Network dynamics underlying OFF responses in the auditory cortex

Giulio Bondanelli [1,3], Thomas Deneux [2], Brice Bathellier [2] Srdjan Ostojic [1]

[1] Laboratoire de Neurosciences Cognitives et Computationelles, Département d'études cognitives, ENS, PSL University, INSERM, Paris, France

[2] Départment de Neurosciences Intégratives et Computationelles (ICN), Institut des Neurosciences Paris-Saclay (NeuroPSI), UMR 9197 CNRS, Université Paris Sud, Gif-sur- Yvette, France

[3] Neural Computation Laboratory, Center for Human Technologies, Istituto Italiano di Tecnologia (IIT), Genoa, Italy

Abstract

Across sensory systems, complex spatio-temporal patterns of neural activity arise following the onset (ON) and offset (OFF) of stimuli. While ON responses have been widely studied, the mechanisms generating OFF responses in cortical areas have so far not been fully elucidated. We examine here the hypothesis that OFF responses are single-cell signatures of network dynamics and propose a network model that generates transient OFF responses through recurrent interactions. To test this model, we performed population analyses of two-photon calcium recordings in the auditory cortex of awake mice listening to auditory stimuli. We found that the network model accounts for the low-dimensional organisation of population responses and their global structure across stimuli, where distinct stimuli activate mostly orthogonal dimensions in the neural state-space. In contrast, a single-cell mechanism explains some prominent features of the data, but does not account for the structure across stimuli and trials captured by the network model.

Introduction

Neural responses within the sensory cortices are inherently transient. In the auditory cortex even the simplest stimulus, for instance a pure tone, evokes neural responses that strongly vary in time following the onset and offset of the stimulus. A number of past studies have reported a prevalence of ON- compared to OFF-responsive neurons in different auditory areas (Phillips et al., 2002; Luo et al., 2008; Fu et al., 2010; Pollak and Bodenhamer, 1981). As a result, the transient onset component has been long considered the dominant feature of auditory responses, and has been extensively studied across the auditory pathway (Liu et al., 2019b; Kuwada and Batra, 1999; Grothe et al., 1992; Guo and Burkard, 2002; Yu et al., 2004; Heil, 1997a,b), with respect to its neurophysiological basis and perceptual meaning (Phillips et al., 2002). In parallel, due to less evidence of OFF-responsive neurons in anaesthetized animals, OFF cortical responses have received comparably less attention. Yet, OFF responses have been observed in awake animals throughout the auditory pathway, and in the mouse auditory cortex they arise in 30% to 70% of the responsive neurons (Scholl et al., 2010; Keller et al., 2018; Joachimsthaler et al., 2014; Liu et al., 2019a; Sollini et al., 2018).

While the generation of ON responses has been attributed to neural mechanisms based on short-term adaptation, most likely inherited from the auditory nerve fibers (Phillips et al., 2002; Smith and Brachman, 1980, 1982), the mechanisms that generate OFF responses are more diverse and seem to be area-specific (Xu et al., 2014; Kopp-Scheinpflug et al., 2018). In subcortical regions, neurons in the dorsal cochlear nucleus (DCN) and in the superior paraolivary nucleus (SPN) of the brainstem nuclei may generate OFF responses by post-inhibitory rebound, a synaptic mechanism in which a neuron emits one or more spikes following the cessation of a prolonged hyperpolarizing current (Suga, 1964; Hancock and Voigt, 1999; Kopp-Scheinpflug et al., 2011). In the midbrain inferior colliculus (IC) and in the medial geniculate body (MGB) of the thalamus, OFF responses appear to be mediated by excitatory inputs from upstream areas and potentially boosted by a post-inhibitory facilitation mechanism (Kasai et al., 2012; Vater et al., 1992; Yu et al., 2004; He, 2003). Unlike in subcortical areas, OFF responses in auditory cortex do not appear to be driven by hyperpolarizing inputs during the presentation of the stimulus, since synaptic inhibition has been found to be only transient with respect to the stimulus duration (Qin et al., 2007; Scholl et al., 2010). The precise cellular or network mechanisms underlying transient OFF responses in cortical areas therefore remain to be fully elucidated.

Previous studies investigating the transient responses in the auditory system mostly adopted a single-neuron perspective (Henry, 1985; Scholl et al., 2010; Qin et al., 2007; He, 2002; Wang et al., 2005; Wang, 2007). However, in recent years, population approaches to neural data have proven valuable for understanding the role of transients dynamics in various cortical areas (Buonomano and Maass, 2009; Remington et al., 2018; Saxena and Cunningham, 2019). Work in the olfactory system has shown that ON and OFF responses encode the stimulus identity in the dynamical patterns of activity across the neural population (Mazor and Laurent, 2005; Stopfer et al., 2003; Broome et al., 2006; Friedrich and Laurent, 2001; Saha et al., 2017). In motor and premotor cortex, transient responses during movement execution form complex population trajectories (Churchland and Shenoy, 2007; Churchland et al., 2010, 2012) that have been hypothesised to be generated by a mechanism based on recurrent network dynamics (Shenoy et al., 2011; Hennequin et al., 2014; Sussillo et al., 2015; Stroud et al., 2018). In the auditory cortex, previous works have suggested a central role of the neural dynamics across large populations for the coding of different auditory features (Deneux et al., 2016; Saha et al., 2017; Lim et al., 2016), yet how these dynamics are generated remains an open question.

Leveraging the observation that the auditory cortex constitutes a network of neurons connected in a recurrent fashion (Linden and Schreiner, 2003; Oswald and Reyes, 2008; Oswald et al., 2009; Barbour and Callaway, 2008; Bizley et al., 2015), in this study we test the hypothesis that transient OFF responses are generated by a recurrent network mechanism broadly analogous to the motor cortex (Churchland et al., 2006; Hennequin et al., 2014). We first analyse OFF responses evoked by multiple auditory stimuli in large neural populations recorded using calcium imaging in the mouse auditory cortex (AC) (Deneux et al., 2016). These analyses identified three prominent features of the AC data: (i) OFF responses correspond to transiently amplified trajectories of population activity; (ii) for each stimulus, the corresponding trajectory explores a low-dimensional subspace; (iii) responses to different stimuli lie mostly in orthogonal subspaces. We then introduce a simple recurrent network model, and show that it can reproduce the three qualitative features of the AC data provided specific constraints on model components are met. These constraints directly lead to a set of new predictions, which we directly verify on the AC data. We finally compare the network model with a simplified single-cell model, and show that the single-cell model can reproduce the first two features of population dynamics in response to individual stimuli, but in contrast to the network model cannot capture the structure across stimuli and single trias.

Results

ON/OFF responses in auditory cortex reflect transiently amplified population dynamics

We analyse the population responses of 2343 cells from the auditory cortex of 3 awake mice recorded using calcium imaging techniques (data from Deneux et al. (2016)). The neurons were recorded while the mice passively listened to randomized presentations of different auditory stimuli. In this study we consider a total of 16 stimuli, consisting of two groups of intensity modulated UP- or DOWN-ramping sounds. In each group, there were stimuli with different frequency content (either 8 kHz pure tones or white noise sounds), different durations (1 or 2 s) and different intensity slopes (either 50-85 dB or 60-85 dB and reversed, see Table stimuli in Materials and methods).

We first illustrate the responses of single cells to the presentation of different auditory stimuli, focusing on the periods following the onset and offset of the stimulus. We observe that the activity of individual neurons to different stimuli was highly heterogeneous. In response to a single stimulus, we found individual neurons that were strongly active only during the onset of the stimulus (ON responses), or only during the offset (OFF responses), while other neurons in the population responded to both stimulus onset and offset, consistent with previous analyses (Deneux et al., 2016). Importantly, across stimuli some neurons in the population exhibited ON and/or OFF responses only when specific stimuli were presented, showing stimulus-selectivity of transient responses, while others strongly responded at the onset and/or at the offset of multiple stimuli (Fig. 1A).

Because of the intrinsic heterogeneity of single-cell responses, we examined the structure of the transient ON and OFF responses to different stimuli using a population approach (Buonomano and Maass, 2009; Saxena and Cunningham, 2019). The temporal dynamics of the collective response of all the neurons in the population can be represented as a sequence of states in a high-dimensional state space, in which the i-th coordinate corresponds to the firing activity $r_i(t)$ of the i-th neuron in the population. At each time point, the population response is described by a population activity vector $\mathbf{r}(t)$ which draws a *neural trajectory* in the state space (Fig. 1B).

To quantify the strength of the population transient ON and OFF responses, we computed the distance of the population activity vector $\mathbf{r}(t)$ from baseline (average firing rate before stimulus presentation), corresponding to the norm of the population activity vector $||\mathbf{r}(t) - \mathbf{r}_B||$ (Mazor and Laurent, 2005). This revealed that the distance from baseline computed during ON and OFF responses was larger than the distance computed for the state at the end of stimulus presentation (Fig. 1B). We refer to this feature of the population transient dynamics as the transient amplification of ON and OFF responses.

To examine what the transient amplification of ON and OFF responses implies in terms of stimulus decoding, we trained a simple decoder to classify pairs of stimuli that differed in their frequency content, but had the same intensity modulation and duration. We found that the classification accuracy was highest during the transient phases corresponding to ON and OFF responses, while it decreased at the end of stimulus presentation (Fig. 1C). This result revealed a robust encoding of the stimulus features during ON and OFF responses, as previously found in the locust olfactory system (Mazor and Laurent, 2005; Saha et al., 2017).

OFF responses rely on orthogonal low-dimensional subspaces

To further explore the structure of the neural trajectories associated with the population OFF response to different stimuli, we analysed neural activity using dimensionality reduction techniques (Cunningham and Yu, 2014). We focused in particular on the OFF responses within the period starting 50 ms before stimulus offset to 300 ms after stimulus offset.

By performing principal component analysis (PCA) independently for the responses to individual stimuli, we found that the dynamics during the transient OFF responses to individual stimuli explored only about five dimensions, as 80% of the variance of the OFF responses to individual stimuli was explained on average by the first 5 principal components (Fig. 2A; see Fig. S1 for cross-validated controls; note that, given the temporal resolution, the maximal dimensionality of responses to individual stimuli was 13). The projection of the low-dimensional OFF responses to each stimulus onto the first two principal components revealed circular activity patterns, where the population vector smoothly rotated between the two dominant dimensions (Fig. 2B).

A central observation revealed by the dimensionality reduction analysis was that the OFF response trajectories relative to stimuli with different frequency content spanned orthogonal low-dimensional subspaces. For instance, the response to the 8 kHz sound was poorly represented on the plane defined by the two

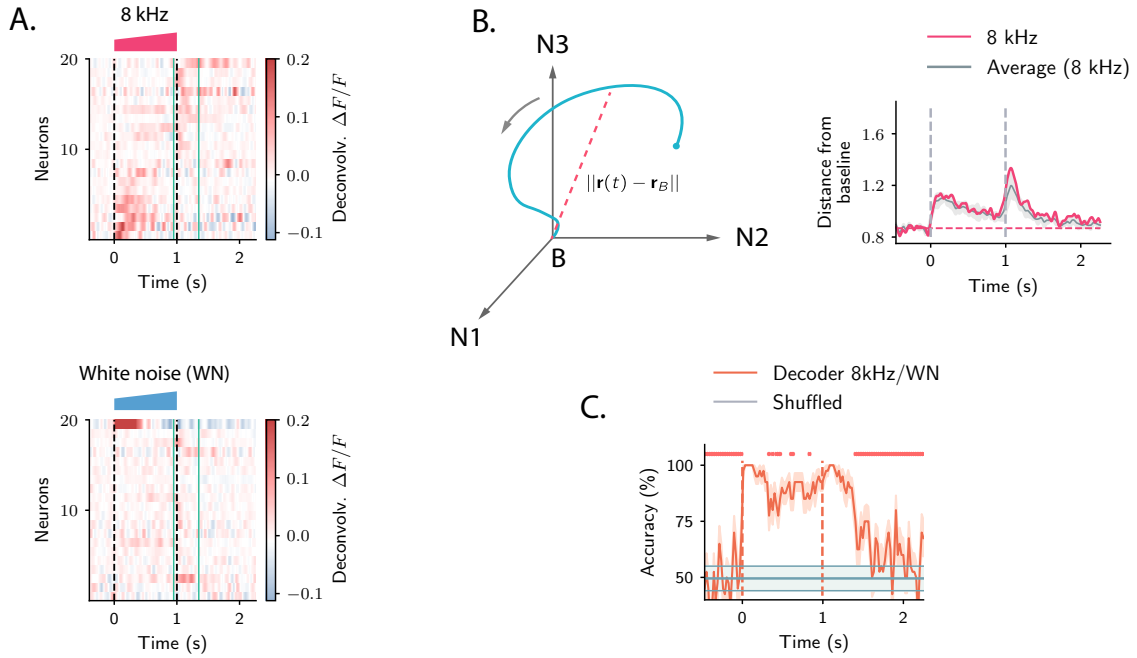

Figure 1: **Strong transient ON and OFF responses in auditory cortex of mice passively listening to different sounds and stimulus decoding. A.** *Top*: deconvolved calcium signals averaged over 20 trials showing the activity (estimated firing rate) of 20 out of 2343 neurons in response to a 8 kHz 1s UP-ramp with intensity range 60-85 dB. We selected neurons with high signal-to-noise ratios (ratio between the peak activity during ON/OFF responses and the standard deviation of the activity before stimulus presentation). Neurons were ordered according to the difference between peak activity during ON and OFF response epochs. *Bottom*: activity of the same neurons as in the top panel in response to a white noise sound with the same duration and intensity profile. In all panels dashed lines indicate onset and offset of the stimulus, green solid lines show the temporal region where OFF responses are analysed (from 50 ms before stimulus offset to 300 ms after stimulus offset). **B.** *Left*: cartoon showing the OFF response to one stimulus as a neural trajectory in the state space, where each coordinate represents the firing rate of one neuron (with respect to the baseline B). The length of the dashed line represents the distance between the population activity vector $\mathbf{r}(t)$ and the baseline firing rate $\mathbf{r}_B$. *Right*: the red trace shows the distance from baseline $||\mathbf{r}(t) - \mathbf{r}_B||$ computed for the population response to the 8 kHz sound in **A**. The grey trace shows the distance from baseline averaged over the 8 kHz sounds of 1 s duration (4 stimuli). Gray shading represents ± 1 standard deviation. The dashed horizontal line shows the average level of the distance $||\mathbf{r}(t) - \mathbf{r}_B||$ before stimulus presentation (even if baseline-subtracted responses are used, a value of the norm different from zero is expected because of noise in the spontaneous activity before stimulus onset). **C.** Accuracy of stimulus classification between a 8 kHz versus white noise UP-ramping sounds over time based on single trials (20 trials). The decoder is computed at each time step (spaced by ~ 50 ms) and accuracy is computed using leave-one-out cross-validation (LOOCV). Orange trace: average classification accuracy over the cross-validation folds. Orange shaded area corresponds to ± 1 standard error. The same process is repeated after shuffling stimulus labels across trials at each time step (gray trace and shading). The dashed horizontal lines represent the chance level. The red markers on the top indicate the time points where the average classification accuracy is lower than the maximum accuracy during the ON transient response ($P < 0.01$, two-tailed t-test).

principal component of the response to the white noise sound (Fig. 2B), and vice versa, showing that they evolved on distinct subspaces. To quantify the relationship between the subspaces spanned by the OFF responses to different stimuli, we proceeded as follows. We first computed the first 5 principal components for the OFF response to individual stimuli. Therefore, for each stimulus these dimensions define a 5-dimensional subspace. Then, for each pair of stimuli, we computed the relative orientation of the corresponding pair of subspaces, measured by the subspace overlap (Fig. 2D; see Materials and methods).

This approach revealed an interesting structure between the OFF response subspaces for different stimuli (Fig. 2D and Fig. S2). Stimuli with different frequency content evoked in general non-overlapping OFF responses, reflected in low values of the subspace overlap. Two clusters of correlated OFF responses emerged, corresponding to the 8 kHz UP-ramps and white noise UP-ramps of different durations and intensity.

Surprisingly, DOWN-ramps evoked OFF responses that were less correlated than UP-ramps also between sounds with the same frequency content.

The fact that most of the stimuli evoked non-overlapping OFF responses is reflected in the number of principal components that explain 80% of the variance for all OFF responses considered together, which is around 60 (Fig. 2C and Fig. S1). This number is in fact close to the number of dominant components of the joint response to all 16 stimuli (see Table stimuli) that we would expect if the responses to individual stimuli evolved on uncorrelated subspaces (given by #PC per stimulus $\times$ #stimuli ≈ 80). Notably this implies that, while the OFF responses to individual stimuli span low-dimensional subspaces, the joint response across stimuli shows high-dimensional structure.

