Python toolbox Phy [88].

Human experiment

Participants. This study was conducted under an Investigational Device Exemption from the U.S. Food and Drug Administration and approved by the Institutional Review Boards at the Universities of Pittsburgh and Chicago. The clinical trial is registered at ClinicalTrials.gov (NCT01894802). Informed consent was obtained before any study procedures were conducted. Participant C1 (male, m), 55-60 years old at the time of implant, presented with a C4-level ASIA D spinal cord injury (SCI) that occurred 35 years prior to implant. C1 retains partial function of the contralateral arm and risk but can only achieve hand movements with tendonesis grip. Participant C2 (male), 60-65 years old at time of implant, presented with a C4-level ASIA D spinal cord injury (SCI), along with a right brachial plexopathy, that occurred 4 years prior. C2 retains some shoulder function and wrist function.

Cortical implants. We implanted four microelectrode arrays (Blackrock Neurotech, Salt Lake City, UT, USA) in each participant. The two arrays (one medial and one lateral array) in Brodmann's area 1 of S1 were 2.4 mm x 4 mm with sixty 1.5 mm long electrode shanks wired in a checkerboard pattern such that 32 electrodes could be recorded from or stimulated. The two arrays in motor cortex were 4 mm x 4 mm with one-hundred 1.5 mm long electrode shanks wired such that 96 electrodes could be used to monitor neuronal activity. The inactive shanks were located at the corners of these arrays. Two percutaneous connectors, each connected to one sensory array and one motor array, were fixed to the participant's skull. We targeted array placement during surgery based on functional neuroimaging (fMRI) or magnetoencephalography (MEG) of the participants attempting to make movements of the hand and arm within the constraints of anatomical features such as blood vessels and cortical topography – see [89] for a full description of the implantation approach.

Center-out reaching task. To facilitate arm movements given the reduced motor function of each participant, each participant's arm was rest upon a forearm support with ball bearings to reduce friction and the need to lift the arm during movement. Participant C1 was provided 8 target positions equally spaced around a circle 30 cm in diameter. Due to reduced shoulder function of participant C2, 4 targets were sub-selected (45, 90, 135, 225 degrees). Each trial was composed of 5 phases, all 1 second in duration and indicated by auditory tones. The 'cue phase' was indicated with a long auditory tone at which point the trial target would be indicated to the participant via text on a TV (e.g. "top-left", "bottom-right"). After 1 second the 'movement phase' began as indicated by a short tone at which point each participant would then slowly move their arm to the indicated target over the 1 second phase period such that they arrived at the target when the tone for the next phase occurred. Participants then remained stationary during the 'wait phase' until the 'return phase' tone played at which point they slowly returned to the center position. Finally, the participants waited for another second before the next trial began. Participants were allowed to pause as requested. Each participant completed 11 blocks of reaches in which each trial was repeated 10 times. Consequently, participant C1 completed 880 reaches and participant C2 completed 440 reaches.

Passive 'task'. To contrast with the active nature of the center-out reaching task we also had participants relax and watch TV (a nature documentary) for approximately 60-90 minutes. Participant C1 remained engaged with the documentary for the entire duration and often made comments to the experimenters. Participant C2 was less engaged and fell asleep approximately half way through the documentary.

Data acquisition and processing. Cereplex ES headstages (Blackrock Neurotech, UT, USA) were used to connect each pedestal to a NeuroPort system (Blackrock Neurotech, UT, USA). Voltages were sampled at 30 kHz and high-pass filtered (750 Hz) with a 1st order Chebychev filter. An RMS threshold of -4.5 was used to detect spikes from the continuous signal. Post-hoc offline sorting of single units was performed using Offline Sorter (v4.7.1, Plexon, TX, USA).

Measure of mean response uniformity and effective dimension

The Treves-Rolls measure [25, 26] for quantifying the uniformity of firing rate in a neuronal population is defined as

$$D(R) = \frac{\mathbb{E}(R)^2}{\mathbb{E}(R^2)} = \frac{(\sum_i r_{0i})^2}{N \sum_i r_{0i}^2},$$

where r_{0i} is the firing rate of neuron i. When all neurons have the same firing rate, we have $D(R) = 1$, marking the most uniform activity level. As neurons fire at more and more diverse rates, $D(R)$ becomes lower. This can be seen as an interpolation of the cases where M neurons fire at the same rate while other neurons are silent, which has $D(R) = M/N$.

The participation ratio [27, 28] as a measure of the effective dimension of populatino covariability is defined as

$$D(C) = \frac{\mathbb{E}(\lambda)^2}{\mathbb{E}(\lambda^2)} = \frac{(\sum_i \lambda_i)^2}{N \sum_i \lambda_i^2},$$

where λ_i is the i'th eigenvalue of the spike count covariance matrix C. When variability is isotropic in neuronal space with the same magnitude in all directions, we have $D(C) = 1$, signifying the maximum effective dimension. As the noise is distributed more and more unevenly in different direction, $D(C)$ becomes lower. This can be seen as an interpolation of the cases where M eigenvalues of C have the same positive value while the other eigenvalues are 0, i.e. where the variability is restricted to an M-dimensional subspace, which has $D(R) = M/N$.

The formal similarity of these measure is not a coincidence, as the participation ratio is exactly measuring the uniformity of noise across orthogonal dimensions.

Simulating firing rates with prescribed $D(R)$

Here we describe how to sample firing rates from a log-normal distribution $\mathcal{LN}(\mu, \sigma^2)$ with a given mean $\bar{R}$ and uniformity $D(R)$ as we did in Fig. 2a). For $X \sim \mathcal{LN}(\mu, \sigma^2)$, we have

$$\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \mathbb{E}[X^2] = \exp(2(\mu + \sigma^2)), \quad D(X) = \exp(-\sigma^2).$$

Then

$$\sigma^2 = -\log(D(R)), \quad \mu = \log(\bar{R}) - \sigma^2/2,$$

gives us the desired paramters for the log-normal distribution.

Simulations with spatially structured connectivity and input noise

To make Fig. S3, we perform similar procedures as Fig. 2a except W and C_0 are generated in a spatially dependent manner. Specifically, each neuron in the population is randomly assigned a location in a two-dimensional grid of length 1. For neurons i and j that are distance d_{ij} apart, we let the connection between them be $w_{ij} = w(d_{ij})$ and the input covariance between them be $c^0_{ij} = c_0(d_{ij})$, where w is the Laplacian of a Gaussian with mean 0 and standard deviation α_W, and c_0 is a Gaussian with mean 0 and standard deviation α_{C_0}. The reason we use the Laplacian of a Gaussian for w is to keep the mean of the elements of W near 0. The W and C_0 generated this way are Euclidean random matrices [90]. Note that we are not explicitly modeling the covariance matrix C itself as an Euclidean random matrix, as was done in previous work [51].

Data processing

We describe here the general approach of processing the population electrophysiology data shown in Fig. 3. When there is a trial structured present during the recording such as the presentation of stimuli or the execution of some motor action, we first separate the trials into groups with the same experimental condition. This includes the same sensory stimuli, task cue configuration, behavior of the subject etc. Within each group, we further partition the trials into different epochs delimiting different stages of a trial such as fixation, stimulus presentation, and behavioral response. Finally, we find a 100 ms bin in each epoch where neuronal activity is the most stationary and count the number of spikes in the bin for each neuron. Then we obtain a set of spike count matrices of size (number of trials $\times$ number of neurons) for each experimental condition and trial epoch. Each of these matrices are regarded as samples from some distribution of neuronal response, from which we can calculate the firing rate distribution and the covariance matrix across trials for this operating point of the system. This in turn gives us $D(R)$ and $D(C)$. For example, 4 conditions with 5 epochs in each trial would give us 20 points in the $[D(R), D(C)]$ scatter plot. For resting state data where the subject is spontaneously behaving without any trial structure, we take the system to roughly be at the same operating point in each minute-long recording, and use the slow drift of the system's state across time [91] to sample different $[D(R), D(C)]$ pairs in place of experimental conditions and trial epochs. That is, we calculate the mean responses and covariances for each minute in the recording by binning it into 100 ms spike counts. Then for example, a 30-minute long spontaneous state recording would produce 30 points in the $[D(R), D(C)]$ scatter plot. An additional point to note is, previous work has shown in monkeys that whether the eyes are open or closed can have significant impact on the spiking activity of neurons including its effective dimension [92, 93]. We therefore removed the portion of the data where the monkey has its eyes closed or is making a saccade using eye-tracking data.

Fitting the eigenvalue distribution to data and detecting the outliers. As discussed in the main text, our model only accounts for the bulk of the covariance matrix eigenvalue distribution and we need to remove the outlier eigenvalues before verifying our theory. We do this by fitting our model to the data. Given firing rate distributions R_i where i indexes different operating points (due to different experimental conditions and trial epochs as discussed earlier) and covariance matrix eigenvalues $\lambda_{0i} \in \mathbb{R}^M$ from data

recorded from a population of M neurons, we need to fit two parameters: the connection diversity parameter g and the number of neurons in the underlying population N. Synaptic connectivity is assumed to stay constant in each recording session, so g and N do not depend on the operating point i. Since we are considering spatial sampling here (i.e. $M < N$), there is no analytic solution to the eigenvalue problem (Supplementary Information B.2), so we need to use a search algorithm to find the combination of g and N that best explains the data. We use Bayesian optimization [94] to perform the search. The cost function is defined by

$$L(g, N) = \sum_i AD(\lambda(g, N, R_i), \lambda_{0i}),$$

where AD denotes the Anderson-Darling statistic [95] which measures how close two distributions are to each other and $\lambda(g, N, R_i)$ is the samples generated by first simulating an $N \times N$ covariance matrix with parameter g according to Eq. (1) (with $p = 2$ and $C_0 = I$) using the firing rate distribution R_i and then randomly picking a $M \times M$ principle submatrix to calculate eigenvalues. This procedure is done many times (typically $\sim 100M$) to ensure a good sampling of the model distribution. Each repetition has either a different realization of connectivity W or a different sampled submatrix. There is an unknown scaling parameter as summarized by k_0 in Eq. (S5). We eliminate its effect by matching the means of $\lambda(g, N, R_i)$ and λ_{0i} before calculating the Anderson-Darling statistics. Note that this simulation process is inherently stochastic since the generated eigenvalue samples depend on the realizations of the weight matrix W and which M neurons are sampled from the total size-N population (even after repeating the simulation for many times), but this stochasticity is allowed by the Bayesian optimization algorithm.