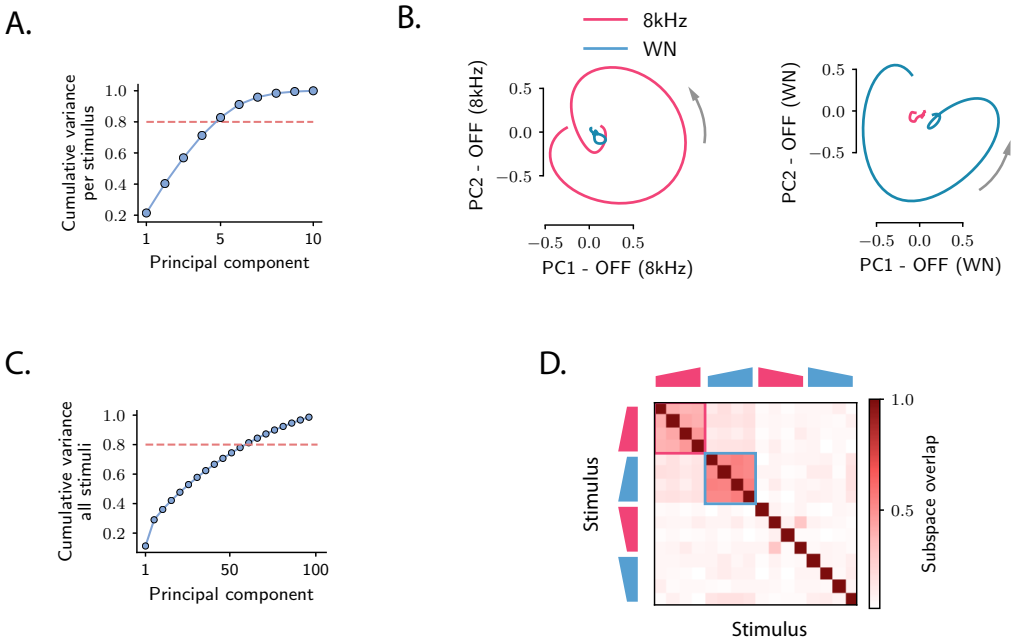

Figure 2: **Low-dimensional structure of population OFF responses. A.** Cumulative variance explained for OFF responses to individual stimuli as a function of the number of principal components. The trace shows the average cumulative variance over all 16 stimuli. Error bars are smaller than the symbol size. **B.** *Left*: projection of the population OFF response to the 8 kHz and white noise sounds on the first two principal components computed for the OFF response to the 8 kHz sound. *Right*: projection of both responses on the first two principal components computed for the OFF response to the white noise sound. PCA was performed on the period from -50 ms to 300 ms with respect to stimulus offset. We plot the response from 50 ms before stimulus offset to the end of the trial duration. **C.** Cumulative variance explained for all the OFF responses to the 16 stimuli together as a function of the number of principal components. **D.** Overlap between the subspaces defined by the first 5 principal components of the OFF responses relative to pairs of stimuli. The overlap is measured by the cosine of the principal angle between these subspaces (see Materials and methods). The red and blue boxes highlight the groups of stimuli corresponding respectively to the 8 kHz and white noise UP-ramps.

A recurrent network model for OFF responses

Our analyses of AC data identified three prominent features of population dynamics: (i) OFF responses correspond to transiently amplified trajectories; (ii) responses to individual stimuli explore low-dimensional subspaces; (iii) responses to different stimuli lie in largely orthogonal subspaces. We next aimed to examine whether these three features could be accounted for by recurrent interactions between neurons. To this end, we exploited a simple network model of recurrently coupled units. In this section we introduce the model and the underlying hypotheses, and determine the sufficient constraints under which the model reproduces qualitatively the three features listed above. The model hypotheses, and the identified constraints directly provide a new set of predictions which we test on the AC data in the following sections.

To model the OFF response dynamics, we used a network of N recurrently coupled linear rate units with

time evolution given by:

$$\dot{r}_i = -r_i + \sum_{j=1}^{N} J_{ij} r_j \tag{1}$$

The quantity $r_i(t)$ represents the deviation of the activity of the unit i from its baseline firing rate, while J_{ij} denotes the effective strength of the connection from neuron j to neuron i (Fig. 3A). We focused on the dynamics of this model following stimulus offset, which we represent as $t = 0$.

The model relied on two hypotheses. Our first hypothesis was that the network received no external input after stimulus offset. The only effect of the preceding stimulus was to set the initial pattern of activity of the network at $t = 0$. The internal recurrent dynamics then fully determined the subsequent evolution of the activity. Each stimulus s was thus represented by an initial state $\mathbf{r}_0^{(s)}$ that was mapped onto a trajectory of activity $\mathbf{r}^{(s)}(t)$.

Our second hypothesis was that the connectivity matrix was of low-rank type (Mastrogiuseppe and Ostojic, 2018), i.e. described by $R \ll N$ modes

$$\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T} + \mathbf{u}^{(2)}\mathbf{v}^{(2)T} + ... + \mathbf{u}^{(R)}\mathbf{v}^{(R)T}, \tag{2}$$

where each mode r was specified by two N-dimensional vectors $\mathbf{u}^{(r)}$ and $\mathbf{v}^{(r)}$, which we denote as the right and left connectivity patterns. Recurrent neural networks based on such connectivity directly generate low-dimensional dynamics: for a given stimulus s the trajectory of dynamics lies in the subspace spanned by the initial state $\mathbf{r}_0^{(s)}$ and the set of right connectivity patterns $\{\mathbf{u}^{(r)}\}_{r=1...R}$ (Mastrogiuseppe and Ostojic, 2018; Bondanelli and Ostojic, 2020; Schuessler et al., 2020). In this class of low-rank networks, the dynamics are therefore fully determined by a set of patterns over the neural population: patterns corresponding to initial states $\mathbf{r}_0$ determined by the stimuli, and patterns corresponding to connectivity vectors $\mathbf{u}^{(r)}$ and $\mathbf{v}^{(r)}$ for $r = 1 \dots R$. Our first goal was to identify sufficient conditions on the relationship between these patterns to reproduce qualitatively the three features of AC population activity summarized above.

The first feature of the data was that the dynamics were transiently amplified, i.e. the OFF responses first deviated from baseline before eventually converging to it. A preliminary requirement to reproduce this feature is that dynamics are stable, i.e. eventually decay to baseline following any initial state. This requirement leads to the usual condition that the eigenvalues of the connectivity matrix $\mathbf{J}$ have real parts less than unity. Provided this basic requirement is met, during the transient dynamics from the initial state at stimulus offset, the distance to baseline can either monotonically decrease, or transiently increase before eventually decreasing. To generate the transient increase, the connectivity matrix needs to belong to the class of non-normal matrices (Trefethen and Embree (2005); Murphy and Miller (2009); Goldman (2009); Hennequin et al. (2012); see Materials and methods), but this is not a sufficient condition. More specifically, the largest eigenvalue of the symmetric part defined by $\mathbf{J}_S = (\mathbf{J} + \mathbf{J}^T)/2$ needs to be larger than unity, while the initial state $\mathbf{r}_0$ needs to belong to a subset of amplified patterns (Bondanelli and Ostojic, 2020).

The second feature of AC data was that each stimulus generated a population response embedded in an approximately 5-dimensional subspace of the full state space. In terms of the low-rank model, this implies that each stimulus activates a small number of modes, and that the dynamics along those modes are transiently amplified. To illustrate the implications on the level of patterns of initial activity and connectivity, it is useful to examine the simplified example of a single stimulus, and single mode in the connectivity matrix, i.e. $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$. Dynamics along this mode will be simultaneously stable and amplified if the patterns $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ are close to orthogonal to each other (see Materials and methods, and (Bondanelli and Ostojic, 2020)). The initial state $\mathbf{r}_0$ set by the stimulus will activate this mode and lead to transient amplification if $\mathbf{r}_0$ strongly overlaps with the left connectivity pattern $\mathbf{v}^{(1)}$, as any component of $\mathbf{r}_0$ orthogonal to $\mathbf{v}^{(1)}$ will monotonically decay. When these conditions for amplification are met, the initial activity $\mathbf{r}_0$ elicits an amplified OFF response (Fig. 3B-C), and the evoked trajectory lies in the plane defined by $\mathbf{r}_0$ and $\mathbf{u}^{(1)}$ (Fig. 3D). The plane defined by $\mathbf{r}_0$ and $\mathbf{u}^{(1)}$ corresponds to the plane spanned by the first two principal components of the population responses following initial activity along $\mathbf{r}_0$ (see Materials and methods). These sufficient conditions identified in the unit-rank example can be extended to the more general case where the connectivity consists of a larger number of modes. Stable and transiently amplified dynamics are obtained when all possible pairs of left and right connectivity vectors are close to mutually orthogonal (see Materials and methods), and an amplified response along these modes will be obtained when the initial state $\mathbf{r}_0$ overlap strongly with the left connectivity patterns $\mathbf{v}^{(r)}$ (see Materials and methods). A direct consequence of these constraints is that the population activity vector at the peak of the transient dynamics lies along a direction that is largely orthogonal to the initial state at stimulus offset (Fig. 3F), see Materials and methods for an upper bound on the correlation between the two).

The third feature observed in AC data was that population responses evoked by different stimuli were orthogonal for many of the considered stimuli. Orthogonal trajectories for different stimuli can be reproduced in the model under the conditions that (i) initial patterns of activity corresponding to different stimuli are orthogonal, (ii) the connectivity matrix is partitioned into a set of orthogonal low-rank terms, each individually leading to transiently amplified dynamics, (iii) each stimulus activates one of the low-rank terms. Altogether, for P stimuli leading to orthogonal responses, the connectivity matrix can be partitioned in P different groups of modes:

$$\mathbf{J} = \sum_{s=1}^{P} \mathbf{J}^{(s)}, \tag{3}$$

with

$$\mathbf{J}^{(s)} = \sum_{r=1}^{R_s} \mathbf{u}^{(r;s)} \mathbf{v}^{(r;s)T} \tag{4}$$

The set of modes indexed by s correspond to the s-th stimulus, so that the pattern of initial activity $\mathbf{r}_0^{(s)}$ generated by stimulus s overlaps only with those modes. Moreover, modes corresponding to different groups are orthogonal, and generate dynamics that span orthogonal subspaces (Fig. 3D). Each term in Eq. (3) can therefore be interpreted as an independent transient coding channel, that can be determined from the OFF-response to stimulus s alone.

Testing the network model on the data

In the previous section, we laid out the assumptions behind our network model, and identified constraints on model components that would be sufficient to reproduce qualitatively the main features of OFF responses observed in the AC data. We next set out to test directly on the data the model assumptions and predictions resulting from the identified constraints on model components.

We first tested the starting assumption that the OFF response trajectories are consistent with linear, autonomous network dynamics driven only by the initial pattern of activity and network interactions. If that is the case, the OFF response dynamics are fully determined by the initial activity at stimulus offset ($t = 0$). This assumption implies that the geometric structure of the initial activity states $\mathbf{r}_0^{(s)}$ fully predicts the geometric structure of the dynamics $\mathbf{r}^{(s)}(t)$ during the OFF responses (see Materials and methods). In particular, two stimuli s_1 and s_2 corresponding to correlated initial states $\mathbf{r}_0^{(s_1)}$ and $\mathbf{r}_0^{(s_2)}$ will lead to correlated trajectories $\mathbf{r}^{(s_1)}(t)$ and $\mathbf{r}^{(s_2)}(t)$ in the model. Conversely, if two OFF-response trajectories $\mathbf{r}^{(s_1)}(t)$ and $\mathbf{r}^{(s_2)}(t)$ are orthogonal, the initial states $\mathbf{r}_0^{(s_1)}$ and $\mathbf{r}_0^{(s_2)}$ are necessarily orthogonal. To test this prediction on the data, we computed the correlation matrix between the initial states $\mathbf{r}_0^{(s)}$ in the data, i.e. the activity states reached at the end of stimulus presentation for all different stimuli (Fig. 4A *left panel*), and compared it with the matrix of trajectory subspace overlaps (Fig. 2D), which quantifies correlations between OFF responses following different stimuli. We found that the correlation structure of initial states matched well with the correlation structure of the dynamics during OFF responses (Fig. 4A *right panel*, Fig. S8), consistent with our assumption of autonomous network dynamics. In contrast, if the dynamics during OFF responses were driven by external inputs to the network, no direct match would necessarily be expected between the structure of initial activity and the subsequent trajectories.

To quantify more directly to which extent the posited recurrent network mechanism accounted for AC dynamics, we next fitted a linear network model to OFF responses, and compared the obtained fits with controls. For each stimulus, the pattern of initial activity $\mathbf{r}_0^{(s)}$ was taken from the data, and the connectivity matrix $\mathbf{J}$ was fitted to the responses to all stimuli at once. The only constraint we imposed on the fitted connectivity matrix was to minimize its rank through rank-reduced regression (Fig. S4, Fig. S7; see Materials and methods). This constraint provides an effective regularization that minimizes over-fitting, and ensures that the fitted connectivity matrix is of the form hypothesized in Eq. (2) with minimal R. Qualitatively, we found that the fit reproduced well the low-dimensional population trajectories (Fig. 4B). We evaluated the goodness of the fit by computing the coefficient of determination R^2, and found that the network model explained a relatively high fraction of the variance of OFF responses ($R^2 = 0.48$, Fig. 4C). To evaluate whether the model captured non-trivial collective dynamics or simply the smoothness of the activity, we followed the approach introduced in Elsayed and Cunningham (2017), and repeated the model fitting on control datasets which kept no additional structure than the one defined by correlations across time, neurons and stimuli of the original neural responses (see Materials and methods). We found that the value of the R^2 obtained from fitting the model to the original dataset was significantly higher than the one obtained from

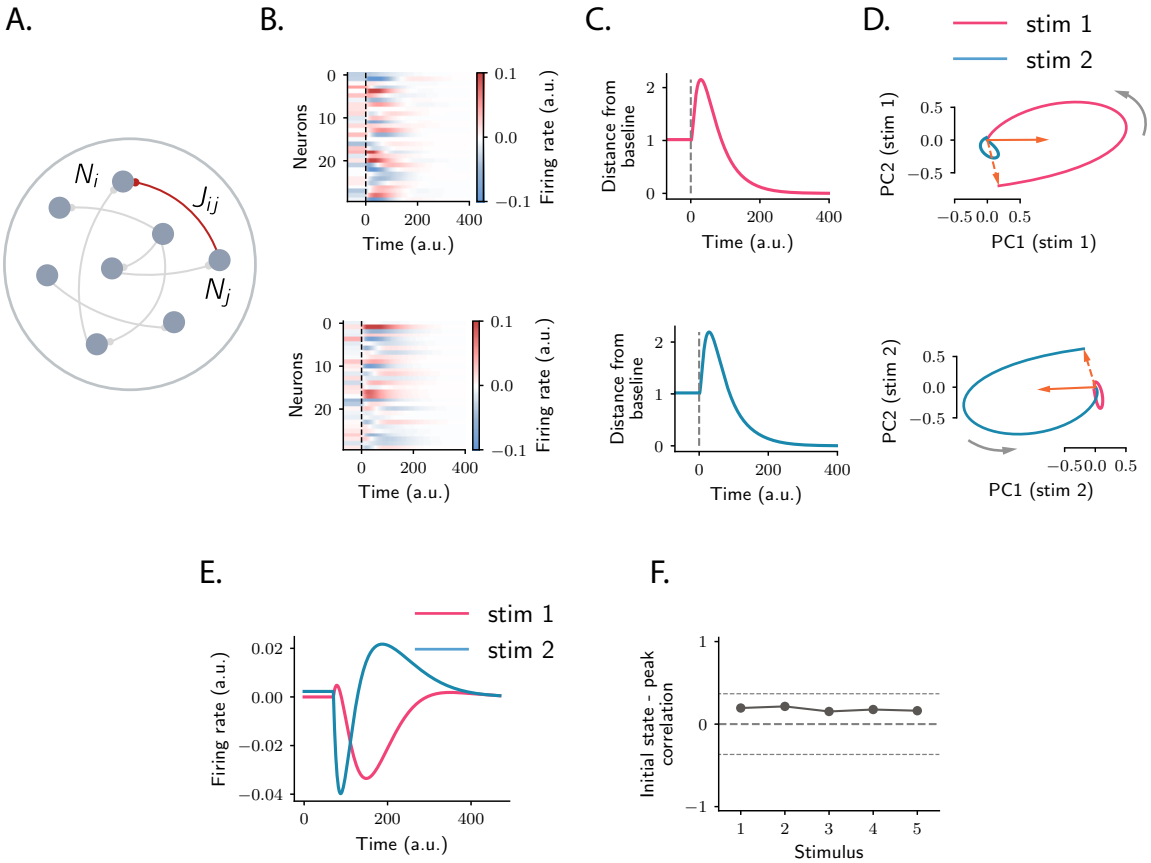

Figure 3: **Population OFF responses in a network model with low-rank structure. A.** Illustration of the recurrent network model. The variables defining the state of the system are the firing rates of the units, denoted by r_i. The strength of the connection from unit j to unit i is denoted by J_{ij}. **B.** Single-unit OFF responses to two distinct stimuli generated from the low-rank network model. The two stimuli were modeled as two different activity states $\mathbf{r}_0^{(1)}$ and $\mathbf{r}_0^{(2)}$ at the end of stimulus presentation ($t = 0$). We simulated a network of 2000 units. The connectivity consisted of 30 unit-rank transient channels of the form given by Eq. (3) (with $P = 30$, $R_1 = R_2 = ... = R_P = 1$ and $||\mathbf{u}^{(r)}|| = 5$ for all r). Here the two stimuli have been chosen along two orthogonal vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Dashed line indicate the time of stimulus offset. **C.** Distance of the population activity vector from baseline during the OFF responses to the two stimuli in **B.** The amplitude of the offset responses, as quantified by the peak value of the distance from baseline, is controlled by the length of the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$. **D.** Projection of the population OFF responses to the two stimuli on the two principal components of either stimuli. The projections of the vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ (resp. $\mathbf{u}^{(2)}$ and $\mathbf{v}^{(2)}$) on the subspace defined by the first two principal components of stimulus 1 (resp. 2) are shown respectively as solid and dashed arrows. The subspaces defined by the vector pairs $\mathbf{u}^{(1)}$-$\mathbf{v}^{(1)}$ and $\mathbf{u}^{(2)}$-$\mathbf{v}^{(2)}$ are mutually orthogonal, so that the OFF responses to stimuli $\mathbf{v}_1$ and $\mathbf{v}_2$ evolve in orthogonal dimensions. **E.** Firing rate of one example neuron in response to two distinct stimuli: in a recurrent network the time course of the activity of one unit is not necessarily the same across stimuli. **F.** Correlation between the state at the end of stimulus presentation and the state at the peak of the OFF responses, for 5 example stimuli. Error bars (smaller than symbol size) represent the standard deviation computed over 100 bootstrap subsamplings of 50% of the units in the population. Thin dashed lines indicates the upper bound on the absolute value of the correlation between initial and peak state obtained theoretically ($= 1/e$, see Eq. (57)).

fitting the control datasets (Fig. 4C,Fig. S5,Fig. S6), confirming that the recurrent network model captured non-trivial collective dynamics in the data beyond the correlations across time, neurons and stimuli.