There is sometimes degeneracy between g and N in that there may be a continuous set of combinations that produce similarly small cost functions. However, for the purpose of getting a null model for the eigenvalue distribution and identifying the outliers in the data, any good pair of $\hat{g}$ and $\hat{N}$ can be used to generate $\lambda(\hat{g}, \hat{N}, R_i)$ that sufficiently captures the bulk of λ_{0i}. Once we have a good sample of $\lambda(g, N, R_i)$, we can identify the eigenvalues in the neuronal data that lie outside of its support as outliers. Alternatively, when M is relatively small, even a data eigenvalue within the null model support may be unexpected if it is very large. Let λ_1 be a large eigenvalue in neuronal data. We can set a reference p-value p_0 and calculate the likelihood of obtaining a sample from $\lambda(\hat{g}, \hat{N}, R_i)$ no smaller than λ_1 in M draws. If that likelihood is smaller than p_0, we say λ_1 is an outlier. In our study we adopt the second method with $p_0 = 0.01$. We typically find 0-2 outliers in data. All of the $D(C)$ values presented in the figures are calculated after removing these outlier eigenvalues except in Fig. 3e,j, in the small boxes in Fig. 5f-h, and in Fig. S7.

In Fig. S8c-d, we compared the fit of our model ("diverse gain") to those of two other models. For the "uniform gain" model [16], we set R to be uniform across neurons when generating $\lambda(g, N, R_i)$ and performed the same optimization procedure as above. In particular, the number of fitted parameters is the same, and the only difference is the "uniform gain" model does not incorporate information from the firing rate distribution. For the "uncoupled" model we simply multiplied the firing rate distribution by a scalar to match its mean with the eigenvalues to get the eigenvalues that would be produced by a population of uncoupled Poisson neurons.

Correcting for temporal sampling. In Supplementary Information B.1, we discuss how a limited number of temporal samples/trials affects the estimation of the effective dimension $D(C)$. To correct for this effect, we use the simple formula

$$D(C) = \frac{1}{1/D(\hat{C}) - \alpha},$$

where $D(\hat{C})$ is the naive sample covariance matrix and $\alpha = N/T$ with T being the number of temporal samples. This recovers the effective dimension of the exact covariance matrix. However, this procedure only corrects for the bias due to temporal sampling, not the variance. If T is too small, the estimate of $D(C)$ would still be very noisy even after the correction.

Shuffled data. We describe here how we generated the null distribution in Fig. 3d, i. What is measured by the correlation between $D(C)$ and $D(R)$ is how these two quantities covary when the operating point of the system changes. However, due to finite-size effect and the discrete nature of spike count data, we may observe spurious correlation between them even when we repeatedly sample spike count data from the same operating point. To control for this, we need to compare with shuffled data. Suppose we have K operating points in the data (i.e. K points in the $[D(R), D(C)]$ scatter plot). We take the operating point with $D(C)$ and $D(R)$ closest to the mean of all $[D(R), D(C)]$ pairs and generate K spike count matrices each having the same number of temporal samples as the data according to the firing rate vector and covariance matrix of this operating point using the algorithm in [96]. This in turn gives us K points in the $[D(R), D(C)]$ plane which can be used to calculate a correlation. This is done 10 times for each brain area and recording session. The resulting set of correlation values compose the null histograms.

Calculation of the discriminability index

For two stimulus responses with means r_1, r_2 and covariance matrices C_1, C_2, the discriminability index d' is [97]

$$(d')^2 = 2(r_2 - r_1)^T (C_1 + C_2)^{-1} (r_2 - r_1).$$

For the simulation in Fig. 5c, we first drew r_1 from some log-normal distribution with a certain mean and $D(R)$ as we did in Fig. 2a. We then sample r_2 in the same manifold S (Fig. 5a at a distance $d_0(1/D(R) - 1)$ from r_1 in a random direction. Here d_0 is a scaling constant and the distance gets larger at the same pace as the overall size of S (Supplementary Information C). This is achieved by first sampling a random unit vector in the neuronal space, then projecting it to the linear subspace orthogonal to both the vector of ones (to ensure it stays in the simplex Δ^N) and the vector r_1 (to ensure it roughly stays in the sphere S_0), and finally scaling it by $d_0(1/D(R) - 1)$ to obtain $r_2 - r_1$. C_1 and C_2 are again generated according to Eq. (1) with $p = 2$ and $C_0 = I$. Then it is straightforward to calculate d'.

To estimate d' in neuronal data in Fig. 5d-e, we fit linear classifiers (logistic regression) to each pair of stimulus responses and get the cross-validated classification accuracy a. d' is related to a by [97]

$$d' \approx 2Q(a),$$

where Q is the quantile function of the standard normal.

Data availability. Will be available for public use upon manuscript publication.

Code availability. All modelling and analysis codes are provided at https://github.com/brain-math/fr-diversity-lowers-cov-dimensionality

Acknowledgments. We thank Michael Boninger, MD, for serving as the sponsor-investigator of the BCI clinical trial conducted under an FDA Investigational Device Exemption. C.G. and J.D. are supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (NIH) under award numbers UH3 NS107714, R35 NS122333, and R01 NS130302. D.J.F is supported by NIH grant R01EY019041 and a Vannevar Bush Faculty Fellowship (N000141912001). O.Z. is supported by NIH grants F30EY033648 and T32GM007281. B.D. is supported by the NIH grants 1U19NS107613-01, R01NS133598 and CRCNS-R01EY034723, a Vannevar Bush faculty fellowship (N000141812002), and the Simons Foundation Collaboration on the Global Brain. This research benefited from Physics Frontier Center for Living Systems funded by the National Science Foundation (PHY-2317138). Financial support was provided via the National Institute from Mathematics and Theory in Biology (Simons Foundation award MP-TMPS-00005320 and NSF award DMS-2235451). Cartoon schematics created with BioRender.

Author contributions. G.J.T. and B.D. conceived of the project, theory, and data analysis. G.J.T. performed the analytic calculations and the data analysis. O.Z., V.S. and D.J.F. performed the monkey electrophysiology recordings (areas LIP, FEF, and SC). C.M.G. and J.E.D. performed the human participant electrophysiology recordings. All authors were involved in the writing of the manuscript.

Competing interests. None to declare.

Supplementary Information

A Derivation of main result

A.1 Expression for the covariance matrix

To study the effective dimension $D(C)$, we need to understand C. We do this in a mechanistic model of a recurrent circuit. Consider a network of N neurons

$$\tau \frac{dr_i}{dt} = -r_i + f\left(\mu_i + \sum_j w_{ij} r_j + \xi_i\right), \quad i = 1, \ldots, N, \tag{S1}$$

where r_i is the response of neuron i, τ is the neuronal time constant, f is the input-output transfer function, μ_i is the mean external input, w_{ij} is the connection weight from neuron j to neuron i, and ξ_i is trial-to-trial input noise. In vectorized notation we can write

$$\tau \frac{dr}{dt} = -r + f\left(\mu + Wr + \xi\right),$$

where $r := (r_i)_i, \mu := (\mu_i)_i, W := (w_{ij})_{i,j}, \xi := (\xi_i)_i$, and f is applied entry-wise.

In each trial, assume the network evolves to a steady state

$$r = f\left(\mu + Wr + \xi\right). \tag{S2}$$

Here r is a function of ξ, which is a random variable. When the noise ξ is small, we can linearize to get the approximation

$$r \approx r_0 + \nabla_\xi r \cdot \xi, \tag{S3}$$

where r_0 is the steady state when there is no noise, i.e.

$$r_0 = f\left(\mu + Wr_0\right).$$

Differentiating Eq. (S2) gives us

$$\nabla_\xi r = L(W\nabla_\xi r + I) \quad \Rightarrow \quad \nabla_\xi r = (I - LW)^{-1} L,$$

where $L := \mathrm{diag}(l_1, \ldots, l_N)$ is the diagonal matrix of neuronal gains

$$l_i := f'\left(\mu_i + \sum_j w_{ij} r_{0j}\right).$$

Then

$$r - r_0 \approx (I - LW)^{-1} L\xi,$$

and the covariance matrix C is

$$C = \langle (r - r_0)(r - r_0)^T \rangle = (L^{-1} - W)^{-1} C_0 (L^{-1} - W)^{-T},$$

where $C_0 = \langle \xi\xi^T \rangle$ is the covariance matrix of the input noise, which we assume to be $C_0 = I$ in the following. This is exactly Eq. (1).