Fitting the network model to the data provided us with an estimate of the effective connectivity matrix $\mathbf{J}$. We next used this estimate to test for the predictions on $\mathbf{J}$ resulting from the conditions we identified as sufficient for reproducing the three main qualitative features of the OFF responses. We first tested the condition for amplified transient responses, i.e. that $\mathbf{J}$ belongs to a specific subset of non-normal matrices. To this end, we examined the spectra of the full connectivity $\mathbf{J}$ and of its symmetric part $\mathbf{J}_S$ (Bondanelli and Ostojic, 2020). The eigenvalues of the fitted connectivity matrix mostly had real parts smaller than

one, indicating stable dynamics, and large imaginary parts (Fig. 4D *left panel*). Our analytical criterion predicted that, for OFF responses to be amplified, the spectrum of the symmetric part $\mathbf{J}_S$ must have at least one eigenvalue larger than unity. Consistent with the criterion derived theoretically, we found that the symmetric part of the connectivity had indeed a large number of eigenvalues larger than one (Fig. 4D *right panel*), and could therefore produce amplified responses (Fig. 4B).

We next examined the assumption that the connectivity is low-rank, and tested the ensuing predictions for the relationship between patterns of initial activity $\mathbf{r}_0^{(s)}$ and connectivity vectors $\mathbf{u}^{(r)}$ and $\mathbf{v}^{(r)}$ determined from the connectivity matrix $\mathbf{J}$ fitted to each stimulus independently. Minimizing the rank of the fitted connectivity matrices, we found an optimal rank of 6 per stimulus consistent with the PCA analysis of the trajectories. We first focused on the relative arrangement between the left and right connectivity vectors obtained from the singular-value decomposition of $\mathbf{J}$, and tested the prediction that the reciprocal overlap between all pairs of left and right connectivity patterns should be sufficiently small to generate stable amplification (Eq. (41), see Materials and methods). Consistent with this prediction we found that, across stimuli, most of left and right connectivity vectors were close to mutually orthogonal (Fig. 4E). We then tested the second prediction that initial states should be strongly aligned with the with the vectors $\mathbf{v}^{(r)}$, more precisely, that the components of $\mathbf{r}_0^{(s)}$ along the vectors $\mathbf{v}^{(r)}$ weighted by the norms $||\mathbf{u}^{(r)}||$ should be larger than a critical value (Eq. (47), see Materials and methods). This prediction was also borne out, as Eq. (47) held across all stimuli (Fig. 4F *right panel*), and the initial states had large overlap with the right connectivity vectors $\mathbf{v}^{(r)}$ (Fig. 4F *left panel*). Additionally the network model predicted that, when the sufficient conditions described above are met, for each stimulus, the population activity state at the peak of the transient phase is essentially orthogonal to the state at the end of stimulus presentation, resulting in values of the correlation smaller than $1/e \approx 0.35$ (Eq. (57)). Computing the correlation between the peak state and the initial state at the end of stimulus presentation we found that this correlation took values lower than the theoretical bound for all stimuli, consistent with the prediction of the network model (Fig. 4G, Fig. S9).

Importantly, we noticed systematic deviations from the expected low correlation between left and right connectivity vectors of $\mathbf{J}$ (Fig. 4E). These exceptions involved left and right connectivity vectors that were highly correlated across adjacent modes. This particular correlation structure gives rise to a series of largely independent channels characterized by strong anti-symmetric components that generate pairs of complex conjugate eigenvalues with large imaginary eigenvalues (see Materials and methods), as observed in the full spectrum (Fig. 4D). The strong anti-symmetric component of the fitted connectivity $\mathbf{J}$ is consistent with the large imaginary eigenvalues observed in the full spectrum. This anti-symmetric component is responsible for generating the rotational dynamics observed in the population OFF responses, but cannot alone generate amplified dynamics (see Materials and methods), for which it needs to be compensated by the strong symmetric part of the connectivity (Fig. 4D).

Finally we examined whether the fitted connectivity $\mathbf{J}$ consisted of independent low-rank coding channels (Eq. (3)), a constraint that would generate orthogonal responses to different stimuli as observed in the data. If $\mathbf{J}$ consisted of fully independent channels as postulated in Eq. (3), than it could be equivalently estimated by fitting the recurrent model to the OFF responses to each stimulus s independently, leading to one matrix $\mathbf{J}^{(s)}$ for each stimulus. The hypothesis of fully independent channels then implies that (i) the connectivity vectors for the different connectivity matrices are mutually orthogonal, (ii) the full connectivity matrix can be reconstructed by summing the matrices $\mathbf{J}^{(s)}$ estimated for individual stimuli. To test these two predictions, we estimated the connectivity matrices $\mathbf{J}^{(s)}$ (the rank parameter for each stimulus was set to $R = 6$). We then computed the overlaps between the connectivity vectors corresponding to different stimuli (see Materials and methods). We found that the overlap between the transient channels matched the overlap between the response dynamics and population states at the end of stimulus presentation (Fig. 4H, left panel). This indicated that some of the stimuli, but not all, corresponded to orthogonal coding channels. We further reconstructed the OFF responses using the matrix $\mathbf{J}_{\text{Sum}} = \sum_s \mathbf{J}^{(s)}$. We then compared the values of R^2 obtained using the matrices $\mathbf{J}_{\text{Sum}}$, $\mathbf{J}_{\text{Full}}$ and multiple controls where the elements of the matrix $\mathbf{J}_{\text{Sum}}$ were randomly shuffled. While the fraction of variance explained when using the matrix $\mathbf{J}_{\text{Sum}}$ was necessarily lower than the one computed using $\mathbf{J}_{\text{Full}}$, the model with connectivity $\mathbf{J}_{\text{Sum}}$ could still explain a consistent fraction of the total variance, and yielded values of R^2 significantly higher than the ones obtained from the shuffles (Fig. 4H, right panel). These results together indicated that the full matrix $\mathbf{J}_{\text{Full}}$ can indeed be approximated by the sum of the low-rank channels represented by the $\mathbf{J}^{(s)}$.

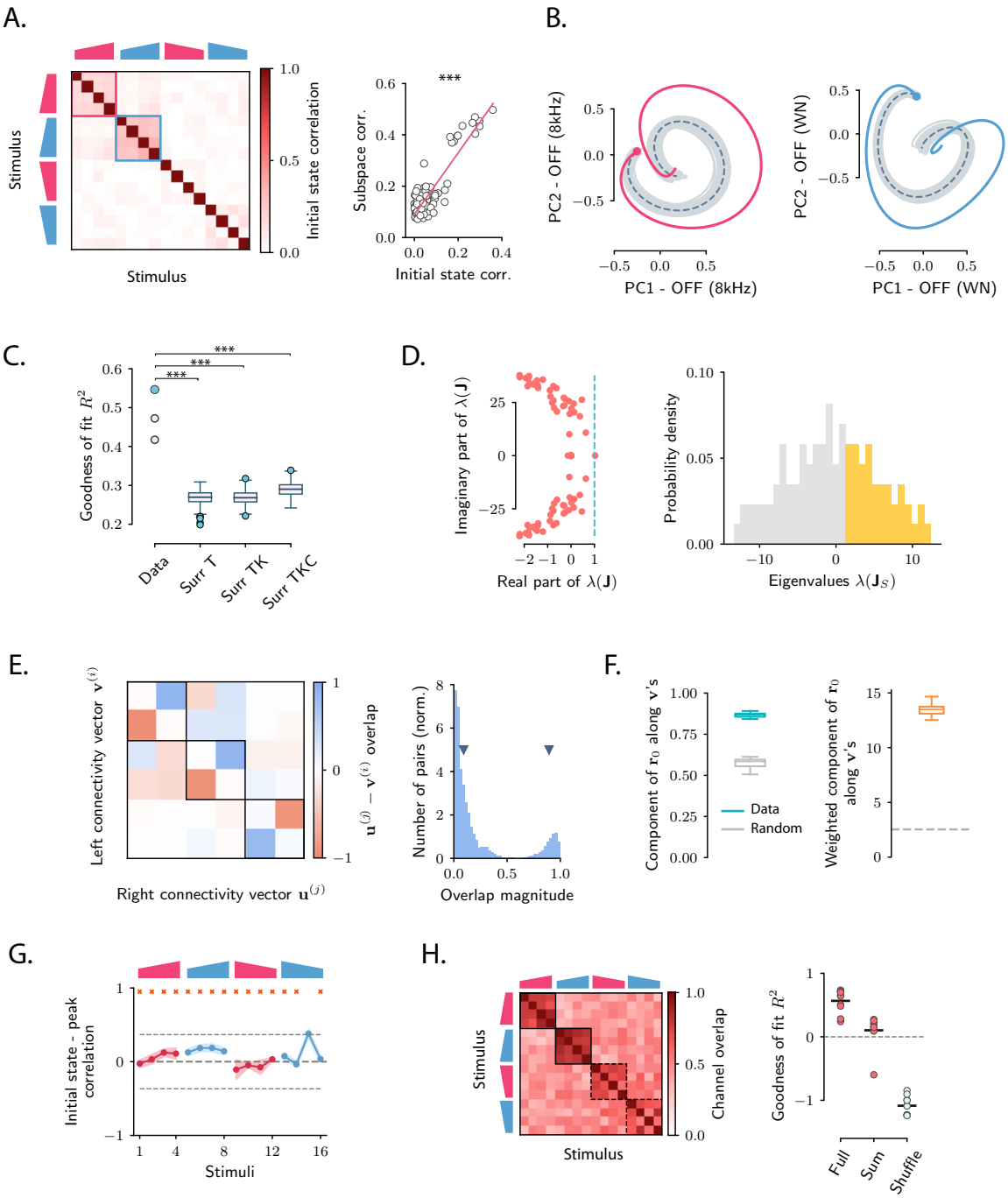
A.
Stimulus
Stimulus
Initial state correlation
1.0
0.5
0.0
Subspace corr.
Initial state corr.
***
B.
PC2 - OFF (8kHz)
PC1 - OFF (8kHz)
PC2 - OFF (WN)
PC1 - OFF (WN)
C.
Goodness of fit R^2
Data
Surr T
Surr TK
Surr TKC
D.
Imaginary part of $\lambda(\mathbf{J})$
Real part of $\lambda(\mathbf{J})$
Probability density
Eigenvalues $\lambda(\mathbf{J}_S)$
E.
Left connectivity vector $\mathbf{v}^{(i)}$
Right connectivity vector $\mathbf{u}^{(j)}$
$\mathbf{u}^{(j)} - \mathbf{v}^{(i)}$ overlap
Number of pairs (norm.)
Overlap magnitude
F.
Component of $\mathbf{r}_0$ along $\mathbf{v}$'s
Data
Random
Weighted component of $\mathbf{r}_0$ along $\mathbf{v}$'s
G.
Initial state - peak correlation
Stimuli
H.
Stimulus
Stimulus
Channel overlap
Goodness of fit R^2
Full
Sum
Shuffle

Figure 4

Figure 4 *(previous page)*: **Testing the predictions of the network model on auditory cortical data. A.** *Left*: correlation matrix between the initial population activity states $\mathbf{r}_0^{(s)}$ at the end of stimulus presentation (50 ms before offset) for each pair of stimuli. *Right*: linear correlation between subspace overlaps (Fig. 2D) and the overlap between initial states $\mathbf{r}_0^{(s)}$ for each stimulus pair. The component of the dynamics along the corresponding initial states was subtracted before computing the subspace overlaps. **B.** Colored traces: projection of the population OFF response to a 8 kHz and a white noise sound on the first two principal components. The responses are shown for the period from -50 ms to 300 ms with respect to stimulus offset. Light grey traces: projections of multiple trajectories generated by the recurrent model using the connectivity matrix fitted to all stimuli. Initial conditions were chosen in a neighbourhood of the population activity vector 50 ms before sound offset, indicated with a dot. 100 traces are shown. Dashed trace: projection of the average trajectories. The ridge and rank parameter are respectively $\lambda = 2$ and $R = 70$. Before the fit, the number of principal components was reduced to 100. **C.** Fraction of variance explained by a linear dynamical system for the original dataset and for three surrogate datasets, as quantified by the R^2 value. For the original dataset, the three dots show the average R^2 over the cross-validation folds for the three mice; the colored dot refers to mouse 1. For the surrogate datasets, the average values of R^2 over cross-validation folds are shown for multiple surrogates of mouse 1 data. This difference is significantly positive for the three types of surrogates, meaning that the linear dynamical system better fits the original dataset than the surrogate datasets (Here, *, **, *** correspond respectively to $P < 0.01$, $P < 0.005$, $P < 0.001$; upper tail test using the mean value of R^2 over the cross-validations folds for each dataset; 500 surrogates are shown). Box-and-whisker plots show the distribution of the values for the three types of surrogates (Tukey convention: box lower end, middle line and upper end represent the first quartile, median and third quartile of the distributions; lower and upper whisker extends for 1.5 times the interquartile range). Fitting is performed on the population OFF responses to all 16 stimuli together. For each dataset, the goodness of the fit R^2 is computed using 10-fold cross-validation in the time domain. The parameters used for reduced-rank ridge regression are the same as in **B. D.** Left: eigenvalues of the effective connectivity matrix $\mathbf{J}$ found by fitting a linear dynamical system to the population OFF responses. The dashed line marks the stability boundary given by $\mathfrak{Re}\lambda(\mathbf{J}) = 1$. *Right*: probability density distribution of the eigenvalues of the symmetric part of the effective connectivity, $\mathbf{J}_S$. Eigenvalues larger than unity determine strongly non-normal dynamics and are highlighted in yellow. The parameters used for reduced-rank ridge regression are the same as in **B. E.** *Left*: overlap between left and right connectivity patterns of the connectivity $\mathbf{J}$ fitted on one example stimulus. The color code shows strong (and opposite in sign) correlation values between left and right connectivity patterns belonging to adjacent modes. Weak coupling is instead present between the distinct rank-2 channels highlighted by dark boxes. *Right*: histograms of the absolute values of the correlation between right and left connectivity patterns, across all stimuli, and across 10 random subsamplings of 50% of the neurons in the population for each stimulus. Most of the pairs of left and right connectivity vectors present low values of the correlation, while a small fraction of these pairs have high correlation, and correspond to the anti-symmetric couplings within a rank-2 channel. The two markers indicate the average values of the low and high correlation values. When fitting individual stimuli the rank parameter R and the number of principal components are set respectively to 6 and 50. **F.** *Left*: component of the initial state $\mathbf{r}_0$ on the left connectivity vectors, i.e. $(\sum_r \alpha_r^2)^{1/2}$, obtained from the fitted connectivity $\mathbf{J}$ to individual stimuli (in green). The initial condition has a strong component along the left connectivity vectors $\mathbf{v}^{(r)}$'s. For comparison the component of random vectors along the left connectivity vectors is shown in grey. *Right*: component of the initial state $\mathbf{r}_0$ on the right connectivity vectors, α_r, weighted by the norms $||\mathbf{u}^{(r)}||$ (i.e. $\sqrt{\sum_r \alpha_r^2 ||\mathbf{u}^{(r)}||^2}$, see Eq. (47)), obtained from the fitted connectivity $\mathbf{J}$ to individual stimuli. The grey dashed line correspond to the lower bound for this value derived theoretically (Eq. (47)). In both panels the boxplots show the distributions across all stimuli of the corresponding average values computed over 10 random subsamplings of 50% of the neurons. The rank parameter and the number of principal component is the same as in **E. G.** For each stimulus, we show the correlation between the state at the end of stimulus presentation and the state at the peak of the OFF response, defined as the time of maximum distance from baseline. Error bars represent the standard deviation computed over 2000 bootstrap subsamplings of 50% of the neurons in the population (2343 neurons). Grey dashed lines indicate the upper bounds on the absolute value of the correlation between initial and peak state derived theoretically (Eq. (57)). Red markers on the top indicate the stimuli for which the correlation is smaller than the theoretical upper bound ($P < 0.05$, upper tail test between the upper bound and the absolute values of the correlation across all 2000 bootstrap subsamplings). **H.** *Left*: overlap between the transient channels relative to each stimulus, as quantified by the overlap between the left connectivity vectors $\mathbf{u}^{(r;s)}$ of the connectivities $\mathbf{J}^{(s)}$ fitted to each individual stimulus s (see Materials and methods). The rank parameter and the number of principal component is the same as in **E.** *Right*: goodness of fit, as quantified by the R^2 value, when reconstructing the population OFF responses using the connectivity $\mathbf{J}_{\text{Full}}$ or the connectivity given by the sum of the individual channels $\mathbf{J}_{\text{Sum}}$. These values are compared with the values of the R^2 obtained by shuffling the elements of the matrix $\mathbf{J}_{\text{Sum}}$(Shuffled). Different points correspond to the value of the R^2 computed for each cross-validation partition (10-fold). In all panels, the fit is computed using reduced-rank ridge regression, where the best ridge and rank hyperparameters are found using cross-validation (see Materials and methods and Fig. S4). The number of principal component is set to 100, if not differently stated.

Comparison with a single-cell model for OFF response generation

The auditory cortex is not the first stage where OFF responses arise. Robust OFF responses are found throughout the auditory pathway, and in subcortical areas the generation of OFF responses most likely relies on mechanisms that depend on the interaction between the inhibitory and excitatory synaptic inputs to single cells (Kopp-Scheinpflug et al., 2018).

To examine the possibility that OFF responses in auditory cortex themselves are consistent with a similar single-cell mechanism, we considered a simplified model for OFF response generation that mimics single-cell activation after stimulus offset (Fig. 2 in Kopp-Scheinpflug et al. (2018)). We assume that the response of neuron i after stimulus offset is specified by a characteristic linear filter $L_i(t)$ (Meyer et al., 2017), which describes the cell's intrinsic response generated by intracellular or synaptic mechanisms (Fig. 5A). Since the shape of this temporal response is assumed to be set by intrinsic properties, it is identical across different stimuli. Each stimulus modulates the response of each neuron in a multiplicative way depending on its activity at stimulus offset, so that the OFF response of neuron i to stimulus s can be written as:

$$r_i^{(s)}(t) = r_{0,i}^{(s)}\, L_i(t) \tag{5}$$

where $r_{0,i}^{(s)}$ is the modulation factor for stimulus s and neuron i. Without loss of generality, we assume that the values of the $L_i(0)$ are equal to unity, so that we can interpret the vector of modulation factors $\mathbf{r}_0^{(s)}$ as the vector of initial conditions $r_i^{(s)}(0)$. We estimated the single-cell responses $L_i(t)$ from the data by fitting basis functions to the responses of all the neurons subject to prior normalization by the initial conditions $r_{0,i}^{(s)}$ (see Materials and methods).

The single-cell model is able to qualitatively account for some of the prominent features observed in the data. Because the single-cell responses $L_i(t)$ are in general non-monotonic, the distance from baseline of the population activity vector can show amplified dynamics (Fig. 5B). At the population level, the trajectories generated by the single-cell model span at least two dimensions (Fig. 5C) and provide good fits to the trajectories in the data if the response to a single stimulus is considered (Fig. 5C).

Although the single-cell model was qualitatively consistent with transiently amplified, low-dimensional trajectories, we found that it could not account for other features in the data. A major discrepancy between the single-cell model and the AC data was the structure of responses across stimuli. While in the single-cell model the temporal shape of the response of each neuron is the same across all stimuli, this was not necessarily the case in the data (Fig. 5D). As a consequence, while fitting the single-cell model to OFF responses to a single stimulus led to an excellent match with AC data, increasing the number of simultaneously fitted stimuli significantly decreased the goodness of fit (Fig. 5B,C,D). In contrast, in the recurrent model, single-cell responses in general can differ across stimuli (Fig. 3E), so that the goodness of fit remained high as the number of simultaneously fitted stimuli was increased (Fig. 3D). Computing the goodness of fit by taking into account the number of parameters of the recurrent and single-cell models led to the same conclusion (Fig. S10).

A second qualitative difference concerned the decorrelation between the state at the end of stimulus presentation and the state at the peak of the transient phase. We found that when fitting the single-cell model to a progressively increasing number of stimuli, the value of the correlation between the initial and peak states gradually deviated from the experimental value, in contrast to the recurrent model, which predicted values of the correlation in agreement with the data for any number of fitted stimuli (Fig. 5F).

A third major difference between the recurrent and single-cell models appears when considering single-trial dynamics. So far we examined the activity obtained by averaging for each neuron the response to the same stimulus over multiple trials. The two models however also make predictions for the structure of trial-to-trial variability of simultaneously recorded neurons. Specifically, both models predict that neural dynamics shape the variability between stimulus offset and the peak of the OFF response, in such a way that variability is amplified along particular directions of the state-space (Fig. 6A). In the recurrent model, the noise correlation matrix directly inherits the low-rank structure of the connectivity matrix (see Materials and methods), so that the amplification of the variability between stimulus offset and the OFF response peak is stronger along the direction of maximum amplification in trial-averaged responses (aligned with the left connectivity vectors $\mathbf{u}^{(r)}$) than along random directions in the state-space (Fig. 6B,C,D, see Materials and methods, Eq. (78), (77), (79)). In the single-cell model, the direction of maximal variability in general depends on the temporal responses $L_i(t)$ of individual neurons, but the variability is again larger along the direction of maximum amplification in trial-averaged responses than along a random direction (Fig. 6C,D). Both models therefore predict an amplification of the variability along a specific direction of the state-space, however the underlying mechanisms are different. In the recurrent model, amplified variability is a consequence of

inter-neuron correlations generated by recurrent interactions, and is thus significantly reduced after removing noise correlations by shuffling trial labels independently for each cell (Fig. 6C-D *left panels*). In contrast, in the single-cell model, inter-neuron correlations are absent, and amplified variability is instead due to the heterogeneous responses of individual units, and is therefore not modified when considering shuffled responses (Fig. 6C-D *center panels*).

Inspecting trial-to-trial variability in the calcium activity data, across all stimuli we found that the variability amplification was higher for the amplified direction in trial-averaged responses than for the random direction (Fig. 6C-D *right panels*), which was consistent with both models. The difference in variability amplification was however significantly stronger when considering the original responses with respect to the shuffled responses where noise correlations had been removed (Fig. 6C-D *right panels*), in agreement with predictions of the recurrent, but not the single-cell model. This difference in variability amplification is not a straight-forward consequence of noise-correlations known to be present in the auditory cortex (Rothschild et al., 2010), but instead could have been generated by single-cell variability alone, as demonstrated by the single-cell model (Fig. 6C-D *center panels*). Our results therefore indicate that the recurrent interactions between neurons in the network play a key role in shaping trial-to-trial response variability.

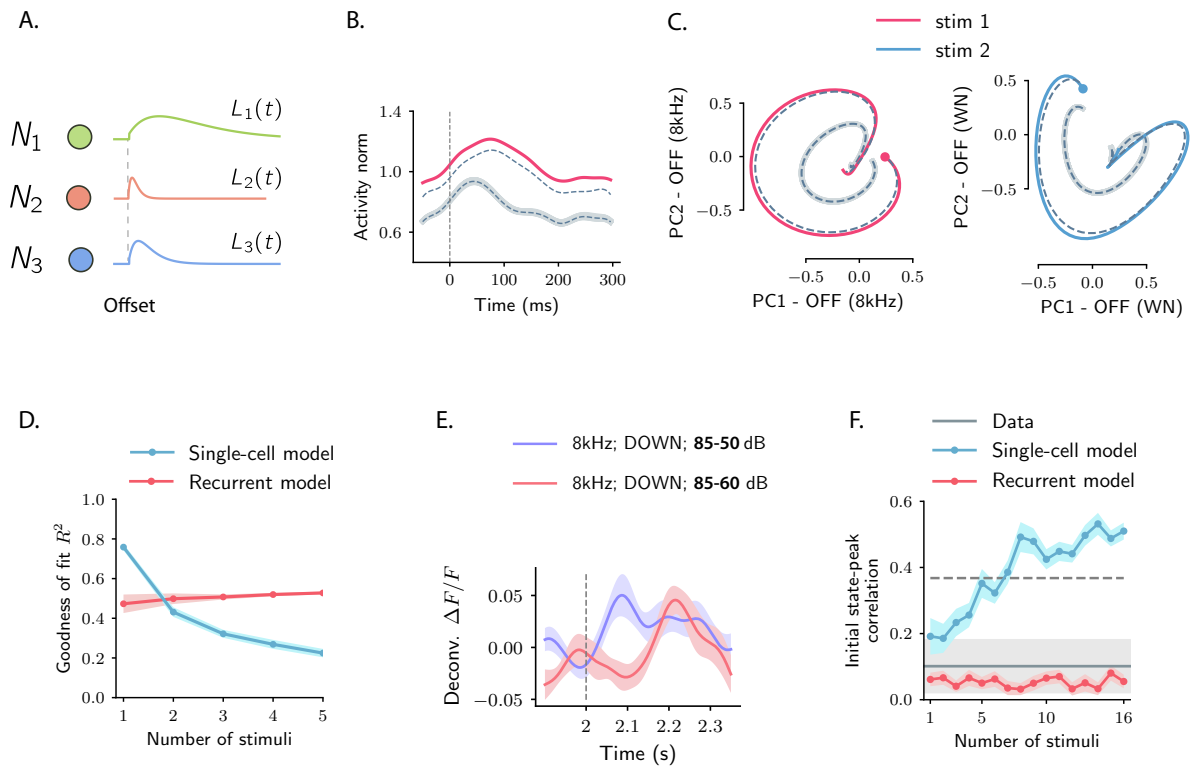

Figure 5: **Comparison between single-cell and network models for OFF response generation. A.** Cartoon illustrating the single-cell model defined by Eq. (5). The activity of each unit is described by a characteristic response filer $L_i(t)$ (colored traces). Across stimuli, only the relative activation of the units changes, but the shape of the responses of each unit $L_i(t)$ does not vary. **B.** Distance from baseline of the population activity vector during the OFF response to one example stimulus (red trace; same stimulus as in panel **C.** *Left*). The dashed vertical line indicate the time of stimulus offset. Grey dashed lines correspond to the norms of the population activity vectors obtained from fitting the single-cell model respectively to the OFF response to a single stimulus, and simultaneously to OFF responses to two stimuli (in this case, the two stimuli are the ones considered in panel **C.**) In light grey we plot the norm of 100 realization of the fitted OFF response by simulating different initial states distributed in the vicinity of the fitted initial state. **C.** Colored traces: projection of the population OFF responses to two distinct example stimuli on the two principal components of either stimulus. As in panel **B.** grey dashed traces correspond to the projection of the OFF responses obtained when fitting the single-cell model to one or two stimuli at the same time. **D.** Relative variance explained, as quantified by R^2, when fitting the single-cell model (green trace) and the recurrent model (red trace) to the AC calcium activity data, as a function of the number of stimuli included in the fit. Both traces show the cross-validated value of R^2 (10-fold CV in the time domain). Error bars represent the standard deviation over multiple samples of the stimuli. **E.** Example of an auditory cortical neuron that responds at the offset of two distinct stimuli with different response profiles, a feature consistent with the recurrent model (see Fig. 3E.), but not with the single-cell model. **F.** Correlation between initial state and peak state obtained from fitting the single-cell (in green) and recurrent model (in red) to a progressively increasing number of stimuli. For the fitted responses the peak state is defined as the population vector at the time of maximum distance from baseline of the original response. The colored shaded areas correspond to the standard error over 50 subsamplings of a fixed number of stimuli (on the abscissa) out of 16 stimuli. For each subsampling of stimuli, the correlation is computed for one example stimulus. In panels **B.**, **C.**, **E.** and **F.** the fit of the single-cell and recurrent models are performed on the first 40% most responding units, to avoid normalizing the responses by very low values when fitting basis functions (see Materials and methods). The fit of the single-cell model is performed using 10 Gaussian basis functions (see Materials and methods).

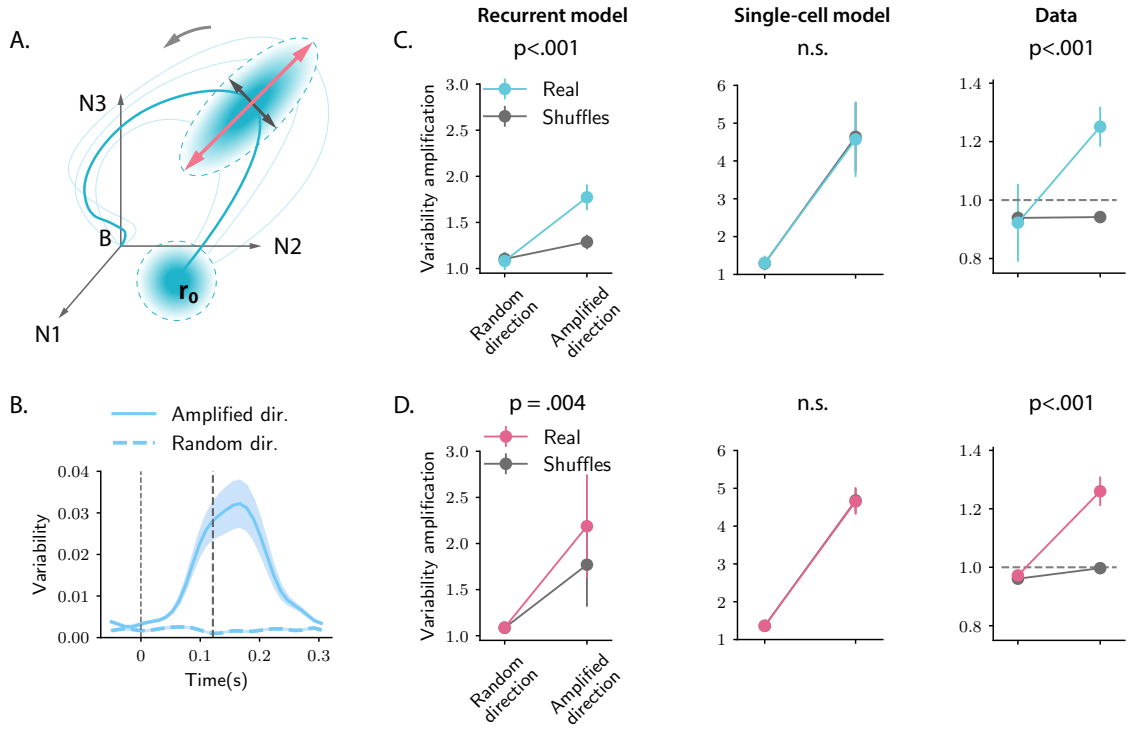

Figure 6: **Structure of trial-to-trial variability generated by the recurrent and single-cell model, compared to the AC data. A.** Cartoon illustrating the structure of single-trial population responses along different directions in the state-space. The thick and thin blue lines represent respectively the trial-averaged and single-trial responses. Single-trial responses start in the vicinity of the trial-averaged activity $\mathbf{r}_0$. Both the recurrent and single-cell mechanism dynamically shape the structure of the single-trial responses, here represented as graded blue ellipses. The red and black lines represent respectively the amplified and random directions considered in the following analyses. **B.** Time course of the variability computed along the amplified direction (solid trace) and along a random direction (dashed trace) for one example stimulus and one example session (287 simultaneously recorded neurons). In this and all the subsequent panels the amplified direction is defined as the population activity vector at the time when the norm of the trial-averaged activity of the pseudo-population pooled over sessions reaches its maximum value (thick dashed line). Thin dashed lines denote stimulus offset. Shaded areas correspond to the standard deviation computed over 20 bootstrap subsamplings of 19 trials out of 20. **C.** Variability amplification computed for the amplified and random directions on trajectories generated using the recurrent model, the single-cell model (see Materials and methods), and on the calcium activity data, for one example stimulus (same as in **B.**) The recurrent and single-cell models were first fitted to the trial-averaged responses to individual stimuli, independently across all recordings sessions (13 sessions, 180 ± 72 neurons). The single-cell model was fitted on the first 40% most responsive neurons for each stimulus and session, to avoid normalizing for very small values (see Materials and methods). 100 single-trial responses were then generated by applying the fitted models on 100 initial conditions isotropically distributed around the trial-averaged initial state $\mathbf{r}_0$. In the three panels, the values of variability amplification were computed over 50 subsamplings of 90% of the cells (or PC dimensions for the recurrent model), and 50 shuffles. Error bars represent the standard deviation over multiple subsamplings, after averaging over all sessions and shuffles. Significance levels were evaluated by first computing the difference in variability amplification between amplified and random directions (ΔVA) and then computing the p-value on the difference between ΔVA(Real) and ΔVA(Shuffles) across subsamplings (two-sided independent t-test). For the recurrent model, the variability amplification is higher for the amplified direction than for a random direction, and this effect is significantly stronger for the real than for the shuffled responses. Instead, in the single-cell model the values of the variability amplification computed on the real responses are not significantly different from the one computed on the shuffled responses. For this particular stimulus, the variability amplification computed on calcium activity data is significantly different from the one computed on the shuffled responses, and is therefore consistent with the prediction of the recurrent mechanism. **D.** Variability amplification computed on trajectories generated using the recurrent model, the single-cell model, and on the calcium activity data, across all 16 stimuli. Error bars represent the standard error across stimuli. Significance levels were evaluated computing the p-value on the difference between ΔVA(Real) and ΔVA(Shuffles) across stimuli (two-sided Wilcoxon signed-rank test).