A wide range of linear or linearizable systems can lead to Eq. (1) [30]. Here we present an alternative formulation with spiking neurons, namely the Hawkes process model [98, 99]. Let each neuron i generate a Poisson spike train $s_i(t)$ (as a sum of delta functions) according to firing rate $r_i(t)$, which is determined by

$$r_i(t) = \mu_i + \int_0^\infty l_i(t') \sum_j w_{ij} s_j(t - t') dt' = \mu_i + \sum_j w_{ij} (l_i * s_j)(t),$$

where μ_i is a constant baseline, w_{ij} is the connection weight from neuron j to i, l_i is the response kernel that is supported on the positive real line, and $*$ denotes convolution. Taking the Fourier transform of this equation gives us

$$\tilde{r}_i(\omega) = \mu_i + \sum_j w_{ij} \tilde{l}_i(\omega) \tilde{s}_j(\omega) m,$$

which can be written in vector form as

$$\tilde{r} = \mu + \tilde{L}W\tilde{s},$$

where $\tilde{L} = \mathrm{diag}(\tilde{l}_1, \ldots, \tilde{l}_N)$. Averaging both sides gives us

$$\tilde{r}_0 = \mu + \tilde{L}W\tilde{r}_0.$$

Subtracting the above two equations gives us

$$\tilde{r} - \tilde{r}_0 = \tilde{L}W(\tilde{s} - \tilde{r}_0).$$

We then have

$$\tilde{s} - \tilde{r}_0 = \tilde{s} - \tilde{r} + \tilde{L}W(\tilde{s} - \tilde{r}_0) \quad \Rightarrow \quad \tilde{s} - \tilde{r}_0 = (I - \tilde{L}W)^{-1}(\tilde{s} - \tilde{r}).$$

Then the cross-spectrum is

$$\tilde{C}(\omega) = \langle(\tilde{s} - \tilde{r}_0)(\tilde{s} - \tilde{r}_0)^T\rangle = (I - \tilde{L}(\omega)W)^{-1}\tilde{C}_0(\omega)(I - \tilde{L}(\omega)W)^{-T},$$

where $\tilde{C}_0(\omega) = \langle(\tilde{s} - \tilde{r})(\tilde{s} - \tilde{r})^T\rangle$ is the cross-spectrum of the Poisson spike generation process, which is diagonal. This expression highlights the doubly-stochastic nature of the Hawkes model.

By the Wiener-Khinchin theorem, we can Fourier transform the cross-spectrum $\tilde{C}(\omega)$ back to the cross-covariance $C(\tau)$, and the long-time covariance matrix which is the integral of the cross-covariance is just $C := \tilde{C}(0)$. The (diagonal) elements of $L := \tilde{L}(0)$ are the integrated effects of a spike to the post-synaptic firing rate which we refer to as neuronal gains, and those of $C_0 := \tilde{C}_0(0)$ are the variances of the spike generation process. Then

$$C = (I - LW)^{-1}C_0(I - LW)^{-T}.$$

Since the spikes are Poisson, we have $C_0 \propto R$. When $p = 2$ in the transfer function Eq. (2), we have $L^2 \propto R$. Putting these together we have

$$C \propto (R^{-1/2} - W)^{-1}(R^{-1/2} - W)^{-T},$$

which recovers Eq. (1) for $p = 2$ and $C_0 = I$. Notice that we do not need to assume small noise in this formulation. One could also study temporal properties of neuronal response such as the dependence of effective dimension $D(C)$ on the window size used for binning spike counts as we now have access to the full cross-spectrum [100].

A.2 The covariance eigenspectrum

We shall next derive the eigenvalue distribution of the covariance matrix C under the assumptions $w_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, g^2/N)$, that f follows Eq. (2), and that L and W are independent. Then

$$L = pk^{1/p}R^{1-1/p}, \tag{S4}$$

where $R = \mathrm{diag}(r_0)$. (Throughout the paper, we also use R to denote the distribution of firing rates, which is exactly the eigenvalue distribution of R as a diagonal matrix.) Since we are ultimately interested in the effective dimension $D(C)$, which is a normalized quantity that is invariant to a global scaling of C's entries, we can consider

$$C = k_0(R^{1/p-1} - W)^{-1}(R^{1/p-1} - W)^{-T}, \tag{S5}$$

where W has been rescaled by some scalar and all scaling has been absorbed into one variable k_0. We shall use this form for all of our following analyses. We will also write $L = R^{1-1/p}$, which differs from the actual neuronal gain in Eq. (S4) by a constant.

Following standard notation in free probability, we define for any random matrix M of size N

$$\varphi(M) := \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\mathrm{tr}(M)\right],$$

where $\mathrm{tr}(M) = \sum_i M_{ii}$ is the trace. Note that all derivations here are in the large N limit, and we shall not make a clear distinction between a random matrix of finite size with the object it converges to, nor shall we discuss different notions of convergence such as weak and a.s. convergence. We observe in simulations that all of our results converge for N at a few hundred, which is comparable to the size of the neuronal data we analyze.

First, note that Eq. (1) only holds when the spectral radius $\rho(LW) < 1$. Using results in free probability about the product of isotropic matrices [101, 102], we have

$$\rho(LW) = \varphi(W^T L^2 W)^{1/2} = \varphi(g^2 L^2 S)^{1/2} = (g^2 \varphi(L^2)\varphi(S))^{1/2} = g\varphi(R^{2(1-1/p)})^{1/2},$$

where S is a free Poisson element [34] with parameter 1 (i.e. it obeys the Marchenko-Pastur law). The same result was obtained in previous work using mean-field theory [103]. Then the condition becomes

$$g < \varphi(R^{2(1-1/p)})^{-1/2}.$$

When $p = 2$, we in particular have $g < \bar{R}^{-1/2}$ where $\bar{R}$ is the population mean firing rate. If we define the normalized parameter $\hat{g} := g\sqrt{\bar{R}}$, then the condition is simply

$$\hat{g} < 1.$$

We shall next use tools from operator-valued free probability theory to derive the eigenvalue distribution of the covariance matrix C [34] (in particular following pages 245-247). Denote by a, b, and c the corresponding variables in the non-commutative probability space $(\mathcal{C}, \varphi)$ of L^{-1}, W, and $k_0^{-1}C$ respectively. Then Eq. (S5) becomes

$$c = (a-b)^{-1}(a-b)^{-*}.$$

The distribution of c is the same as the -2 power of

$$T = A + B, \quad A = \begin{pmatrix} 0 & a \\ a^* & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & b \\ b^* & 0 \end{pmatrix}. \tag{S6}$$

(b and $-b$ are identical, so we replace b with $-b$ for simplicity.) A and B are free $\mathcal{D}_2$-valued self-adjoint elements in $\mathcal{A} := M_2(\mathcal{C})$. Here $\mathcal{D}_2$ is the set of diagonal 2×2 complex matrices. Explicitly, the conditional expectation $E : \mathcal{A} \to \mathcal{D}_2$ is defined to be

$$E\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \varphi(a_{11}) & 0 \\ 0 & \varphi(a_{22}) \end{pmatrix}.$$

We can take E to have a range in $\mathcal{D}_2$ because T is even due to it having zeros on the diagonal. In particular, B is a $\mathcal{D}_2$-valued semi-circular element. The covariance function of B, $\eta : \mathcal{D}_2 \to \mathcal{D}_2$ is given by

$$\eta(D) = \eta\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} = \begin{pmatrix} d_2\varphi(bb^*) & 0 \\ 0 & d_1\varphi(b^*b) \end{pmatrix} = g^2 D.$$

Let G_X be the $\mathcal{D}_2$-valued Cauchy transform of X. We are ultimately interested in the distribution of T^{-2}. Note that there is a relation

$$G_T(z) = z^{-1}[1 - z^{-2}G_{T^{-2}}(z^{-2})]. \tag{S7}$$

The scalar-valued Cauchy transform $G_c(z)$, which ultimately determines the eigenvalue distribution p_c of c, can be calculated as

$$G_c(z) = \frac{1}{2}\mathrm{tr}(G_{T^{-2}}(z)).$$

Let us write

$$G_{T^{-2}}(z) = \begin{pmatrix} G_1(z) & 0 \\ 0 & G_2(z) \end{pmatrix}, \quad G_T(z) = \begin{pmatrix} G_1'(z) & 0 \\ 0 & G_2'(z) \end{pmatrix}.$$

Using results from [34], we have

$$\begin{aligned}
&\begin{pmatrix} z^{-1} - z^{-3}G_1(z^{-2}) & 0 \\ 0 & z^{-1} - z^{-3}G_2(z^{-2}) \end{pmatrix} \\
=&z^{-1}[1 - z^{-2}G_{T^{-2}}(z^{-2})] = G_T(z) = G_{A+B}(z) = G_A[z - R_B(G_T(z))] \\
=&E\left[(z - R_B(G_T(z)) - A)^{-1}\right] = E\left[(z - \eta(G_T(z)) - A)^{-1}\right] = E\left[\left(z - g^2 G_T(z) - A\right)^{-1}\right] \\
=&\begin{pmatrix} \varphi\left[(z - g^2G_1'(z) - (z - g^2G_2'(z))^{-1}a^2)^{-1}\right] & 0 \\ 0 & \varphi\left[(z - g^2G_2'(z) - (z - g^2G_1'(z))^{-1}a^2)^{-1}\right] \end{pmatrix}.
\end{aligned}$$

where R_B is the R-transform of B. Combining this with Eq. (S7) gives us a system of 2 equations of G_1 and G_2. Note that (from the definition of the Cauchy transform) the self-adjointness of a implies that $G_1(z) = G_2(z) = G_c(z)$, so we get a single equation

$$2(1 - zG_c(z)) = \varphi\left[\frac{1}{1 - g^2 z(1 - zG_c(z)) + z^{1/2}a} + \frac{1}{1 - g^2 z(1 - zG_c(z)) - z^{1/2}a}\right]. \tag{S8}$$

Note that the functional φ is with respect to a, the variable corresponding to L^{-1}, which is in turn related to the firing rates R. Then we can numerically solve this equation for G_c when we have samples from the firing rate distribution. The Cauchy transform is related to the eigenvalue distribution p_c by

$$p_c(x) = -\lim_{y\to 0^+} \frac{1}{\pi}\Im[G_c(x + iy)].$$

Eq. (S8) was verified by simulation data (Fig. 3a, rightmost column). It is consistent with the result from previous work on uniform gain if we set $a = 1$ [16].