Discussion

Adopting a population approach, we showed that strong OFF responses observed in auditory cortical neurons form transiently amplified trajectories that encode individual stimuli within low-dimensional subspaces. A geometrical analysis revealed a clear structure in the relative orientation of these subspaces, where subspaces corresponding to different auditory stimuli were largely orthogonal to each other. We demonstrated that these features of OFF-responses can be generated by recurrent interactions among neurons. Indeed, we showed that a simple, linear recurrent network model accounts for a number of properties of the population OFF responses, notably the low-dimensionality of the transient channels and their global structure across multiple stimuli. In contrast, a single-neuron model captures the response to individual stimuli, but not the structure across stimuli and across individual trials.

In this study, we focused specifically on the responses following stimulus offset. Strong transients during stimulus onset display similar transient coding properties (Mazor and Laurent, 2005), and could be generated by the same network mechanism as we propose for the OFF responses. However, during ON-transients, a number of additional mechanisms are likely to play a role, in particular single-cell adaptation, synaptic depression or feed-forward inhibition. Indeed, recent work has showed that ON and OFF trajectories elicited by a stimulus are orthogonal to each other in the state-space (Saha et al., 2017), and this was also the case in our dataset (Fig. S3). Linear network models instead produce ON and OFF responses that are necessarily correlated, and cannot account for the observed orthogonality of ON and OFF responses for a given stimulus. Distinct ON and OFF response dynamics could also result from intrinsic non-linearities known to play a role in the auditory cortex (Calhoun and Schreiner, 1998; Rotman et al., 2001; Sahani and Linden, 2003; Machens et al., 2004; Williamson et al., 2016; Deneux et al., 2016).

A major assumption of our model is that the auditory cortex does not receive external inputs after the auditory stimulus is removed. Our results show that the recurrent dynamics are able to produce strongly amplified OFF transients in the absence of any external drive, and we found that this mechanism can fully account for auditory cortex data. However, neurons in the auditory cortex receive direct input from the medial geniculate body (MGB) of the thalamus, and indirect input from upstream regions of the auditory pathway, where strong OFF responses have been observed. Thus, in principle, OFF responses observed in auditory cortex could be at least partly inherited from upstream auditory regions (Kopp-Scheinpflug et al., 2018). Disentangling the contributions of upstream inputs and recurrent dynamics is challenging if one has access only to AC activity (but see Seely et al. (2016) for an interesting computational approach). Ultimately, the origin of OFF responses in auditory cortex needs to be addressed by comparing responses between related areas, an approach recently adopted in the context of motor cortical dynamics (Lara et al., 2018). A direct prediction of our model is that the inactivation of recurrent excitation in auditory cortical areas should weaken OFF responses (Li et al., 2013). However, recurrency in the auditory system is not present only within the cortex but also between different areas along the pathway (Ito and Malmierca, 2018; Winer et al., 1998; Lee et al., 2011). Therefore OFF responses could be generated at a higher level of recurrency and might not be abolished by inactivation of AC.

The dimensionality of the activity in large populations of neurons in the mammalian cortex is currently the subject of debate. A number of studies have found that neural activity explores low-dimensional subspaces during a variety of simple behavioral tasks (Gao et al., 2017). In contrast, a recent study in the visual cortex has shown that that the response of large populations of neurons to a large number of visual stimuli is instead high-dimensional (Stringer et al., 2019). Our results provide a potential way of reconciling these two sets of observations. We find that the population OFF responses evoked by individual auditory stimuli are typically low-dimensional, but lie in orthogonal spaces, so that the dimensionality of the responses increases when considering an increasing number of stimuli. Note that in contrast to (Stringer et al., 2019), we focused here on the temporal dynamics of the population response, the dimensionality of which is limited by the temporal resolution (Fig. S1).

The analyses we performed in this study were directly inspired by an analogy between OFF responses in the sensory areas and neural activity in the motor cortices (Churchland and Shenoy, 2007; Churchland et al., 2010, 2012). Starting at movement onset, single-neuron activity recorded in motor areas exhibits strongly transient and multiphasic firing lasting a few hundreds of milliseconds. Population-level dynamics alternate between at least two dimensions, shaping neural trajectories that appear to rotate in the state-space. These results have been interpreted as signatures of an underlying dynamical system implemented by recurrent network dynamics (Churchland et al., 2012; Shenoy et al., 2011), where the population state at movement onset provides the initial condition able to generate the transient dynamics used for movement generation. Computational models have explored this hypothesis (Sussillo et al., 2015; Hennequin et al., 2014; Stroud et al., 2018) and showed that the complex transient dynamics observed in motor cortex can be generated in network models with strong recurrent excitation balanced by fine-tuned inhibition (Hennequin et al., 2014).

Surprisingly, fitting a recurrent network model to auditory cortical data, we found that the arrangement of the eigenvalues of the connectivity matrix was qualitatively similar to the spectrum of this class of networks, suggesting that a common mechanism might account for the responses observed in both areas. However, while models and analyses of motor cortex data have focused on the anti-symmetric part of the connectivity to isolate rotational dynamics, in our analysis we instead pointed out the role of the symmetric part in generating strongly amplified transient activity. Indeed, while we found strong anti-symmetric components in the fitted connectivity matrix (Fig. 4E), these components did not account for transient amplification in the population trajectories (see Materials and methods).

The perceptual significance of OFF responses in the auditory pathway is still matter of ongoing research. Single-cell OFF responses observed in the auditory and visual pathways have been postulated to form the basis of duration selectivity (Brand et al., 2000; Alluri et al., 2016; He, 2002; Aubie et al., 2009; Duysens et al., 1996). In the auditory brainstem and cortex, OFF responses of single neurons exhibit tuning in the frequency-intensity domain, and their receptive field has been reported to be complementary to the receptive field of ON responses (Henry, 1985; Scholl et al., 2010). The complementary spatial organization of ON and OFF receptive fields may result from two distinct sets of synaptic inputs to cortical neurons (Scholl et al., 2010), and has been postulated to form the basis for higher-order stimulus features selectivity in single cells, such as frequency-modulated (FM) sounds (Sollini et al., 2018) and amplitude-modulated (AM) sounds (Deneux et al., 2016), both important features of natural sounds (Sollini et al., 2018; Nelken et al., 1999). Complementary tuning has also been observed between cortical ON and OFF responses to binaural localization cues, suggesting OFF responses may contribute to the encoding of sound-source location or motion (Hartley et al., 2011). At the population level, the proposed mechanism for OFF response generation may provide the basis for encoding complex sequences of sounds. Seminal work in the olfactory system has shown that sequences of odors evoked specific transient trajectories that depend on the history of the stimulation (Broome et al., 2006; Buonomano and Maass, 2009). Similarly, within our framework, different combinations of sounds could bring the activity at the end of stimulus offset to different regions of the state-space, setting the initial condition for the subsequent OFF responses. If the initial conditions corresponding to different sequences are associated with distinct transient coding channels, different sound sequences would evoke transient trajectories along distinct dimensions during the OFF responses, therefore supporting the robust encoding of complex sound sequences.

Materials and methods

Contents

1 Data analysis **19**
- 1.1 The dataset . . . 19
- 1.2 Decoding analysis . . . 19
- 1.3 Principal component analysis . . . 20
- 1.4 Cross-validated PCA . . . 20
- 1.5 Correlations between OFF response subspaces . . . 20
- 1.6 Controls for subspace overlaps and initial state-peak correlations . . . 21
- 1.7 Linear dynamical system fit . . . 21
- 1.8 Linear dynamical fit on PCA-reduced data . . . 22
- 1.9 Control datasets . . . 23
- 1.10 Analysis of the transient channels . . . 24

2 The network model **24**
- 2.1 Normal and non-normal connectivity matrices . . . 24
- 2.2 Sufficient condition for amplified OFF responses . . . 25
- 2.3 Low-rank networks . . . 25
- 2.4 Correlation between structure of initial states and structure of the dynamics . . . 28
- 2.5 Correlation between initial and peak state . . . 29
- 2.6 Principal component analysis of rank-1 dynamics . . . 30
- 2.7 Amplification in a rank-2 anti-symmetric channel . . . 30

3 Single-cell model for OFF response generation **31**
- 3.1 Fitting single-cell responses with basis functions . . . 31
- 3.2 Structure of single-trial responses in the recurrent model . . . 32
- 3.3 Structure of single-trial responses in the single-cell model . . . 34

1 Data analysis

1.1 The dataset

Neural recordings Neural data was recorded and described in previous work (Deneux et al., 2016). We analysed the activity of 2343 neurons in mouse auditory cortex recorded using two-photon calcium imaging while mice were passively listening to different sounds. Each sound was presented 20 times. Data included recordings from 3 mice across 13 different sessions. Neural recordings in the three mice comprised respectively 1251, 636 and 456 neurons. Recordings in different sessions were performed at different depths, typically with a 50μm difference (never less than 20μm). Since the soma diameter is $\sim 15\mu$m, this ensured different cells were recorded in different sessions. We analysed the trial-averaged activity of a pseudo-population of neurons built by pooling across all sessions and animals. The raw calcium traces (imaging done at 31.5 frames per second) were smoothed using a Gaussian kernel with standard deviation $\sigma = 32$ ms.

The stimuli set The stimuli consisted of a randomized presentation of 16 different sounds, 8 UP-ramping sounds and 8 DOWN-ramping sounds. For each type, sounds had different frequency content (either 8kHz or white noise (WN)), different durations (1 or 2 seconds) and different combinations of onset and offset intensity levels (for UP-ramps either 50-85dB or 60-85dB, while for DOWN-ramps 85-50dB or 85-60dB). The descriptions of the stimuli are summarized in Table stimuli.

Stim	Direction	Frequency	Duration (s)	Modulation (dB)
1	UP	8 kHz	1 s	50-85
2	UP	8 kHz	1 s	60-85
3	UP	8 kHz	2 s	50-85
4	UP	8 kHz	2 s	60-85
5	UP	WN	1 s	50-85
6	UP	WN	1 s	60-85
7	UP	WN	2 s	50-85
8	UP	WN	2 s	60-85
9	DOWN	8 kHz	1 s	85-50
10	DOWN	8 kHz	1 s	85-60
11	DOWN	8 kHz	2 s	85-50
12	DOWN	8 kHz	2 s	85-60
13	DOWN	WN	1 s	85-50
14	DOWN	WN	1 s	85-60
15	DOWN	WN	2 s	85-50
16	DOWN	WN	2 s	85-60

Table 1: Stimuli set

1.2 Decoding analysis

To assess the accuracy of stimulus discrimination (8kHz vs. white noise sound) on single-trials, we trained and tested a linear discriminant classifier (Bishop, 2006) using cross-validation. For each trial, the pseudo-population activity vectors were built at each $\sim$ 50ms time bin. We used leave-one-out cross-validation (LOOCV). At each time bin we used 19 out of 20 trials as the training set, and tested the obtained decoder on the remaining trial. The classification accuracy was the average of correctly classified stimuli over all 20 cross-validation folds.

At each time t the decoder for classification between stimuli s_1 and s_2 was trained using the trial-averaged pseudo-population vectors $\mathbf{c}_{1t}$ and $\mathbf{c}_{2t}$. These vectors defined the decoder $\mathbf{w}_t$ and the bias $\mathbf{b}_t$ given by:

$$\mathbf{w}_t = \mathbf{c}_{1t} - \mathbf{c}_{2t}, \qquad \mathbf{b}_t = \frac{\mathbf{c}_{1t} + \mathbf{c}_{2t}}{2} \tag{6}$$

A given population vector $\mathbf{x}$ was classified as either stimulus s_1 or stimulus s_2 according to the value of the function $y(\mathbf{x}) = \mathbf{w}_t^T \mathbf{x} - \mathbf{b}_t$:

$$\begin{cases} \text{if } y(\mathbf{x}) > 0 \text{ then } \mathbf{x} \text{ is classified as stimulus } s_1 \\ \text{if } y(\mathbf{x}) < 0 \text{ then } \mathbf{x} \text{ is classified as stimulus } s_2 \end{cases} \tag{7}$$

Random performance was evaluated by training and testing the classifier using cross-validation on surrogate datasets built by shuffling stimulus labels of single trials at each time bin.

1.3 Principal component analysis

To perform principal component analysis on the population responses to C stimuli $s_{i_1}, ..., s_{i_C}$ we considered the matrix $\mathbf{X} \in \mathbb{R}^{N \times TC}$, where N is the number of neurons and T is the number of time steps. $\mathbf{X}$ contained the population OFF responses to the stimuli $s_{i_1}, ..., s_{i_C}$, centered around the mean over times and stimuli. If we denote by λ_i the i-th eigenvalue of the correlation matrix $\mathbf{X}\mathbf{X}^T$, the percentage of variance explained by the i-th principal component is given by:

$$\mathsf{VAR}(i) = \lambda_i \Big/ \sum_{j=1}^{N} \lambda_j \tag{8}$$

while the cumulative percentage of variance explained by the first M principal components is given by:

$$\mathsf{CUMVAR}(M) = \sum_{j=1}^{M} \lambda_j \Big/ \sum_{j=1}^{N} \lambda_j \tag{9}$$

In Fig. 1A, for each stimulus s we consider the matrix $\mathbf{X}^{(s)} \in \mathbb{R}^{N \times T}$ containing the population OFF responses to stimulus s. We computed the cumulative variance explained as a function of the number of principal components M and then averaged over all stimuli.

1.4 Cross-validated PCA

We used cross validation to estimate the variance of OFF responses to individual stimuli and across stimuli attributable to the stimulus-related component, discarding the component due to trial-to-trial variability (Fig. S1). Specifically, we applied to our calcium activity data the method of cross-validated principal component analysis (cvPCA) developed in Stringer et al. (2019). This method provides an unbiased estimate of the stimulus-related variance by computing the covariance between a training and a test set of responses (e.g. two different trials) to an identical collection of stimuli. Let $\mathbf{X}^{(train)}$ and $\mathbf{X}^{(test)}$ be the $N \times TC$ matrices containing the mean-centered responses to the same C stimuli. We consider the training and test responses to be two distinct trials. We first perform ordinary PCA on the training responses $\mathbf{X}^{(train)}$, and find the principal component $\mathbf{u}_i^{(train)}$ $(i = 1, .., N)$. We then evaluate the cross-validated PCA spectrum $\{\lambda_i\}$ as:

$$\lambda_i = \frac{1}{C} \mathbf{u}_i^{(\text{train})T} \mathbf{X}^{(\text{test})T} \mathbf{X}^{(\text{train})} \mathbf{u}_i^{(\text{train})}. \tag{10}$$

We repeat the procedure for all pairs of trials (i, j) with $i \neq j$ and average the result over pairs. The cross-validated cumulative variance is finally computed as in Eq. (9).

1.5 Correlations between OFF response subspaces

To quantify the degree of correlation between pairs of OFF responses corresponding to two different stimuli, we computed the cosine of the principal angle between the corresponding low-dimensional subspaces. In general, the principal angle θ_P between two subspaces U and V represents the largest possible correlation between any two pairs of vectors in U and V and it is defined by the relation (Bjorck and Golub, 1973; Knyazev and Argentati, 2002)

$$\cos \theta_P = \max_{\mathbf{u} \in U, \mathbf{v} \in V} \mathbf{u}^T \mathbf{v} \tag{11}$$

To compute the correlations between the OFF responses to stimuli s_1 and s_2 we first identified the $K = 5$ dominant principal components of the response to stimulus s_1 and organized them in a $N \times K$ matrix $\mathbf{Q}(s_1)$. We repeated this for stimulus s_2, which yielded a matrix $\mathbf{Q}(s_2)$. Therefore the columns of $\mathbf{Q}(s_1)$ and $\mathbf{Q}(s_2)$ define the two subspaces on which the responses to stimuli s_1 and s_2 live. The cosine of the principal angle between these two subspaces is given by (Bjorck and Golub, 1973; Knyazev and Argentati, 2002):

$$\cos\theta_P(s_1, s_2) = \sigma_1\Big(\mathbf{Q}(s_1)^T\mathbf{Q}(s_2)\Big) \tag{12}$$

corresponding to the largest singular value of the matrix $\mathbf{Q}(s_1)^T\mathbf{Q}(s_2)$. We note that this procedure directly relates to canonical correlation analysis (CCA; see Hotelling (1936); Uurtio et al. (2017)). In particular the first principal angle corresponds to the first canonical weight between the subspaces spanned by the columns of $\mathbf{Q}(s_1)$ and $\mathbf{Q}(s_2)$ (Golub and Zha, 1992; Bjorck and Golub, 1973).

1.6 Controls for subspace overlaps and initial state-peak correlations

We assessed the significance of the orthogonality between OFF response subspaces across stimuli (Fig. 4A) using two different controls, which aim at testing for two distinct null hypotheses (Fig. S2). The first hypothesis is that small subspace overlaps (i.e. low correlations) between OFF responses to different stimuli may be due to the high number of dimensions of the state-space in which they are embedded. To test for this hypothesis we compared the subspace overlap computed on the trial-averaged activity with the subspace overlaps computed on the trial-averaged activity where the stimulus labels had been shuffled across trials. We shuffled stimulus labels multiple times, resulting in one value of the subspace overlap for each shuffle. For each pair of stimuli, significance levels were then computed as the fraction of shuffles for which the subspace overlap was lower than the subspace overlap for the real data (lower tail test; Fig. S2A).