A.3 The effective dimension $D(C)$

We next derive expressions for the first few moments of c and subsequently the participation ratio. Recall the moment generating function $M_c(z)$ and its relation with the Cauchy transform:

$$M_c(z) = \sum_{n=1}^{\infty} \mu_n z^n, \quad z(M_c(z) + 1) = G_c(z^{-1}).$$

Then Eq. (S8) becomes

$$2M_c(z) + \varphi\left[\frac{1}{1 + g^2 z^{-1} M_c(z) + az^{-1/2}} + \frac{1}{1 + g^2 z^{-1} M_c(z) - az^{-1/2}}\right] = 0,$$

i.e.

$$\frac{M_c(z)}{1 + g^2 z^{-1} M_c(z)} + \varphi\left(\frac{1}{[1 + g^2 z^{-1} M_c(z)]^2 - z^{-1}a^2}\right) = 0,$$

Differentiating with respect to z once and twice and setting $z = 0$ gives us the first and second moments of c

$$\mu_1 = \frac{\varphi(a^{-2})}{1 - g^2\varphi(a^{-2})}, \quad \mu_2 = \frac{\varphi(a^{-4})}{[1 - g^2\varphi(a^{-2})]^4}.$$

The participation ratio can then be calculated to be

$$D(C) = \frac{\mu_1^2}{\mu_2} = \frac{\varphi(a^{-2})^2}{\varphi(a^{-4})}[1 - g^2\varphi(a^{-2})]^2 = D\left(R^{2(1-1/p)}\right)\left(1 - g^2\mathbb{E}\left[R^{2(1-1/p)}\right]\right)^2, \tag{S9}$$

When $p = 2$, this simplifies to Eq. (3). This result also agrees with previous work for uniform gain if we set $R = I$ [16].

B Temporal and spatial sampling

The theory in the previous section was derived for the exact covariance matrix of the full network population. In practice, due to constraints of the task and recording time we only have the sample covariance matrix calculated from a finite number of trials or time bins (usually no more than an order of magnitude larger than the number of recorded units), and we only observe a small number of neurons ($\sim 10^2$) from a large underlying population ($> 10^4$). We refer to these limitations as temporal and spatial sampling respectively and shall next examine their implications for the estimate of the effective dimension $D(C)$.

B.1 Temporal sampling

The effect of temporal sampling has been studied thoroughly before. Previous work has derived the relation between the covariance eigenspectrum of the original and time-sampled covariance matrix [16, 104], which also applies to Eq. (S8), but we focus on its effect on the effective dimension here. Let $\hat{C}$ be the sample

covariance matrix. Again since $D(C)$ is normalized, we can without loss of generality assume $k_0 = 1$. We have

$$\hat{C} = \frac{1}{T} X X^T, \quad X \sim \mathcal{N}(0, C).$$

This is equivalent to

$$X = (R^{1/p-1} - W)^{-1} Z, \quad Z_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/T).$$

Then

$$\hat{C} = (R^{1/p-1} - W)^{-1} Z Z^T (R^{1/p-1} - W)^{-T}.$$

This matrix has the same nonzero eigenvalues as $\check{C}Y$, where

$$\check{C} := (R^{1/p-1} - W)^{-T} (R^{1/p-1} - W)^{-1}, \quad Y := ZZ^T.$$

Note that the nonzero eigenvalues of $\check{C}$ is again the same as those of C, and Y is a free Poisson element. Since $\check{C}$ and Y are independent and Y is orthogonally invariant, the product $\check{C}Y$ is free [34]. Then we can calculate [16]

$$\mu_1(\hat{C}) = \mu_1(C), \quad \mu_2(\hat{C}) = \mu_2(C) + \alpha \mu_1(C)^2,$$

where $\alpha = N/T$ measures the extent of temporal sampling. Then we have

$$D(\hat{C}) = \frac{\mu_1(\hat{C})^2}{\mu_2(\hat{C})} = \frac{\mu_1(C)^2}{\mu_2(C) + \alpha \mu_1(C)^2} = \frac{1}{1/D(C) + \alpha}.$$

A larger α (i.e. a smaller number of temporal samples T) reduces the effective dimension of the sample covariance matrix $D(\hat{C})$. Intuitively, a limited amount of samples prevents the system from fully exploring its dimensions.

We can revert the $D(\hat{C})$ we calculate from data to the original $D(C)$ by

$$D(C) = \frac{1}{1/D(\hat{C}) - \alpha}. \tag{S10}$$

Note that we always normalize the participation ratio by N instead of T, which is different from the convention used in [16] when $\alpha > 1$.

B.2 Spatial sampling

Different from the uniform gain case where spatial sampling can be formulated as another free product problem [16], here this is no longer possible because C is not orthogonally invariant due to heterogeneous gain L. However, we can still make a reasonable approximation when the size of the sampled subpopulation is small compared to the whole population. Let $N = N_1 + N_2$ where N_1 the number of observed neurons and N_1 the unobserved ones. Let $f := N_1/N_2$ and assume $0 < f \ll 1$. Recall that after linearization (Eq. (S3)) we have

$$h = L(Wh + \xi), \quad h := r - r_0.$$

This can be written as

$$\begin{aligned} h_1 =& L_1(W_{11} h_1 + W_{12} h_2 + \xi_1), \\ h_2 =& L_2(W_{21} h_1 + W_{22} h_2 + \xi_2), \end{aligned} \tag{S11}$$

where we have partitioned the vectors and matrices into blocks according to the observed and unobserved population. Using the same notation for the partition, we have that C_{11} and C_{22} are the covariance matrices of the sampled and unobserved population respectively. If $f \ll 1$ and the sampling is completely random, we may assume that the term $W_{21} h_1$ is much smaller than $W_{22} h_2$ and ignore it. That is, we are ignoring the contribution of the sampled population's input to the unobserved population. Then the covariance matrix of the sampled population is

$$C_{11} = (L_1^{-1} - W_{11})^{-1} (W_{12} C_{22} W_{12}^T + I)(L_1^{-1} - W_{11})^{-T},$$

which has the same nonzero eigenvalues as

$$C_1(C_2 + I),$$

where we have written

$$C_1 := (L_1^{-1} - W_{11})^{-T} (L_1^{-1} - W_{11})^{-1}, \quad C_2 := W_{12} C_{22} W_{12}^T.$$

Notice that C_1 has the same nonzero eigenvalues as $(L_1^{-1} - W_{11})^{-1}(L_1^{-1} - W_{11})^{-T}$, which is just the covariance matrix of the observed population if it were isolated from the rest of the population, and C_{22} is just the covariance matrix of the unobserved population when we ignore the input of the observed population (i.e., the $W_{21}h_1$ term in (S11)). The behavior of these matrices is known from section A. C_2 has the same nonzero eigenvalues as $C_{22}D$, where $D := W_{12}^T W_{12}$. Since D is orthogonally invariant and independent of C_{22}, their product is free. Similarly, the product of C_1 and $C_2 + I$ is also free. This allows us to derive a theory for the moments of C_{11}, although it is an approximation that is only valid when f is small.

Recall the formulae for any free variables a, b

$$\mu_1(ab) = \mu_1(a)\mu_1(b),$$

$$\mu_2(ab) = \mu_2(a)\mu_1(b)^2 + \mu_1(a)^2[\mu_2(b) - \mu_1(b)^2].$$

This gives us

$$\mu_1(C_{11}) = \mu_1(C_1)[1 + \mu_1(C_2)],$$
$$\mu_2(C_{11}) = \mu_2(C_1)[1 + \mu_1(C_2)]^2 + \mu_1(C_1)^2[\mu_2(C_2) - \mu_1(C_2)^2].$$

Then

$$D(C_{11}) = \frac{\mu_1(C_{11})^2}{\mu_2(C_{11})} = \frac{1}{\frac{1}{D(C_1)} + \frac{\kappa_2(C_2)}{[1+\mu_1(C_2)]^2}},$$

where κ denotes the cumulants. For simplicity, we assume $p = 2$ in Eq. (2) in the following. Then

$$D(C_1) = D(R_1)(1 - g_1^2\bar{R}_1)^2, \quad D(C_{22}) = D(R_2)(1 - g_2^2\bar{R}_2)^2,$$

where g_1, g_2 are rescaled versions of g. Recall that the elements of W are drawn from $\mathcal{N}(0, g^2/N)$. If we want to apply Eq. (3), we need to make sure that the "g" here is with respect to the correct population size. That is,

$$\frac{g^2}{N} = \frac{g_1^2}{N_1} = \frac{g_2^2}{N_2}.$$

Using $N = N_1 + N_2, N_1 = fN_2$, we have

$$g_1^2 = fg_2^2, \quad g_2^2 = \frac{g^2}{1+f}.$$

We can also calculate

$$\mu_1(C_2) = \frac{1}{f}\mu_1(C_{22})\mu_1(D).$$

Here the $1/f$ term is to make sure that we only take into account the nonzero eigenvalues of $C_{22}D$. D is a free Poisson variable with parameter $\lambda = 1/f$ with a scaling factor $fg^2/(1+f)$. The unscaled free Poisson variable follows the Marchenko-Pastur law and has $\mu_1 = 1, \kappa_2 = \lambda$. Then

$$\mu_1(D) = \frac{fg^2}{1+f}, \quad \kappa_2(D) = \frac{fg^4}{(1+f)^2}.$$

We thus have

$$\mu_1(C_2) = \frac{g^2}{1+f}\mu_1(C_{22}),$$

$$\mu_2(C_2) = \frac{1}{f}\mu_2(C_{22}D) = \frac{1}{f}[\mu_2(C_{22})\mu_1(D)^2 + \mu_1(C_{22})^2\kappa_2(D)] = \frac{fg^4}{(1+f)^2}\mu_2(C_{22}) + \frac{g^4}{(1+f)^2}\mu_1(C_{22})^2,$$

$$\kappa_2(C_2) = \mu_2(C_2) - \mu_1(C_2)^2 = \frac{fg^4}{(1+f)^2}\mu_2(C_{22}).$$

Recall

$$\mu_1(C_{22}) = \frac{\bar{R}_2}{1 - g_2^2\bar{R}_2}, \quad \mu_2(C_{22}) = \frac{\mathbb{E}[R_2^2]}{(1 - g_2^2\bar{R}_2)^4}.$$

Putting everything together, we have

$$D(C_{11}) = \frac{1}{\frac{1}{D(C_1)} + \frac{fg_2^4\mu_2(C_{22})}{[1+g_2^2\mu_1(C_{22})]^2}} = \frac{D(R_1)}{\frac{1}{(1-g_1^2\bar{R}_1)^2} + f\frac{D(R_1)}{D(R_2)}\left(\frac{1}{1-g_2^2\bar{R}_2} - 1\right)^2}. \tag{S12}$$

Observe in the denominator that the participation ratios of R_1 and R_2 roughly cancel each other if the sampling is completely random. Then the effective dimension of the observed population is again proportional

to its firing rate uniformity, but the coefficient of this relation now depends on the proportion of neurons sampled from the full population as well as the connectivity and activity of the unobserved population which can depend heavily on which neurons are being recorded, making it difficult to interpret the slope of the $[D(R), D(C)]$ plot from neuronal data. As spatial sampling become more drastic, i.e. as f gets smaller, the effective dimension $D(C_{11})$ becomes larger and eventually approaches $D(R_1)$ when $f \to 0$. (Note that $g_1 \to 0$ as $f \to 0$.)