Alternatively, small subspace overlaps could be an artifact of the trial-to-trial variability present in the calcium activity data. In fact, for a single stimulus, maximum values of the correlation between two different trials were of the order of 0.2 (Deneux et al., 2016). To test for the possibility that small subspace overlaps may be due to trial-to-trial variability, for each pair of stimuli we computed the values of the subspace overlaps by computing the trial-averaged activity on only half of the trials (10 trials), subsampling the set of 10 trials multiple times for each pair of stimuli. This yielded a set of values $\cos\theta_{\text{real}}(s_1, s_2, n)$, where s_1 and s_2 are the specific stimuli considered and $n = 1, ..., N_{\text{shuffle}}$. We then computed the subspace overlaps between the trial-averaged responses to the same stimulus, but averaged over two different sets of 10 trials each, over multiple permutations of the trials, resulting in a set of values $\cos\theta_{\text{shuffle}}(s, n)$, where s is the specific stimulus considered and $n = 1, ..., N_{\text{shuffle}}$. For each pair of stimuli s_1 and s_2, significance levels were computed using two-tailed t-test and taking the maximum between the p-values given by $p(\cos\theta_{\text{real}}(s_1, s_2, :), \cos\theta_{\text{shuffle}}(s_1, :))$ and $p(\cos\theta_{\text{real}}(s_1, s_2, :), \cos\theta_{\text{shuffle}}(s_1, :))$ for those stimulus pairs for which $\overline{\cos\theta_{\text{real}}(s_1, s_2, :)} < \overline{\cos\theta_{\text{shuffled}}(s_1, :)}$ and $\overline{\cos\theta_{\text{real}}(s_1, s_2, :)} < \overline{\cos\theta_{\text{shuffled}}(s_2, :)}$, where the bar symbol indicated the mean over shuffles (Fig. S2B).

The same null hypotheses have been used to test the significance of the orthogonality between initial state and peak state for individual stimuli (Fig. S9). A procedure analogous to the one outlined above has been applied in this case. Here, instead of the subspace overlaps, the quantity of interest is the correlation given by the ordinary scalar product. For the first control shuffled data are obtained by shuffling labels "initial state" and "peak state" across trials. Significance levels were evaluated as outlined above for the first control (Fig. S9A). To test for the second hypothesis, we computed correlations between activity vectors defined at the same time point, but averaged over two different sets of 10 trials each. For each trial permutations we computed these correlations for all time points belonging to the OFF response (35 time steps) and average over the time points. Significance level was then assessed as outlined above for the second control (Fig. S9B).

1.7 Linear dynamical system fit

To fit the linear system $\dot{\mathbf{X}} = \mathbf{X}(\mathbf{J} - \mathbf{I})$ we first computed the velocity of the trajectory as

$$\dot{\mathbf{X}} = \frac{\mathbf{X}(t_{i+1}) - \mathbf{X}(t_i)}{t_{i+1} - t_i}. \tag{13}$$

We used ridge regression and we constrained the rank of the matrix $\mathbf{J}$ by using reduced rank regression (Izenman, 1975; Davies and Tso, 1982). Since ridge regression involves computationally expensive matrix inversion, we first reduced the dimensionality of the original dataset by using principal component analysis and kept a number K of component such that the explained variance was over 90% ($K \gtrsim 100$). Thus, here the data matrix $\mathbf{X} \in \mathbb{R}^{TC\times K}$ contains the activity across T time bins along all K dimensions for a number C of stimuli.

Reduced rank ridge regression Reduced rank regression problems aim at minimizing the squared error $||\dot{\mathbf{X}} - \mathbf{X}(\mathbf{J} - \mathbf{I})||^2$ under a rank constraint on the matrix $\mathbf{J}$, i.e. $\operatorname{rank}\mathbf{J} \leq R$, where R is a hyperparameter of

the model. Instead, ridge regression introduces a penalty for large entries of the matrix $\mathbf{J}$ by minimizing the cost function $||\dot{\mathbf{X}} - \mathbf{X}(\mathbf{J}-\mathbf{I})||^2 + \lambda||(\mathbf{J}-\mathbf{I})||^2$, where λ is the second hyperparameter of the model. Here we combined ridge regression and reduced rank regression in the same framework. To define the problem, it is useful to write the ridge regression optimization problem as:

$$(\mathbf{J}-\mathbf{I})^*_\lambda = \underset{\mathbf{J}-\mathbf{I}}{\operatorname{argmin}} ||\dot{\mathbf{X}}_\lambda - \mathbf{X}_\lambda(\mathbf{J}-\mathbf{I})||^2 \tag{14}$$

where we defined $\dot{\mathbf{X}}_\lambda = (\dot{\mathbf{X}}, \mathbf{0})$ and $\mathbf{X}_\lambda = (\mathbf{X}, \sqrt{\lambda}\mathbf{I})$. The reduced rank ridge regression problem with hyperparameters R and λ is therefore defined by (Mukherjee et al., 2015):

$$(\mathbf{J}-\mathbf{I})^*_{r,\lambda} = \underset{\operatorname{rank}\mathbf{J}\leq R}{\operatorname{argmin}} ||\dot{\mathbf{X}}_\lambda - \mathbf{X}_\lambda(\mathbf{J}-\mathbf{I})||^2 \tag{15}$$

To solve Eq. (15) we used the solution to Eq. (14) given by:

$$\mathbf{J}^*_\lambda = \mathbf{I} + (\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T\dot{\mathbf{X}}_\lambda. \tag{16}$$

Supposing that the matrix $\mathbf{X}_\lambda\mathbf{J}^*_\lambda$ has singular value decomposition given by $\mathbf{X}_\lambda\mathbf{J}^*_\lambda = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then it can be shown that the solution to the reduced rank ridge regression problem Eq. (15) can be written as:

$$\mathbf{J}^*_{r,\lambda} = \mathbf{J}^*_\lambda \sum_{i=1}^{R} \mathbf{V}_i\mathbf{V}_i^T \tag{17}$$

We note that each term in the sum of Eq. (17) has unit rank, so that the resulting matrix $\mathbf{J}^*_{R,\lambda}$ has rank equal to R.

Selection of hyperparameters To find the hyperparameters λ and R, we fitted the linear dynamical system to the data with hyperparameters (λ, R) and computed the goodness of the fit $R^2(\lambda, R)$ using cross-validation. We repeated the process for a range of values of (λ, R). We observed that, independently of the value of λ, the function $R^2(\cdot, R)$ saturates at a particular value of the rank R^*, but does not exhibit a clear maximum. We took the value R^* as the minimal rank hyperparameter, while we defined the best ridge parameter as $\lambda^* = \operatorname{argmax}_\lambda R^2(\lambda, R^*)$ (Fig. S4).

We used K-fold cross-validation, with $K = 10$. When fitting multiple stimuli at once, for each stimulus we partitioned the temporal activity into K chunks, resulting in a total of KC chunks. At the i-th iteration of the cross-validation procedure, we leave out the i-th partition for each stimulus to construct the training set ($(K-1)C$ folds) and test the fit on the remaining C folds.

1.8 Linear dynamical fit on PCA-reduced data

In this section we examine the connectivity matrix that results from fitting a dynamical system to population responses on which dimensionality reduction (PCA) has been applied in a previous step. Suppose that $\mathbf{Y} \in \mathbb{R}^{TC\times N}$ contains the population dynamics of N neurons across T timepoints and C stimuli and is generated by recurrent dynamics, therefore satisfying the equation

$$\dot{\mathbf{Y}} = \mathbf{Y}(\mathbf{M}-\mathbf{I}), \tag{18}$$

where $\mathbf{M}$ represents the true connectivity matrix. Let $\mathbf{Q}$ be the orthogonal $N \times N$ matrix containing the principal component of the population dynamics given by $\mathbf{Y}$. We reduce the dimensionality of $\mathbf{Y}$ by projecting the dynamics onto the first K principal components as:

$$\mathbf{X}_{:,0:K} = \mathbf{Y}\mathbf{Q}_{:,0:K} \tag{19}$$

where the notation $:, 0:K$ denotes the first K columns of the matrix. We then perform the linear dynamical system fit on the reduced dynamics given by the matrix $\mathbf{X}_{:,0:K}$ by fitting the equation:

$$\dot{\mathbf{X}}_{:,0:K} = \mathbf{X}_{:,0:K}(\mathbf{J}-\mathbf{I}) \tag{20}$$

In the following, we derive the relation between the connectivity matrix $\mathbf{J}$ resulting from fitting Eq. (20) and the full connectivity $\mathbf{M}$. We first find the expression of the velocity of the population trajectory given by $\mathbf{X}$. Using Eq. (19) and Eq. (18) and the orthogonality of $\mathbf{Q}$, we write:

$$\dot{\mathbf{X}} = \mathbf{X}\mathbf{Q}^T(\mathbf{M} - \mathbf{I})\mathbf{Q} \tag{21}$$

As in Eq. (20), we want to express $\dot{\mathbf{X}}_{:,0:K}$ as a function of $\mathbf{X}_{:,0:K}$. Using Eq. (21) we obtain:

$$\dot{\mathbf{X}}_{:,0:K} = \mathbf{X}_{:,0:K}\,(\mathbf{Q}^T_{0:K,:}\,\mathbf{M}\,\mathbf{Q}_{:,0:K} - \mathbf{I}) + \boldsymbol{\epsilon} \tag{22}$$

where the second term on the right hand side is given by $\epsilon = \mathbf{X}_{:,K:N}\,\mathbf{Q}^T_{K:N,:}\,(\mathbf{M} - \mathbf{I})\,\mathbf{Q}_{,0:K}$ and acts effectively as noise. We therefore conclude that, assuming that the global dynamics is generated by a dynamical system, the connectivity matrix $\mathbf{J}$ obtained by fitting the dynamics onto the first K principal components (Eq. (20)) corresponds to the projection of the original connectivity $\mathbf{M}$ onto the first K principal components (Fig. S7):

$$\mathbf{J} \simeq \mathbf{J}_{\mathrm{PC}} = \mathbf{Q}^T_{0:K,:}\,\mathbf{M}\,\mathbf{Q}_{:,0:K} \tag{23}$$

1.9 Control datasets

As control dataset for the dynamical system hypothesis, we employed a recent method based on a maximum entropy model (Savin and Tkačik, 2017) and described in (Elsayed and Cunningham, 2017). This method, called Tensor Maximum Entropy, allows to construct surrogate datasets that are maximally random (entropy maximization) but constrained so that their marginal means and covariances are the same as the original dataset.

Marginal means and covariances Let the temporal activity along all K dimensions for all C stimuli be organized in a tensor $\mathbf{Z} \in \mathbb{R}^{T\times K\times C}$. The mean tensor $\mathbf{M}$ is defined as the tensor that makes all the marginal means of $\mathbf{Z}$ vanish. Specifically, if $\bar{\mathbf{Z}} = \mathbf{Z} - \mathbf{M}$, the tensor $\mathbf{M}$ is such that:

$$\sum_{k=1}^{K}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tkc} = 0, \quad \sum_{t=1}^{T}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tkc} = 0, \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\bar{\mathbf{Z}}_{tkc} = 0 \tag{24}$$

The marginal covariances of the tensor $\bar{\mathbf{Z}}$ across times, neural dimensions and stimuli are therefore defined as:

$$\begin{cases} \boldsymbol{\Sigma}^T_{ij} = \displaystyle\sum_{k=1}^{K}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{ikc}\bar{\mathbf{Z}}_{jkc} \\ \boldsymbol{\Sigma}^K_{ij} = \displaystyle\sum_{t=1}^{T}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tic}\bar{\mathbf{Z}}_{tjc} \\ \boldsymbol{\Sigma}^C_{ij} = \displaystyle\sum_{t=1}^{T}\sum_{k=1}^{K}\bar{\mathbf{Z}}_{tki}\bar{\mathbf{Z}}_{tkj} \end{cases} \tag{25}$$

Tensor maximum entropy method The method generates the desired number of surrogate datasets $\mathbf{S}^{(i)} \in \mathbb{R}^{T\times K\times C}$. Each of these surrogates is randomly drawn from a probability distribution that assumes *a priori* no structure apart from that expected from the marginal means and covariances of the original data. Let $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}^T$, $\boldsymbol{\Lambda}^K$ and $\boldsymbol{\Lambda}^C$ be the marginal means and covariances of surrogate $\mathbf{S}$. The method computes the probability $P(\mathbf{S})$ over the surrogates that maximizes the entropy function

$$P(\mathbf{S}) = \underset{y(\mathbf{S})}{\operatorname{argmax}}\Big[-\int_{\mathbf{S}} y(\mathbf{S})\log y(\mathbf{S})\mathrm{d}\mathbf{S}\Big], \qquad \text{with} \int_{\mathbf{S}} P(\mathbf{S})\mathrm{d}\mathbf{S} = 1 \tag{26}$$

subject to the constraints

$$\mathbb{E}_P[\boldsymbol{\mu}] = \mathbf{M}, \quad \mathbb{E}_P[\Lambda^T] = \boldsymbol{\Sigma}^T, \quad \mathbb{E}_P[\Lambda^K] = \boldsymbol{\Sigma}^K, \quad \mathbb{E}_P[\Lambda^C] = \boldsymbol{\Sigma}^C, \tag{27}$$

where $\mathbb{E}_P[\cdot]$ denotes the expectation over the probability density P. We used three types of surrogate datasets, denoted as T, TK and TKC. All the three types of surrogates obey to the first constraint in Eq. (27) on the marginal means. In addition, surrogates of type T obey the constraint on the time covariances, surrogates of type TK on time and dimension covariance, while surrogates TKC obey all the three covariance constraints.

1.10 Analysis of the transient channels

Let $\mathbf{J}_{\text{Full}}$ be the connectivity matrix resulting from fitting the recurrent model to the responses to all stimuli at once, and $\mathbf{J}^{(s)}$ the connectivity obtained from fitting the model to the response to stimulus s (we imposed different values of the rank parameter R^* for the two matrices: we set $r = 70$ for $\mathbf{J}_{\text{Full}}$ and $r = 6$ for $\mathbf{J}^{(s)}$, see Fig. 4). Using the singular value decomposition, we can write the matrices $\mathbf{J}^{(s)}$ as:

$$\mathbf{J}^{(s)} = \hat{\mathbf{U}}^{(s)}\mathbf{\Sigma}^{(s)}\mathbf{V}^{(s)T}, \tag{28}$$

where $\hat{\mathbf{U}}^{(s)}$ and $\mathbf{V}^{(s)}$ contain respectively the R_s right and left connectivity vectors as columns, while $\mathbf{\Sigma}^{(s)}$ contains the norms of $\mathbf{U}^{(s)}$ on the diagonal (see Eq. (3)). The transient dynamics elicited by stimulus s has a strong component along a single dimension specified by the right connectivity vectors $\mathbf{u}^{(r;s)}$ (see also Eq. (35)). We therefore define the overlap between the transient channels corresponding to pairs of stimuli s_1 and s_2 (Fig. 4H *left panel*) as the principal angle between the subspaces defined by $\hat{\mathbf{U}}^{(s_1)}$ and $\hat{\mathbf{U}}^{(s_2)}$ (see Eqs. (11)-(12)).

To test whether the connectivity fitted on all stimuli at once $\mathbf{J}_{\text{Full}}$ consisted of the sum of low-rank transient coding channels, we defined the matrix $\mathbf{J}_{\text{Sum}}$ as the sum of the individual transient channels for all stimuli (see Eq. (3)):

$$\mathbf{J}_{\text{Sum}} = \hat{\mathbf{U}}^{(1)}\mathbf{\Sigma}^{(1)}\mathbf{V}^{(1)T} + \hat{\mathbf{U}}^{(2)}\mathbf{\Sigma}^{(2)}\mathbf{V}^{(2)T} + ... + \hat{\mathbf{U}}^{(P)}\mathbf{\Sigma}^{(P)}\mathbf{V}^{(P)T} \tag{29}$$

We then compared goodness of the fit of the population OFF responses using the full matrix $\mathbf{J}_{\text{Full}}$ and the matrix $\mathbf{J}_{\text{Sum}}$ (Fig. 4H *right panel*).

2 The network model

We study a recurrent network of N randomly coupled linear rate units. Each unit i is described by the time-dependent variable $r_i(t)$, which represents the difference between the firing rate of neuron i at time t and its baseline firing level $r_{i,B}$. The equation governing the temporal dynamics of the network reads:

$$\tau \dot{r}_i = -r_i + \sum_{j=1}^{N} J_{ij} r_j, \tag{30}$$

where τ represents the membrane time constant (fixed to unity), and J_{ij} is the effective synaptic strength from neuron j to neuron i. The system has only one fixed point corresponding to $r_i = 0$ for all i. To have stable dynamics, we require that the real part of the eigenvalues of the connectivity matrix $\mathbf{J}$ is smaller than unity, i.e. $\mathfrak{Re}\lambda_{\max}(\mathbf{J}) < 1$.

We model each stimulus as the state of the system reached at the end of stimulus presentation, which we denote by the vector $\mathbf{r}_0$. This is equivalent to setting the initial condition of the dynamics to $\mathbf{r}(0) = \mathbf{r}_0$. The dynamics following a specific initial condition $\mathbf{r}_0$ corresponds therefore to the OFF response to the stimulus associated with the state $\mathbf{r}_0$ at the time of stimulus offset. We assume that during the OFF response the network receives no external input. Under these conditions the linear dynamics following the initial state $\mathbf{r}_0$ can be formally written as $\mathbf{r}(t) = \exp(t(\mathbf{J} - \mathbf{I}))\mathbf{r}_0$. The time-dependent matrix $\exp(t(\mathbf{J} - \mathbf{I}))$ is called the propagator of the system, and maps the state of the network at time $t = 0$ into the state at time t.

2.1 Normal and non-normal connectivity matrices

We define amplified OFF responses in recurrent networks by focusing on the temporal dynamics of the distance from baseline, defined as the norm of the population activity vector $||\mathbf{r}(t)||$ (Hennequin et al., 2014). The network generates an amplified OFF response to the stimulus associated with the initial condition $\mathbf{r}_0$ when the value of $||\mathbf{r}(t)||$ transiently increases before decaying to its baseline level $||\mathbf{r}_B||$. Note that having a transiently increasing value of the distance from baseline implies that the OFF response $r_i(t)$ of at least one unit displays non-monotonic temporal dynamics. Importantly, the transient behaviour of $||\mathbf{r}(t)||$ depends on the stimulus through $\mathbf{r}_0$, and on the properties of the connectivity matrix $\mathbf{J}$, in particular on the relationship between its eigenvectors.

Connectivity matrices for which the eigenvectors are orthogonal to each other are called *normal* matrices and they are defined by the equation $\mathbf{J}\mathbf{J}^T = \mathbf{J}^T\mathbf{J}$. Networks with normal connectivity cannot produce amplified OFF responses, as defined by a transiently increasing $||\mathbf{r}(t)||$. In such networks, any stimulus $\mathbf{r}_0$

evokes an OFF response for which the distance from baseline decays monotonically to the baseline level. Note that any symmetric matrix is normal.

On the other hand, connectivity matrices for which some eigenvectors are not mutually orthogonal are called *non-normal* (Trefethen and Embree, 2005), and they consist of all connectivity $\mathbf{J}$ for which $\mathbf{J}\mathbf{J}^T \neq \mathbf{J}^T\mathbf{J}$. It is well known that non-normal networks can lead to transiently increasing values of $||\mathbf{r}(t)||$, therefore producing amplified OFF responses. However, the non-normality of the network connectivity $\mathbf{J}$ constitutes only a necessary but not a sufficient condition for the generation of amplified responses.

2.2 Sufficient condition for amplified OFF responses

To find the necessary and sufficient condition for the generation of amplified responses in linear recurrent networks, we write the differential equation governing the dynamics of the distance from baseline as (Neubert and Caswell, 1997; Bondanelli and Ostojic, 2020):

$$\frac{1}{||\mathbf{r}||}\frac{\mathrm{d}||\mathbf{r}||}{\mathrm{d}t} = \frac{\mathbf{r}^T(\mathbf{J}_S - \mathbf{I})\mathbf{r}}{||\mathbf{r}||^2}, \qquad \mathbf{J}_S = \frac{\mathbf{J} + \mathbf{J}^T}{2}, \tag{31}$$

where we denote by $\mathbf{J}_S$ the symmetric part of the connectivity $\mathbf{J}$. A linear recurrent network exhibits amplified responses when the rate of change of the distance from baseline, $\mathrm{d}||\mathbf{r}||/\mathrm{d}t$, takes positive values at time $t = 0$. The right hand side of Eq. (31) takes its largest value when the initial condition $\mathbf{r}_0$ is aligned with the eigenvector of $\mathbf{J}_S$ associated with the largest eigenvalue $\lambda_{\max}(\mathbf{J}_S)$. In this case, the rate of change of the distance from baseline at time $t = 0$ takes the value $\lambda_{\max}(\mathbf{J}_S) - 1$. From Eq. (31) it is possible to show that the necessary and sufficient condition for the generation of amplified responses in a recurrent networks with connectivity $\mathbf{J}$ is given by

$$\lambda_{\max}(\mathbf{J}_S) > 1 \tag{32}$$

This criterion defines two classes of networks based on the properties of the connectivity matrix: networks in which amplified responses cannot be evoked by any stimulus, and networks able to generate amplified responses to at least one stimulus.

2.3 Low-rank networks

In the following section we apply the results outlined above, valid for any linear recurrent network, to a network model with low-rank connectivity of the form given by Eq. (2).

Low-dimensional dynamics We consider low-rank networks in which the initial state is set to $\mathbf{r}(0) = \mathbf{r}_0$ and no external input acts on the system at later times $t > 0$. By construction, the autonomous dynamics generated by low-rank networks are constrained by the rank R of the connectivity matrix and are therefore low-dimensional when $R \ll N$. We first illustrate the linear dynamics in the case of a unit-rank connectivity ($R = 1$), given by $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$. The vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ are respectively the right and left eigenvectors of $\mathbf{J}$, and in the following we refer to them respectively as the right and left connectivity vectors. In this case the autonomous dynamics following the initial state $\mathbf{r}_0$ can be explicitly written as (Bondanelli and Ostojic, 2020):

$$\mathbf{r}(t) = e^{-t}\,\mathbf{r}_0 + e^{-t}a(t)\,\mathbf{u}^{(1)}, \tag{33}$$

where $a(t) = (\mathbf{v}^{(1)T}\mathbf{r}_0)\,(e^{\lambda t} - 1)/\lambda$, and λ is the only non-zero eigenvalue of $\mathbf{J}$, given by

$$\lambda = ||\mathbf{u}^{(1)}||\cos\theta, \quad \cos\theta = \frac{\mathbf{u}^{(1)}\cdot\mathbf{v}^{(1)}}{||\mathbf{u}^{(1)}||||\mathbf{v}^{(1)}||}. \tag{34}$$

Since $\mathbf{u}^{(1)}$ is the right eigenvector of $\mathbf{J}$ corresponding to λ, from Eq. (33) we note that, when $\mathbf{r}_0$ is fully aligned with $\mathbf{u}^{(1)}$ the dynamics are one-dimensional and exhibit a monotonic decay along the same dimension. Instead, when the initial state is not fully aligned with $\mathbf{u}^{(1)}$, the dynamics are confined to the plane defined by $\mathbf{r}_0$ and $\mathbf{u}^{(1)}$. In this case, while the component of the dynamics along $\mathbf{r}_0$ decays exponentially as a function of time, the component along the direction of $\mathbf{u}^{(1)}$ increases initially in proportion to the value of its norm, $||\mathbf{u}^{(1)}||$ (since at time zero, $\mathrm{d}e^{-t}a(t)/\mathrm{d}t = ||\mathbf{u}^{(1)}||$). Eventually, the activity decays to the baseline level given by $\mathbf{r}_B = \mathbf{r}(t \to \infty) = 0$. Therefore in a unit-rank network the dynamics draw a neural trajectory that explores at most two-dimensions in the state space.

These observations can be extended to the general case of rank-R connectivity matrices. For simplicity we re-write Eq. (2) as $\mathbf{J} = \mathbf{U}\mathbf{V}^T$, where the matrices $\mathbf{U}$ and $\mathbf{V}$ contain respectively the R right and left connectivity vectors as columns. The linear dynamics following the initial state $\mathbf{r}_0$ can be written as

$$\begin{aligned}\mathbf{r}(t) = \exp(t(\mathbf{J}-\mathbf{I}))\mathbf{r}_0 &= e^{-t}\left(\mathbf{I} + \sum_{m=1}^{\infty} \frac{(\mathbf{U}\mathbf{V}^T)^m}{m!}\right)\mathbf{r}_0 \\ &= e^{-t}\,\mathbf{r}_0 + e^{-t}\,\mathbf{U}\,\mathbf{a}(t),\end{aligned} \tag{35}$$

where we defined the column vector $\mathbf{a}(t) = \left(\mathbf{V}^T\mathbf{U}\right)^{-1}\left[\exp(t\mathbf{V}^T\mathbf{U}) - \mathbf{I}\right]\left(\mathbf{V}^T\mathbf{r}_0\right)$ in analogy with Eq. (33). Therefore, in the case of rank-R connectivity matrix, the dynamics evolve in a $(R+1)$-dimensional space specified by the initial state $\mathbf{r}_0$ and the right connectivity vectors $\mathbf{u}^{(r)}$ (columns of $\mathbf{U}$). We conclude that low-rank connectivity matrices of the form given by Eq. (2) with $R \ll N$ are sufficient to generate low-dimensional dynamics that explore at most $R+1$ dimensions.

Conditions for stable amplified OFF responses in low-rank networks In this paragraph we examine the conditions required to generate stable amplified dynamics in networks with low-rank connectivity $\mathbf{J}$ and initial state $\mathbf{r}_0$. Specifically, two sets of conditions need to be satisfied. The first set of conditions is directly derived by applying the general criterion for stable amplified dynamics given by Eq. (32) to low-rank networks, and effectively constrains the connectivity $\mathbf{J}$ through the relative arrangement of the connectivity vectors $\mathbf{u}^{(r)}$ and $\mathbf{v}^{(r)}$ $(r = 1, ..., R)$. When this criterion is satisfied, amplified dynamics can be generated only if the initial condition $\mathbf{r}_0$ is aligned with specific directions in the state-space. We thus examine a second set of conditions on the initial state $\mathbf{r}_0$ for which amplified trajectories can be evoked, and express these conditions in terms of relationship between $\mathbf{r}_0$ and the modes $\mathbf{u}^{(r)}$-$\mathbf{v}^{(r)}$. Without loss of generality, we assume that the norm of the vectors $\mathbf{v}^{(r)}$ is equal to one, while the norm of the vectors $\mathbf{u}^{(r)}$ can vary.

Conditions on the modes $\mathbf{u}^{(r)}$-$\mathbf{v}^{(r)}$ We first consider a network with unit-rank connectivity $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$. In unit-rank networks, the dynamics is stable only if the non-zero eigenvalue λ is smaller than one. This yields the stability condition

$$\cos\theta < \frac{1}{||\mathbf{u}^{(1)}||}, \tag{36}$$

where θ is given by Eq. (34). The condition for the generation of amplified responses given by Eq. (32) can be derived in terms of $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ by computing the eigenvalues of the symmetric part of the connectivity, given by $\mathbf{J}_S = (\mathbf{u}^{(1)}\mathbf{v}^{(1)T} + \mathbf{v}^{(1)}\mathbf{u}^{(1)T})/2$. The matrix $\mathbf{J}_S$ has two non-trivial eigenvalues, given by $\lambda_{\max,\min}(\mathbf{J}_S) = ||\mathbf{u}^{(1)}||(\cos\theta \pm 1)/2$. Therefore the condition for amplified OFF responses Eq. (32) can be written in terms of the length of the vector $\mathbf{u}^{(1)}$ as

$$||\mathbf{u}^{(1)}|| > \frac{2}{\cos\theta + 1}. \tag{37}$$

Previous analyses have shown that amplification in a unit-rank network can take arbitrarily large values only if $0 \leq \cos\theta < 1/||\mathbf{u}^{(1)}||$ (Bondanelli and Ostojic, 2020). As a result, stable and unbounded amplified dynamics can be generated in a unit-rank network if and only if the norm of $\mathbf{u}^{(1)}$ is sufficiently large (Eq. (37)) and the correlation between the connectivity vectors is positive and sufficiently small (Eq. (36)). In the case where $\cos\theta = 0$, Eq. (37) reduces to $||\mathbf{u}^{(1)}|| > 2$, and the dynamics of the network can be written as:

$$\mathbf{r}(t) = e^{-t}\,\mathbf{r}_0 + te^{-t}\,\mathbf{u}^{(1)}\mathbf{v}^{(1)T}\mathbf{r}_0. \tag{38}$$

For a general rank-R connectivity matrix $\mathbf{J} = \mathbf{U}\mathbf{V}^T$, deriving necessary and sufficient conditions for the stability and amplification of the dynamics remains an open problem, as it requires evaluating the eigenvalues of $\mathbf{J}$ and $\mathbf{J}_S$ which depend in a complex manner on the connectivity vectors $\mathbf{u}^{(r)}$ and $\mathbf{v}^{(r)}$. Nevertheless sufficient conditions for stable amplification can be obtained. To derive a sufficient condition for stability we write Eq. (2) as $\mathbf{J} = \hat{\mathbf{U}}\mathbf{S}\mathbf{V}^T$, where $\hat{\mathbf{U}}$ contains the normalized right connectivity vectors $\mathbf{u}^{(r)}/||\mathbf{u}^{(r)}||$ and $\mathbf{S} = \mathrm{diag}(||\mathbf{u}^{(1)}||, ..., ||\mathbf{u}^{(R)}||)$. We define the subspace overlap θ_P between the linear subspace spanned by $\hat{\mathbf{U}}$ and the one spanned by $\mathbf{V}$ as in Eq. (11):

$$\cos\theta_P = \max_{\mathbf{x}\in\hat{U},\mathbf{y}\in V} \mathbf{x}^T\mathbf{y} = \sigma_{\max}\left(\mathbf{V}^T\hat{\mathbf{U}}\right). \tag{39}$$

A sufficient condition for stability can be obtained by bounding the modulus of the largest eigenvalue to one. All the eigenvectors of a rank-R matrix $\mathbf{J}$ can be written as linear combinations of the normalized left connectivity vectors. Therefore, for the eigenvector corresponding to the largest eigenvalue we can write $\mathbf{e} = \hat{\mathbf{U}}\mathbf{\Lambda}$, where $\mathbf{\Lambda}^T = (\alpha_1, \alpha_2, ...)$ with $\sum_k \alpha_k^2 = 1$. The sufficient condition for stability then becomes $||\mathbf{J}\hat{\mathbf{U}}\mathbf{\Lambda}|| < 1$, where $||\cdot||$ denotes the Frobenius norm (Horn and Johnson, 2012). By the property of matrix norms (Horn and Johnson, 2012) we can write:

$$\begin{aligned} ||\mathbf{J}\hat{\mathbf{U}}\mathbf{\Lambda}|| &= ||\hat{\mathbf{U}}\mathbf{S}\mathbf{V}^T\hat{\mathbf{U}}\mathbf{\Lambda}|| \leq ||\hat{\mathbf{U}}||\,||\mathbf{S}||\,||\mathbf{V}^T\hat{\mathbf{U}}|| = \\ &= \sqrt{R}\,||\mathbf{S}||\sqrt{\sigma_{\max}(\mathbf{V}^T\hat{\mathbf{U}})^2 + \sigma_2(\mathbf{V}^T\hat{\mathbf{U}})^2 + ... + \sigma_R(\mathbf{V}^T\hat{\mathbf{U}})^2} \leq R||\mathbf{S}||\cos\theta_P \end{aligned} \tag{40}$$

where for the first and second inequality we used respectively the fact that $||\mathbf{\Lambda}|| = 1$ and the definition given in Eq. (39). From Eq. (40) we can write the sufficient condition for stability in terms of the principal angle between the subspaces defined by $\hat{\mathbf{U}}$ and $\mathbf{V}$ as

$$\cos\theta_P < \frac{1}{R||\mathbf{S}||} = \frac{1}{R\sqrt{||\mathbf{u}^{(1)}||^2 + ||\mathbf{u}^{(2)}||^2 + ... + ||\mathbf{u}^{(R)}||^2}}. \tag{41}$$

By following the same steps as in Eq. (40), we can write a less stringent sufficient condition for the case where $||\mathbf{u}^{(i)}|| = \Delta$ for all i, given by $\cos\theta_P < 1/R\Delta$.

Equation (41) provides a sufficient condition for the stability of the system in terms of the correlations between all possible pairs of left and right connectivity vectors, reflected in the principal angle θ_P. We next examine the case in which at least one right connectivity pattern $\mathbf{u}^{(m)}$ has large norm $||\mathbf{u}^{(m)}||$, and specifically $R||\mathbf{u}^{(m)}|| \gg 1$. The sufficient condition for stability Eq. (41) then implies that the overlaps between all pairs of left and right connectivity vectors should be sufficiently small ($\cos\theta_P \ll 1$). This condition is in particular satisfied when all vectors are mutually orthogonal, in which case Eq. (35) becomes

$$\mathbf{r}(t) = e^{-t}\,\mathbf{r}_0 + te^{-t}\,\mathbf{S}\hat{\mathbf{U}}\mathbf{V}^T\mathbf{r}_0. \tag{42}$$

Equation (42) is effectively equivalent to the dynamics of R uncoupled rank-1 channels, each with dynamics given by Eq. (38). As a result, amplification is controlled by the norms of the right connectivity vectors, $||\mathbf{u}^{(r)}||$ ($r = 1, ..., R$), and can therefore exhibit amplification if $||\mathbf{u}^{(m)}|| \gg 2$ for at least one connectivity pattern $\mathbf{u}^{(m)}$. Thus, Eq. (41) provides a link between the amplification of the system, given in terms of the norms $||\mathbf{u}^{(r)}||$ and the stability of the network. Importantly, Eq. (41) is a sufficient but not necessary condition, and represents therefore a very stringent constraint on the reciprocal overlap between pairs of left and right connectivity vectors.

Conditions on the initial state $\mathbf{r}_0$ In this paragraph we examine the conditions on the initial state $\mathbf{r}_0$ for having amplified dynamics. In the following analysis we assume that the correlation between all pairs of left and right connectivity vectors is close to zero (Eq. (41)).