We verified Eq. (S12) using simulation (Fig. S4). The theory agrees well with simulation when f is small, and captures the trend as f gets larger. When we record ~ 100 neurons from an underlying population of $> 10^4$ neurons, we have $f < 0.01$, which is well within the range where our approximation here is valid.

C Implications for decoder performance in discrimination tasks

In this section, we explore how firing rate uniformity $D(R)$ affects neuronal coding in the context of fine discrimination tasks. Consider two stimulus responses with means $r_1, r_2 \in \mathbb{R}^N$ and covariance matrices $C_1, C_2 \in \mathbb{R}^{N^2}$ which are related by Eq. (1) (where $R_i = \mathrm{diag}(r_i), i = 1, 2$). The mean responses have population average $\bar{R}$ and uniformity $D(R)$. The discriminability index d' is defined by [97]

$$(d')^2 = 2(r_2 - r_1)^T (C_1 + C_2)^{-1} (r_2 - r_1).$$

Consider that r_1 and r_2 lie on the simplex

$$\Delta^N = \left\{ (x_1, \ldots, x_N) \in \mathbb{R}^N : \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{R}, x_j \geq 0, j = 1, \ldots, N \right\}.$$

That is, the firing rates are nonnegative and have mean $\bar{R}$. Then the set of points on this simplex with a certain $D(R)$ is just the intersection of Δ^N and a $(N-1)$-sphere of radius $\bar{R}\sqrt{N/D(R)}$ (which we denote with S_0). The intersection $S := S_0 \cap \Delta^N$ is codimension 2 manifold that is a subset of a $(N-2)$-sphere of radius $\sqrt{N\mathbb{V}[R]}$, where $\mathbb{V}[R]$ is the variance of firing rate across neurons (Fig. 5a).

Let us take a brief detour and look at the topological structure of the manifold S. When $D(R)$ is close to 1, $S_0 \subset \Delta^N$ and so $S = S_0$ is just an $(N-2)$-sphere which is $(N-3)$-connected. When $D(R)$ decreases past $(N-1)/N$, S_0 intersects with the $(N-2)$-faces of Δ^N, creating a $(N-2)$-dimensional hole and making S_0 $(N-4)$-connected. As $D(R)$ gets even smaller, S_0 intersects with faces of Δ^N that are of lower and lower dimensions. Observe that for the interval of $D(R)$ where S_0 goes from intersecting the centers of the M-faces to intersecting the centers of the $(M-1)$-faces of Δ^N, the problem of the connectedness of S boils down to understanding the situation on each M-face, which is again how the intersection of a $(M-1)$-sphere with the boundaries of a M-simplex affects its connectivity. Then it is easy to show by induction that the homotopical connectivity of S reduces by 1 each time it starts to intersect with faces of Δ^N of a lower dimension (beginning at the centers of each face) as $D(R)$ decreases. The center of an M-face of Δ^N is just the center of mass of $M+1$ coordinate vertices, which has $M+1$ equal nonzero coordinates and $N-M-1$ zero coordinates and thus has a $D(R)$ value of $(M+1)/N$. Then we can summarize

$$S \text{ is } (K-2)\text{-connected when } \quad \frac{K}{N} < D(R) < \frac{K+1}{N}, \quad K = 1, \ldots, N-1. \tag{S13}$$

In particular, S is disconnected with N components when $D(R) < 2/N$.

Going back to the problem of discriminating between two stimulus responses, we see that a smaller $D(R)$ implies both a smaller $D(C)$ according to Eq. (3) and a larger space S to code in, which allows for larger distance $v := r_2 - r_1$ between the stimulus means, which intuitively are both beneficial to decoding. Another perspective is that given a lower bound on d', a smaller $D(R)$ allows us to pack more stimuli into the space S without affecting decoding performance.

Continuing with our assumption that R and W are independent, we assume that C is independent of v. Since this is a fine discrimination task, r_1 is close to r_2. In addition, the covariance matrices are smooth functions of the mean responses (Eq. (1)), so C_1 and C_2 must be close as well. We thus assume for simplicity that $C := C_1 = C_2$, but will allow $C_1 \neq C_2$ in simulations. Denote the eigenvalue decomposition of C as $C = U^T \Lambda U$, where U is an orthogonal matrix. Since C and v are independent, $u := Uv$ is a completely random rotation of v which follows uniform distribution on the sphere of radius $|v|$. Then

$$(d')^2 = \sum_{i=1}^{N} \lambda_i^{-1} u_i^2 \approx |v|^2 \mathbb{E}[\lambda^{-1}],$$

where we used $|u| = |v|$. Let us examine the two factors $|v|^2$ and $\mathbb{E}[\lambda^{-1}]$ separately.

For $|v|^2$, since the size of S grows as $D(R)$ becomes smaller, it is reasonable to assume that $|v|$ grows linearly with the radius of S_0. Then we have $|v|^2 \propto \mathbb{V}[R] = 1/D(R) - 1$, which increases as $D(R)$ gets smaller.

For $\mathbb{E}[\lambda^{-1}]$, we need to understand the spectrum of $P := C^{-1}$. We are interested in $G_{T^2}(z)$ where T is as in Eq. (S6). Recall

$$G_T(z) = zG_{T^2}(z^2).$$

Then

$$zG_P(z^2) = \varphi\left[\frac{1}{z - g^2 zG_P(z^2) - (z - g^2 zG_P(z^2))^{-1}a^2}\right],$$

$$2zG_P(z) = \varphi\left[\frac{1}{1 - g^2 G_P(z) + z^{-1/2}a} + \frac{1}{1 - g^2 G_P(z) - z^{-1/2}a}\right]$$

Recall the moment generating function $M_P(z)$ and its relation with the Cauchy transform:

$$M_P(z) = \sum_{n=1}^{\infty} \mu_n z^n, \quad z(M_P(z) + 1) = G_P(z^{-1}).$$

Then

$$\frac{M_P(z) + 1}{1 - g^2 z(M_P(z) + 1)} = \varphi\left[\frac{1}{[1 - g^2 z(M_P(z) + 1)]^2 - za^2}\right].$$

Differentiating with respect to z gives us

$$\mathbb{E}[\lambda^{-1}] = g^2 + \varphi(a^2) = g^2 + \mathbb{E}(R^{-1}).$$

To be more concrete, we assume r_1 and r_2 are drawn from a log-normal distribution $\mathcal{LN}(\mu, \sigma^2)$. Recall that for $X \sim \mathcal{LN}(\mu, \sigma^2)$, we have

$$\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \mathbb{E}[X^2] = \exp(2(\mu + \sigma^2)), \quad D(X) = \exp(-\sigma^2).$$

Then

$$\sigma^2 = -\log(D(R)), \quad \mu = \log(\bar{R}) - \sigma^2/2.$$

When $R \sim \mathcal{LN}(\mu, \sigma^2)$, we have $\mathbb{E}(R^{-1}) = \exp\left(-\mu + \sigma^2/2\right)$. Then

$$\mathbb{E}[\lambda^{-1}] = g^2 + \frac{1}{\bar{R}D(R)},$$

which also increases as $D(R)$ gets smaller.

Combining the previous sections, we have

$$(d')^2 \approx |v|^2 \left(g^2 + \frac{1}{\bar{R}D(R)}\right) \propto \left(\frac{1}{D(R)} - 1\right)\left(g^2 + \frac{1}{\bar{R}D(R)}\right). \tag{S14}$$

The two factors embody the two ways a decrease in $D(R)$ can benefit decoding: by enlarging the coding space and lowering the effective dimension of variability. This result is verified in simulations (Fig. 5c).

D Supplementary figures

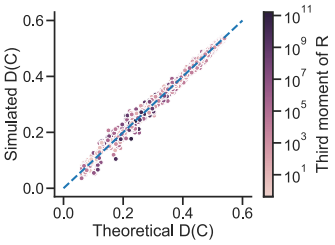

Fig. S1 Verification that Eq. (3) does not depend on the exact distribution of firing rates R beyond $D(R)$. Here we generated the firing rates R by randomly choosing parameters for the generalized gamma distribution [105], which has many common nonnegative continuous distributions as special cases or limiting cases including the gamma distribution, Weibull distribution, the half-normal distribution, and the log-normal distribution which was used in Figs. 1d and 2a. $D(C)$ is simulated according to Eq. (1) with $g = 0.15$ and $\bar{R} = 10$. The dashed line is the identity line. It can be seen that our theory predicts $D(C)$ regardless of the specific distribution of R beyond $D(R)$, which varies dramatically in our simulation as demonstrated by the wide range of third moments of R.

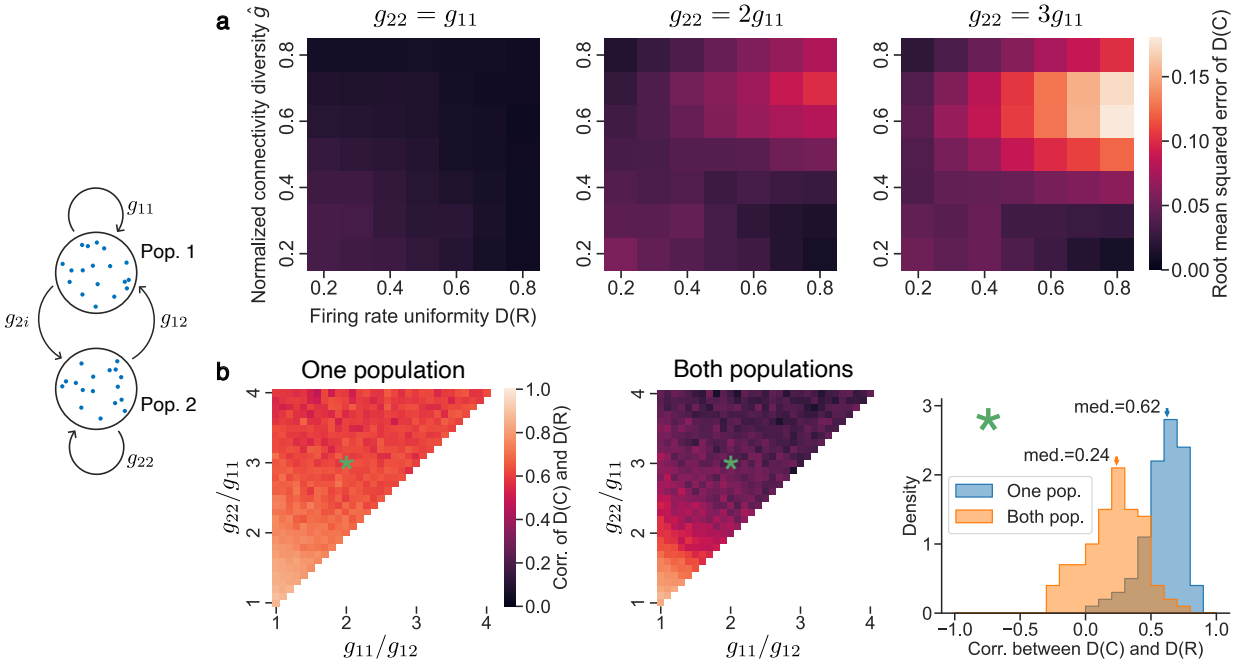

Fig. S2 Further simulation results on the two-population model in Fig. 1d. (**a**) The magnitude of error in predicting $D(C)$ using Eq. (3) as $D(R)$ and $\hat{g}$ vary. (**b**) $[D(R), D(C)]$ correlation. The overall variance of the elements of W is held constant ($\hat{g} = 0.6$) while g_{22}/g_{11} and g_{11}/g_{12} are varied. We also keep $g_{21} = g_{12}$. To calculate the $[D(R), D(C)]$ correlation, 20 $D(R)$ values was uniformly sampled in the range $[0.3, 0.7]$ to match neuronal data. This process was repeated 100 times, the averages from which are plotted on the heatmaps. The heatmap on the left shows the result for each population (averaged across the two populations) while the one on the right shows the case when the activity from the two populations are combined. An example distribution of the $[D(R), D(C)]$ correlation across realizations of W and R when g_{22}/g_{11} and g_{11}/g_{12} are at the green asterisk is shown on the right.

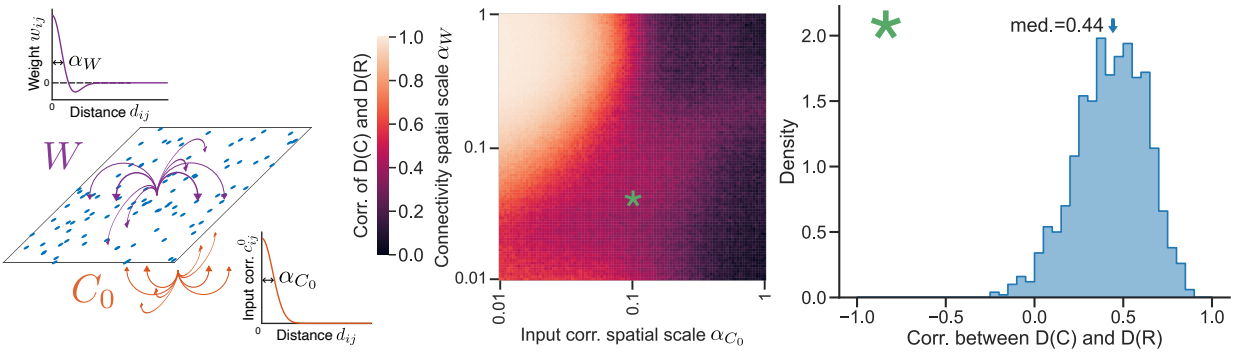

Fig. S3 Simulations with spatially dependent connectivity W and input noise covariance C_0. Neurons are randomly placed on a plane with those close to each other having both larger connection weights and external input correlations. α_W and α_{C_0} quantify the widths of the connectivity profile and the input correlation profile respectively. The simulated population size is 100. Middle: heatmap of the $D(C)$–$D(R)$ correlation as α_W and α_{C_0} change. Note that the axes are in log scale. Right: example distribution of $D(C)$–$D(R)$ correlation across realizations of W and C_0 when α_W and α_{C_0} are at the green asterisk in the heatmap. As can be seen in the heatmap, there is a large range of parameters that produce a moderate level of correlation between $D(C)$ and $D(R)$ comparable to data (Fig. 3d, i), which is likely to cover typical values of real neuronal circuits.

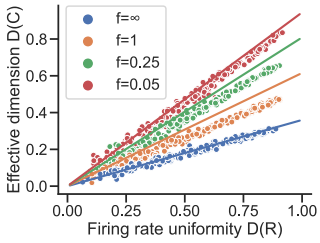

Fig. S4 The effect of spatial sampling. The plot shows $D(C)$ and $D(R)$ of the recorded population. f is the ratio of the size of the recorded population and the unobserved population. $f = \infty$ corresponds to the case where there is no spatial sampling and all neurons in the network are observed. The dots are from simulation and the lines are the theoretical predictions from Eq. (S12). The population size is 1000.

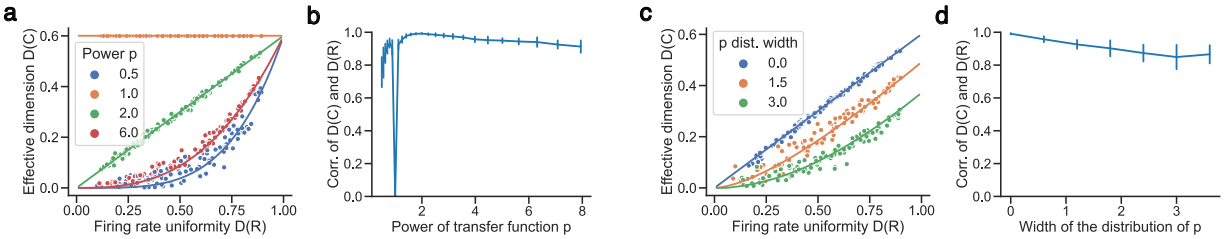

Fig. S5 The effect of different transfer functions. All simulations have population size 1000. (**a**) $[D(R), D(C)]$ plot when the power p in Eq. (2) varies. (**b**) The correlation between $D(C)$ and $D(R)$ as p varies. $D(R)$ was sampled in the range $[0.3, 0.7]$ to match neuronal data. (**c**) Same as (a) but for the heterogeneity of p. The power p for each neuron is randomly assigned according to a uniform distribution centered at 2 with different widths (meaning the distance between the left and right edges of the uniform distribution). (**d**) Same as (b) but for heterogeneity of p.

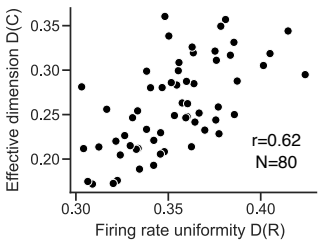

Fig. S6 Data recorded from M1 of another human participant as Fig. 3k.

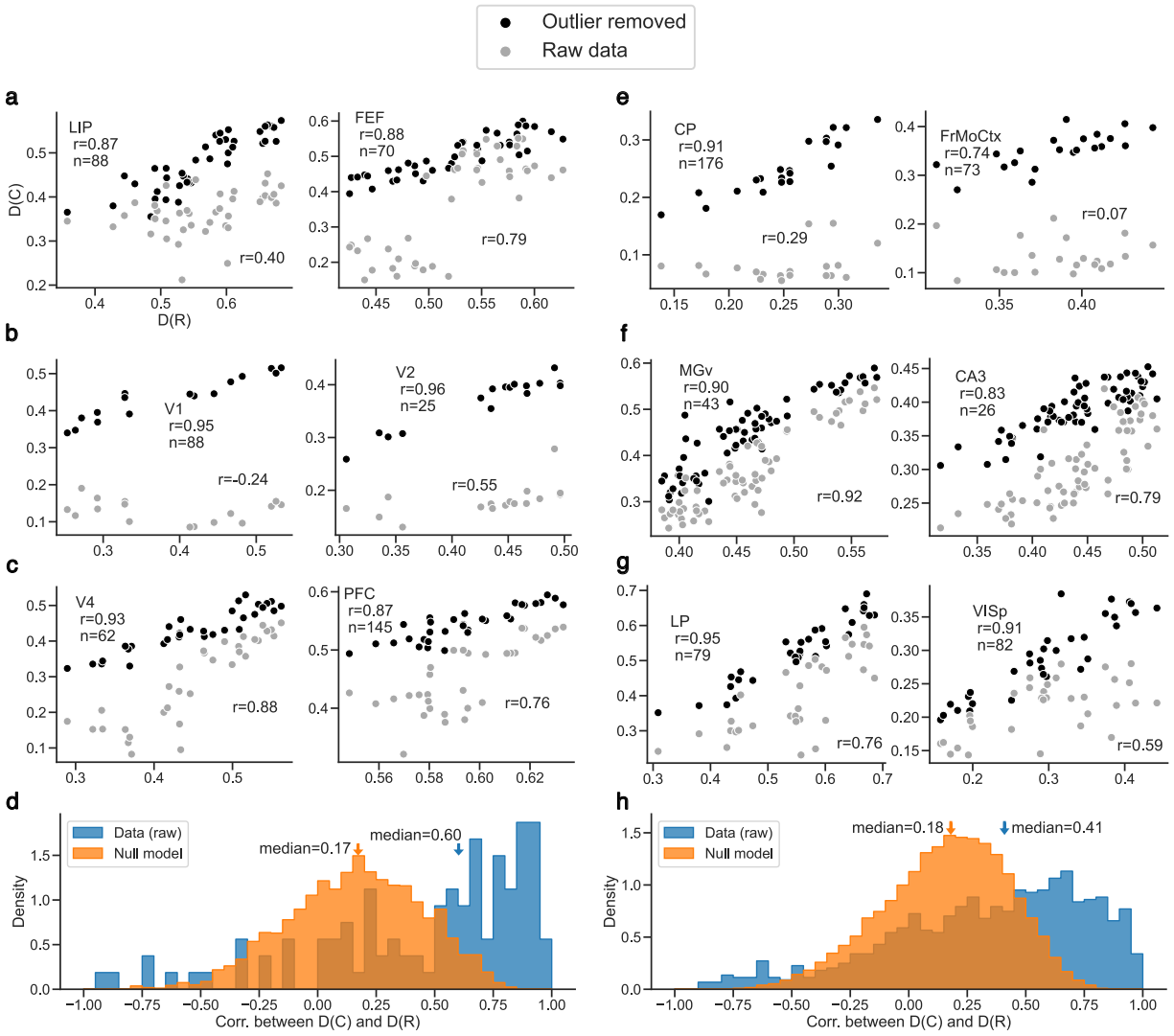

Fig. S7 The same as Fig. 3, but compared with the raw data where the outlier eigenvalues are not removed. (a-d) correspond to Fig. 3a-d and (e-h) correspond to Fig. 3f-i. Note that Fig. 3e and j are already the original raw data, and no outlier was identified when producing Fig. 3k.

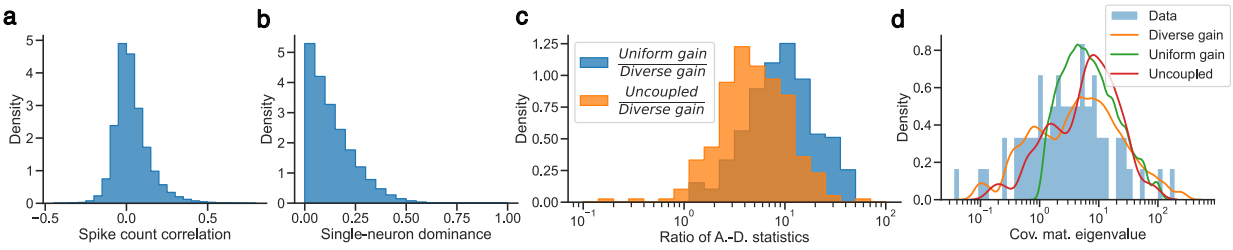

Fig. S8 All data in this figure come from dataset [42] and the distributions are over all brain areas and all sessions. (**a**) Distribution of spike count correlation. (**b**) Distribution of the measure for single-neuron dominance of eigenvectors of covariance matrices. Let v be an eigenvector and let v_1 and v_2 be the elements of v with the largest and second largest absolute values. The measure is defined as $(|v_1| - |v_2|)/|v_1|$, with a value close to one signifying the dominance of one neuron over the rest in this mode and the opposite showing that this mode is shared among multiple neurons. The fact that the histogram concentrates near 0 suggests that there is no correspondence between eigenvectors and neurons, showing that network activity is not a perturbation to uncoupled neurons in neuronal data. (**c**) The distribution of the ratios of the Anderson-Darling (A.-D.) statistics between different fitting methods of the covariance matrix eigenvalue distribution. The A.-D. statistics quantifies the quality of the fit, with a smaller value being better (see Methods). It can be seen that our model (i.e. the "diverse gain" model) almost always fits much better than the "uniform gain" model [16] from previous work and the naive "uncoupled" Poisson-like neurons model which simply scales the firing rate distribution to match the eigenvalue distribution, as the ratios are much larger than 1. (**d**) An example showing the performance of different fitting methods.

References

[1] Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).

[2] Moran, P. A. The statistical analysis of the canadian lynx cycle. *Australian Journal of Zoology* **1**, 291–298 (1953).

[3] Grenfell, B. *et al.* Noise and determinism in synchronized sheep dynamics. *Nature* **394**, 674–677 (1998).

[4] Font-Muñoz, J. S. *et al.* Advection by ocean currents modifies phytoplankton size structure. *Journal of the royal society interface* **14**, 20170046 (2017).

[5] Yang, X., Wang, K. & Shamma, S. A. Auditory representations of acoustic signals. *IEEE transactions on information theory* **38**, 824–839 (1992).

[6] Ermentrout, G. B. & Cowan, J. D. A mathematical theory of visual hallucination patterns. *Biological cybernetics* **34**, 137–150 (1979).

[7] Billock, V. A. & Tsou, B. H. Elementary visual hallucinations and their relationships to neural pattern-forming mechanisms. *Psychological bulletin* **138**, 744 (2012).

[8] Cavagna, A. *et al.* Marginal speed confinement resolves the conflict between correlation and control in collective behaviour. *Nature Communications* **13**, 2315 (2022).

[9] Marder, E. & Bucher, D. Central pattern generators and the control of rhythmic movements. *Current biology* **11**, R986–R996 (2001).

[10] Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology* **2**, 38 (2014).

[11] Li, S. & Wager, S. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics* **50**, 2334–2358 (2022).

[12] Miller, M. B. & Bassler, B. L. Quorum sensing in bacteria. *Annual Reviews in Microbiology* **55**, 165–199 (2001).

[13] Markram, H. *et al.* Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**, 456–492 (2015).

[14] Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods* **15**, 805–815 (2018).

[15] Mejias, J. & Longtin, A. Optimal heterogeneity for coding in spiking neural networks. *Physical Review Letters* **108**, 228102 (2012).

[16] Hu, Y. & Sompolinsky, H. The spectrum of covariance matrices of randomly connected recurrent neuronal networks with linear dynamics. *PLoS computational biology* **18**, e1010327 (2022).

[17] Gast, R., Solla, S. A. & Kennedy, A. Neural heterogeneity controls computations in spiking neural networks. *Proceedings of the National Academy of Sciences* **121**, e2311885121 (2024).

[18] Dahmen, D. *et al.* Strong coupling and local control of dimensionality across brain areas. *Biorxiv* 2020–11 (2020).

[19] Hutt, A., Rich, S., Valiante, T. A. & Lefebvre, J. Intrinsic neural diversity quenches the dynamic volatility of neural networks. *Proceedings of the National Academy of Sciences* **120**, e2218841120 (2023).

[20] Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nature reviews neuroscience* **7**, 358–366 (2006).

[21] Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. Correlations and neuronal population information. *Annual review of neuroscience* **39**, 237–256 (2016).

[22] Panzeri, S., Moroni, M., Safaai, H. & Harvey, C. D. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience* **23**, 551–567 (2022).

[23] Ferster, D. & Miller, K. D. Neural mechanisms of orientation selectivity in the visual cortex. *Annual review of neuroscience* **23**, 441–471 (2000).

[24] Ocker, G. K. *et al.* From the statistics of connectivity to the statistics of spike times in neuronal networks. *Current opinion in neurobiology* **46**, 109–119 (2017).

[25] Treves, A. & Rolls, E. T. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems* **2**, 371 (1991).

[26] Willmore, B. & Tolhurst, D. J. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems* **12**, 255 (2001).

[27] Simpson, E. H. Measurement of diversity. *nature* **163**, 688–688 (1949).

[28] Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).

[29] Ginzburg, I. & Sompolinsky, H. Theory of correlations in stochastic neural networks. *Physical review E* **50**, 3171 (1994).

[30] Grytskyy, D., Tetzlaff, T., Diesmann, M. & Helias, M. A unified view on weakly correlated recurrent networks. *Frontiers in computational neuroscience* **7**, 131 (2013).

[31] Buzsáki, G. & Mizuseki, K. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience* **15**, 264–278 (2014).

[32] Priebe, N. J. & Ferster, D. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron* **57**, 482–497 (2008).

[33] Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M. & Miller, K. D. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron* **98**, 846–860 (2018).

[34] Mingo, J. A. & Speicher, R. *Free probability and random matrices* Vol. 35 (Springer, 2017).

[35] Yuste, R., Cossart, R. & Yaksi, E. Neuronal ensembles: Building blocks of neural circuits. *Neuron* (2024).

[36] Rossi, L. F., Harris, K. D. & Carandini, M. Spatial connectivity matches direction selectivity in visual cortex. *Nature* **588**, 648–652 (2020).

[37] Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E. & Doiron, B. The spatial structure of correlated neuronal variability. *Nature neuroscience* **20**, 107–114 (2017).

[38] Bittner, S. R. *et al.* Population activity structure of excitatory and inhibitory neurons. *PloS one* **12**, e0181773 (2017).

[39] Zandvakili, A. & Kohn, A. Coordinated neuronal activity enhances corticocortical communication. *Neuron* **87**, 827–839 (2015).

[40] Snyder, A. C., Byron, M. Y. & Smith, M. A. A stable population code for attention in prefrontal cortex leads a dynamic attention code in visual cortex. *Journal of Neuroscience* **41**, 9163–9176 (2021).

[41] Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).

[42] Allen Institute MindScope Program. Allen brain observatory – neuropixels visual coding [dataset] (2019). Available from brain-map.org/explore/circuits.

[43] Allen Institute MindScope Program. Allen brain observatory – neuropixels visual behavior [dataset] (2022). Available from brain-map.org/explore/circuits.

[44] Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).

[45] Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).

[46] Mazzucato, L., Fontanini, A. & La Camera, G. Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience* **10**, 11 (2016).

[47] Bennett, C., Arroyo, S. & Hestrin, S. Subthreshold mechanisms underlying state-dependent modulation of visual responses. *Neuron* **80**, 350–357 (2013).

[48] Polack, P.-O., Friedman, J. & Golshani, P. Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience* **16**, 1331–1339 (2013).

[49] Sommers, H. J., Crisanti, A., Sompolinsky, H. & Stein, Y. Spectrum of large random asymmetric matrices. *Physical review letters* **60**, 1895 (1988).

[50] Rajan, K. & Abbott, L. F. Eigenvalue spectra of random matrices for neural networks. *Physical review letters* **97**, 188104 (2006).

[51] Wang, Z. *et al.* The geometry and dimensionality of brain-wide activity. *eLife* **14** (2025).

[52] Harris, I. D., Meffin, H., Burkitt, A. N. & Peterson, A. D. Effect of sparsity on network stability in random neural networks obeying dale's law. *Physical Review Research* **5**, 043132 (2023).

[53] Gudowska-Nowak, E., Nowak, M. A., Chialvo, D. R., Ochab, J. K. & Tarnowski, W. From synaptic interactions to collective dynamics in random neuronal networks models: critical role of eigenvectors and transient behavior. *Neural Computation* **32**, 395–423 (2020).

[54] Luo, F., Zhong, J., Yang, Y., Scheuermann, R. H. & Zhou, J. Application of random matrix theory to biological networks. *Physics Letters A* **357**, 420–423 (2006).

[55] Allesina, S. & Tang, S. The stability–complexity relationship at age 40: a random matrix perspective. *Population Ecology* **57**, 63–75 (2015).

[56] Gao, P. *et al.* A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv* 214262 (2017).

[57] Recanatesi, S., Ocker, G. K., Buice, M. A. & Shea-Brown, E. Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity. *PLoS computational biology* **15**, e1006446 (2019).

[58] Kao, T.-C., Sadabadi, M. S. & Hennequin, G. Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. *Neuron* **109**, 1567–1581 (2021).

[59] Trägenap, S., Whitney, D. E., Fitzpatrick, D. & Kaschube, M. The developmental emergence of reliable cortical representations. *Nature Neuroscience* 1–12 (2025).

[60] Shymkiv, Y., Hamm, J. P., Escola, S. & Yuste, R. Slow cortical dynamics generate context processing and novelty detection. *Neuron* (2025).

[61] Monsalve-Mercado, M. M. & Miller, K. D. The geometry of the neural state space of decisions. *bioRxiv* 2025–01 (2025).

[62] Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience* **17**, 1500–1509 (2014).

[63] Altan, E., Solla, S. A., Miller, L. E. & Perreault, E. J. Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. *PLoS computational biology* **17**, e1008591 (2021).

[64] Susman, L., Mastrogiuseppe, F., Brenner, N. & Barak, O. Quality of internal representation shapes learning performance in feedback neural networks. *Physical Review Research* **3**, 013176 (2021).

[65] Clark, D. G., Abbott, L. & Litwin-Kumar, A. Dimension of activity in random neural networks. *Physical Review Letters* **131**, 118401 (2023).

[66] Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Current opinion in neurobiology* **14**, 481–487 (2004).

[67] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).

[68] Failor, S. W., Carandini, M. & Harris, K. D. Visual experience orthogonalizes visual cortical stimulus responses via population code transformation. *Cell Reports* **44** (2025).

[69] Homann, J., Kim, H., Tank, D. W. & Berry, M. J. Passive exposure sparsifies neural activity in the primary visual cortex. *bioRxiv* 2021–11 (2021).

[70] Gdalyahu, A. *et al.* Associative fear learning enhances sparse network coding in primary sensory cortex. *Neuron* **75**, 121–132 (2012).

[71] Chu, M. W., Li, W. L. & Komiyama, T. Balancing the robustness and efficiency of odor representations during learning. *Neuron* **92**, 174–186 (2016).

[72] Kato, H. K., Gillet, S. N. & Isaacson, J. S. Flexible sensory representations in auditory cortex driven by behavioral relevance. *Neuron* **88**, 1027–1039 (2015).

[73] Ratzon, A., Derdikman, D. & Barak, O. Representational drift as a result of implicit regularization. *Elife* **12**, RP90069 (2024).

[74] Lim, S. *et al.* Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience* **18**, 1804–1810 (2015).

[75] Rumyantsev, O. I. *et al.* Fundamental bounds on the fidelity of sensory cortical coding. *Nature* **580**, 100–105 (2020).

[76] Ebrahimi, S. *et al.* Emergent reliability in sensory cortical coding and inter-area communication. *Nature* **605**, 713–721 (2022).

[77] Moreno-Bote, R. *et al.* Information-limiting correlations. *Nature neuroscience* **17**, 1410–1417 (2014).

[78] Kanitscheider, I., Coen-Cagli, R. & Pouget, A. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences* **112**, E6973–E6982 (2015).

[79] Doiron, B., Litwin-Kumar, A., Rosenbaum, R., Ocker, G. K. & Josić, K. The mechanics of state-dependent neural correlations. *Nature neuroscience* **19**, 383–393 (2016).

[80] Kanashiro, T., Ocker, G. K., Cohen, M. R. & Doiron, B. Attentional modulation of neuronal variability in circuit models of cortex. *Elife* **6**, e23978 (2017).

[81] Huang, C. *et al.* Circuit models of low-dimensional shared variability in cortical networks. *Neuron* **101**, 337–348 (2019).

[82] Asaad, W. F., Santhanam, N., McClellan, S. & Freedman, D. J. High-performance execution of psychophysical tasks with complex visual stimuli in matlab. *Journal of neurophysiology* **109**, 249–260 (2013).

[83] Hwang, J., Mitz, A. R. & Murray, E. A. Nimh monkeylogic: Behavioral control and data acquisition in matlab. *Journal of neuroscience methods* **323**, 13–21 (2019).

[84] Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology* **61**, 331–349 (1989).

[85] Hikosaka, O. & Wurtz, R. H. Modification of saccadic eye movements by gaba-related substances. i. effect of muscimol and bicuculline in monkey superior colliculus. *Journal of neurophysiology* **53**, 266–291 (1985).

[86] Mohan, K., Zhu, O. & Freedman, D. J. Interaction between neuronal encoding and population dynamics during categorization task switching in parietal cortex. *Neuron* **109**, 700–712 (2021).

[87] Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).

[88] Rossant, C. *et al.* Spike sorting for large, dense electrode arrays. *Nature neuroscience* **19**, 634–641 (2016).

[89] Downey, J. E. *et al.* A roadmap for implanting microelectrode arrays to evoke tactile sensations through intracortical microstimulation. *medRxiv* (2024).

[90] Bordenave, C. Eigenvalues of euclidean random matrices. *Random Structures & Algorithms* **33**, 515–532 (2008).

[91] Cowley, B. R. *et al.* Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex. *Neuron* **108**, 551–567 (2020).

[92] Dabrowska, P. A. *et al.* On the complexity of resting state spiking activity in monkey motor cortex. *Cerebral cortex communications* **2**, tgab033 (2021).

[93] Morales-Gregorio, A. *et al.* Neural manifolds in v1 change with top-down signals from v4 targeting the foveal region. *Cell Reports* **43** (2024).

[94] Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **25** (2012).

[95] Anderson, T. W. & Darling, D. A. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics* 193–212 (1952).

[96] Macke, J. H., Berens, P., Ecker, A. S., Tolias, A. S. & Bethge, M. Generating spike trains with specified correlation coefficients. *Neural computation* **21**, 397–423 (2009).

[97] Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *Journal of neurophysiology* **95**, 3633–3644 (2006).

[98] Hawkes, A. G. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **33**, 438–443 (1971).

[99] Pernice, V., Staude, B., Cardanobile, S. & Rotter, S. How structure determines correlations in neuronal networks. *PLoS computational biology* **7**, e1002059 (2011).

[100] Bair, W., Zohary, E. & Newsome, W. T. Correlated firing in macaque visual area mt: time scales and relationship to behavior. *Journal of Neuroscience* **21**, 1676–1697 (2001).

[101] Feinberg, J. & Zee, A. Non-gaussian non-hermitian random matrix theory: phase transition and addition formalism. *Nuclear Physics B* **501**, 643–669 (1997).

[102] Burda, Z. *Free products of large random matrices–a short review of recent developments*, Vol. 473, 012002 (IOP Publishing, 2013).

[103] Schuecker, J., Goedeke, S. & Helias, M. Optimal sequence memory in driven random networks. *Physical Review X* **8**, 041029 (2018).

[104] Burda, Z., Görlich, A., Jarosz, A. & Jurkiewicz, J. Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications* **343**, 295–310 (2004).

[105] Stacy, E. W. A generalization of the gamma distribution. *The Annals of mathematical statistics* 1187–1192 (1962).