We first consider a unit-rank network. If Eq. (36) holds, the temporal dynamics of the network can be approximated by Eq. (38). Under these conditions, the initial state $\mathbf{r}_0$ that elicits the most amplified dynamics is aligned with the left connectivity vector $\mathbf{v}^{(1)}$, and the level of amplification scales as the norm of the right connectivity vector, $||\mathbf{u}^{(1)}||$ (Bondanelli and Ostojic, 2020). Adding to $\mathbf{r}_0$ a component orthogonal to the vector $\mathbf{v}^{(1)}$ may change the overall level of amplification, but the network may still be able to generate amplified dynamics. Let the initial state be given by $\mathbf{r}_0 = \alpha\,\mathbf{v}^{(1)} + \beta\,\mathbf{u}^{(1)} + \gamma\,\mathbf{z}^{(\perp)}$, where $\mathbf{z}^{(\perp)}$ is a vector orthogonal to both left and right connectivity vectors, and $\alpha^2 + \beta^2 + \gamma^2 = 1$. The norm of the population activity vector can then be written as

$$||\mathbf{r}(t)||^2 = e^{-2t}\alpha^2 + e^{-2t}(\beta + \alpha t||\mathbf{u}^{(1)}||)^2 + e^{-2t}\gamma^2 \tag{43}$$

For simplicity we consider the amplification at time $t^* = 1$, corresponding to the peak amplification time when $\beta \approx \gamma \approx 0$ and $||\mathbf{u}^{(1)}|| \gg 2$ (strongly amplified dynamics; Bondanelli and Ostojic (2020)). By requiring that the norm of the population activity vector evaluated at time $t^* = 1$ is larger than unity, we can obtain a sufficient condition on the initial state $\mathbf{r}_0$ to generate amplified dynamics in terms of its component on the left connectivity vector:

$$\alpha^2||\mathbf{u}^{(1)}||^2 + 2\alpha\beta > e^2 - 1. \tag{44}$$

which is satisfied when

$$|\alpha| \gtrsim \frac{\sqrt{e^2-1}}{||\mathbf{u}^{(1)}||} \tag{45}$$

Equation (45) shows that in a strongly amplified network (large $||\mathbf{u}^{(1)}||$) the initial state $\mathbf{r}_0$ can have a significant component orthogonal to both connectivity vectors and yet be able to generate amplified dynamics. However, from Eq. (43) we observe that for a fixed value of $||\mathbf{u}^{(1)}||$ the amplification decreases when the component along $\mathbf{v}^{(1)}$, i.e. α, decreases. Therefore, to have amplification the component of $\mathbf{r}_0$ along $\mathbf{u}^{(1)}$ should be sufficiently strong (e.g. from Eq. (45) we have that $\alpha \gtrsim 0.6$ for a moderately amplified network with $||\mathbf{u}^{(1)}|| = 4$).

By following the same steps we can generalize the analysis above to the case of a rank-R network with mutually orthogonal left and right connectivity vectors. Indeed, if Eq. (41) holds, we can approximate the dynamics of the system with Eq. (42). To examine how the amplification of the system changes as a function of the initial state, we consider the case given by $\mathbf{r}_0 = \alpha_1\mathbf{v}^{(1)} + ... + \alpha_R\mathbf{v}^{(R)} + \beta_1\mathbf{u}^{(1)} + ... + \beta_R\mathbf{u}^{(R)} + \gamma\mathbf{z}^{(\perp)}$, where $\mathbf{z}^{(\perp)}$ is a vector orthogonal to all the left and right connectivity vectors. The norm of the population activity vector with initial condition $\mathbf{r}_0$ can be written as

$$||\mathbf{r}(t)||^2 = e^{-2t}\sum_{r=1}^{R}\left[\alpha_r^2 + (\beta_r + \alpha_r t||\mathbf{u}^{(r)}||)^2 + \gamma^2\right] \tag{46}$$

We consider the amplification at time $t^* = 1$. Asking that the norm Eq. (46) evaluated at time $t^* = 1$ is larger than unity yields

$$\sum_r \left[\alpha_r^2||\mathbf{u}^{(r)}||^2 + 2\alpha_r\beta_r||\mathbf{u}^{(r)}||\right] > e^2 - 1 \tag{47}$$

Equation (47) is analogous to Eq. (44) for the rank-1 network, and provides an implicit condition over the α_r's, i.e. the correlations between the initial state $\mathbf{r}_0$ and the left connectivity vectors $\mathbf{v}^{(r)}$'s. As in the unit-rank network, in a strongly amplified rank-R system ($||\mathbf{u}^{(r)}|| \gg 2$), the initial state $\mathbf{r}_0$ can have a significant component orthogonal to the left connectivity vectors $\mathbf{v}^{(r)}$ and yet be able to generate amplified dynamics. From Eq. (46) we note that, for fixed values of the $||\mathbf{u}^{(r)}||$, amplification decreases by decreasing the components on the right connectivity vectors. Therefore these components should be nonetheless sufficiently strong to generate amplified dynamics, thus requiring sufficiently high correlation between $\mathbf{r}_0$ and the $\mathbf{v}^{(r)}$'s (e.g. assuming $\beta_r \approx 0$ and $||\mathbf{u}^{(r)}|| = \Delta$ for all r, from Eq. (47) we have that $(\sum_r \alpha_r^2)^{1/2} \gtrsim 0.6$ for an amplified network with $\Delta = 4$).

2.4 Correlation between structure of initial states and structure of the dynamics

In this section we show that, in the low-rank network model, the pattern of correlations between initial states across different stimuli predicts the patterns of correlations between the corresponding dynamical trajectories. To examine the structure of the dynamics across stimuli we focus on the correlation between the peak states of different trajectories. We study the case of two trajectories generated by two correlated stimuli, modeled as two correlated initial conditions $\mathbf{r}_0^{(1)}$ and $\mathbf{r}_0^{(2)}$. We assume that Eq. (41) is satisfied, so that the dynamics can be written as in Eq. (42), and that the initial state is strongly aligned with the left connectivity vectors $\mathbf{v}^{(r)}$(see Materials and methods). We set the first initial state $\mathbf{r}_0^{(1)}$ to a linear combination of the left connectivity vectors $\mathbf{v}^{(r_1;1)}$ belonging to a given channel:

$$\begin{aligned}\mathbf{r}_0^{(1)} &= \mathbf{V}^{(1)}\mathbf{A}^{(1)} \\ &= \sum_{r_1=1}^{R_1}\mathbf{v}^{(r_1;1)}\alpha_{r_1}^{(1)}.\end{aligned} \tag{48}$$

where the matrix $\mathbf{V}^{(1)}$ contains the vectors $\mathbf{v}^{(1,r_1)}$, and $\mathbf{A}^{(1)} = (\alpha_1^{(1)}, \alpha_1^{(1)}, ...)^T$ with $\alpha_1^{(1)2} + \alpha_2^{(1)2} + ... = 1$. We set the second initial state to a linear combination of the right connectivity vectors belonging to two orthogonal channels $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$:

$$
\begin{aligned}
\mathbf{r}_0^{(2)} &= \mathbf{V}^{(1)}\mathbf{B}^{(1)} + \mathbf{V}^{(2)}\mathbf{B}^{(2)} \\
&= \sum_{r_1=1}^{R_1} \mathbf{v}^{(r_1;1)}\beta_{r_1}^{(1)} + \sum_{r_2=1}^{R_2} \mathbf{v}^{(r_2;2)}\beta_{r_2}^{(2)},
\end{aligned} \tag{49}
$$

where all the right vectors in $\mathbf{V}^{(1)}$ are orthogonal to the vectors in $\mathbf{V}^{(2)}$ (see Eq. (3)), $\mathbf{B}^{(1)} = (\beta_1^{(1)}, \beta_1^{(1)}, ...)^T$ and $\mathbf{B}^{(2)} = (\beta_1^{(2)}, \beta_2^{(2)}, ...)^T$ with $\beta_1^{(1)2} + \beta_2^{(1)2} + ... + \beta_1^{(2)2} + \beta_2^{(2)2} + ... = 1$. Therefore the correlation between the two initial conditions is given by:

$$
c_0 = \mathbf{r}_0^{(1)} \cdot \mathbf{r}_0^{(2)} = \mathbf{A}^{(1)T}\mathbf{B}^{(1)} = \sum_{r_1} \alpha_{r_1}^{(1)}\beta_{r_1}^{(1)}. \tag{50}
$$

Next we compute the correlation between the peak states for the trajectories elicited by the two initial conditions $\mathbf{r}_0^{(1)}$ and $\mathbf{r}_0^{(2)}$. Using Eq. (42) we can approximate the peak states as:

$$
\begin{aligned}
\mathbf{r}_{\text{peak}}^{(1)} &\approx \mathbf{S}^{(1)}\hat{\mathbf{U}}^{(1)}\mathbf{A}^{(1)} \\
\mathbf{r}_{\text{peak}}^{(2)} &\approx \mathbf{S}^{(2)}\hat{\mathbf{U}}^{(2)}\mathbf{B}^{(2)} + \mathbf{S}^{(1)}\hat{\mathbf{U}}^{(1)}\mathbf{B}^{(1)}.
\end{aligned} \tag{51}
$$

Assuming for simplicity that $\mathbf{S}_i^{(1)} = \Delta_1$ and $\mathbf{S}_j^{(2)} = \Delta_2$ for all $i = 1, ..., R_1$ and $j = 1, ..., R_2$ we can write the correlation between the peak states as:

$$
c_{\text{peak}} = \frac{\mathbf{r}_{\text{peak}}^{(1)} \cdot \mathbf{r}_{\text{peak}}^{(2)}}{||\mathbf{r}_{\text{peak}}^{(1)}||\,||\mathbf{r}_{\text{peak}}^{(2)}||} \approx \frac{\sum_{r_1} \alpha_{r_1}^{(1)}\beta_{r_1}^{(1)}}{\sqrt{\sum_{r_1} \beta_{r_1}^{(1)2} + \frac{\Delta_2^2}{\Delta_1^2}\sum_{r_2} \beta_{r_2}^{(2)2}}} = \frac{c_0}{\sqrt{\sum_{r_1} \beta_{r_1}^{(1)2} + \frac{\Delta_2^2}{\Delta_1^2}\sum_{r_2} \beta_{r_2}^{(2)2}}}, \tag{52}
$$

where the correlation between the initial states appears directly in the numerator. Notably, in the case where $\Delta_1 \approx \Delta_2$, equation (52) reduces to

$$
c_0 = c_{\text{peak}}. \tag{53}
$$

2.5 Correlation between initial and peak state

In this section we derive the expression for the correlation between the initial state $\mathbf{r}_0$ and the state at the peak of the transient dynamics, in the case where left and right connectivity vectors are mutually orthogonal (see Eq. (41)). We set the initial condition to $\mathbf{r}_0 = \sum_r \alpha_r \mathbf{v}^{(r)} + \sum_r \beta_r \mathbf{u}^{(r)} + \gamma \mathbf{z}^{(\perp)}$ and evaluate the state at the peak at time $t^* = 1$ (corresponding to the peak time when $\mathbf{r}_0$ has a strong component on the $\mathbf{v}^{(r)}$'s and the dynamics is strongly amplified, i.e. $||\mathbf{u}^{(r)}|| \gg 2$, see Bondanelli and Ostojic (2020)). The initial condition and the peak state are therefore given by:

$$
\begin{aligned}
\mathbf{r}(0) &= \sum_r \alpha_r \mathbf{v}^{(r)} + \sum_r \beta_r \hat{\mathbf{u}}^{(r)} + \gamma \mathbf{z}^{(\perp)} \\
\mathbf{r}(1) &= \frac{1}{e}\Big[\sum_r \alpha_r\, \mathbf{v}^{(r)} + \sum_r (\alpha_r ||\mathbf{u}^{(r)}|| + \beta_r)\, \hat{\mathbf{u}}^{(r)} + \gamma \mathbf{z}^{(\perp)}\Big]
\end{aligned} \tag{54}
$$

From the mutual orthogonality between $\hat{\mathbf{u}}^{(r)}$, $\mathbf{v}^{(r)}$ and $\mathbf{z}^{(\perp)}$ it follows that

$$
\begin{aligned}
||\mathbf{r}(0)||^2 &= 1 \\
||\mathbf{r}(1)||^2 &= \frac{1}{e^2}\Big[\sum_r \alpha_r^2 + \sum_r (\alpha_r ||\mathbf{u}^{(r)}|| + \beta_r)^2 + \gamma^2\Big]
\end{aligned} \tag{55}
$$

Therefore we can write the correlation between the initial condition and the peak state as

$$
\frac{\mathbf{r}(0) \cdot \mathbf{r}(1)}{||\mathbf{r}(0)||\,||\mathbf{r}(1)||} = \frac{1 + \sum_r \alpha_r \beta_r ||\mathbf{u}^{(r)}||}{\sqrt{1 + \sum_r \alpha_r^2 ||\mathbf{u}^{(r)}||^2 + 2\sum_r \alpha_r \beta_r ||\mathbf{u}^{(r)}||}} \tag{56}
$$

In the case of mutually orthogonal left and right connectivity vector (Eq. (41)), if the initial state $\mathbf{r}_0$ has a strong component on the right connectivity vectors $\mathbf{v}^{(r)}$, the component of $\mathbf{r}_0$ on the $\mathbf{u}^{(r)}$, i.e. β_r, can be neglected. Under these assumptions, the sufficient condition Eq. (47) implies the correlation between the initial and peak state must satisfy the upper bound given by

$$\frac{\mathbf{r}(0)\cdot\mathbf{r}(1)}{||\mathbf{r}(0)||\,||\mathbf{r}(1)||} < \frac{1}{e} \approx 0.35 \tag{57}$$

2.6 Principal component analysis of rank-1 dynamics

Here we show the link between the connectivity vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ of a unit-rank connectivity matrix $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$ and the principal components of the OFF response with initial condition $\mathbf{r}_0$. We examine the case in which the left and right connectivity vectors are approximately orthogonal to each other, and we make the additional assumption that the initial state can be written as $\mathbf{r}_0 = \alpha\mathbf{v}^{(1)} + \beta\mathbf{z}^{(\perp)}$, where $\mathbf{z}^{(\perp)}$ is orthogonal to both connectivity vectors. In this case we can write the OFF response Eq. (33) as:

$$\mathbf{r}(t) = f_1(t)\mathbf{r}_0 + f_2(t)\mathbf{u}^{(1)}, \tag{58}$$

where $f_1(t) = e^{-t}$ and $f_2(t) = te^{-t}$. To compute the principal components of the population OFF response for a single stimulus we need to compute the eigenvalues and eigenvectors of the covariance matrix defined as:

$$\mathbf{C} = \int_0^{\tau_0} \big(\mathbf{r}(t) - \boldsymbol{\mu}\big)\big(\mathbf{r}(t) - \boldsymbol{\mu}\big)^T \,\mathrm{d}t, \qquad \boldsymbol{\mu} = \frac{1}{\tau_0}\int_0^{\tau_0} \mathbf{r}(t)\,\mathrm{d}t \tag{59}$$

where $\boldsymbol{\mu}$ is the temporal mean and τ_0 is the timescale on which the transient dynamics occurs. This cutoff is necessary to capture the principal components of the transient dynamics and avoid overweighting the dynamics near the fixed point. Using Eq. (58) we can re-write the covariance matrix as:

$$\mathbf{C} = \alpha\,\mathbf{r}_0\mathbf{r}_0^T + \beta\,(\mathbf{u}^{(1)}\mathbf{r}_0^T + \mathbf{r}_0\mathbf{u}^{(1)T}) + \gamma\,\mathbf{u}^{(1)}\mathbf{u}^{(1)T} \tag{60}$$

where $\alpha = \int f_1^2 - \tau_0^{-1}\,(\int f_1)^2$, $\beta = \int f_1 f_2 - \tau_0^{-1}\int f_1 \int f_2$ and $\gamma = \int f_2^2 - \tau_0^{-1}\,(\int f_2)^2$. The coefficient α, β and γ are of order 1. We examine the two limit cases of weakly amplified dynamics, $||\mathbf{u}^{(1)}|| \ll 2$, and strongly amplified dynamics $||\mathbf{u}^{(1)}|| \gg 2$. In the case of weakly amplified dynamics we can write the covariance matrix as:

$$\mathbf{C} = \alpha\,\mathbf{r}_0\mathbf{r}_0^T + O(||\mathbf{u}^{(1)}||/\sqrt{N}) \tag{61}$$

Therefore the first principal component of weakly amplified trajectory is strongly aligned with the vector $\mathbf{r}_0$. It follows that the second principal component is strongly correlated with the vector $\mathbf{u}^{(1)}$. In case of strongly amplified dynamics the covariance matrix can be written as:

$$\mathbf{C} = \gamma\,||\mathbf{u}^{(1)}||^2\hat{\mathbf{u}}^{(1)}\hat{\mathbf{u}}^{(1)T} + O(||\mathbf{u}^{(1)}||/\sqrt{N}) \tag{62}$$

where we defined the normalized vector $\hat{\mathbf{u}}^{(1)} = \mathbf{u}^{(1)}/||\mathbf{u}^{(1)}||$. The first principal component is therefore strongly correlated with $\hat{\mathbf{u}}^{(1)}$, while the second one with $\mathbf{r}_0$.

2.7 Amplification in a rank-2 anti-symmetric channel

When fitting a low-rank linear dynamical system to the OFF responses to individual stimuli we observed a specific structure in the pattern of correlations between right and left connectivity vectors $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(j)}$ of the fitted connectivity $\mathbf{J}$. This pattern showed low values of the correlation for almost all pairs of connectivity vectors, except for pairs of left and right vectors across adjacent modes, i.e. $\mathbf{u}^{(1)}$-$\mathbf{v}^{(2)}$, $\mathbf{u}^{(2)}$-$\mathbf{v}^{(1)}$, $\mathbf{u}^{(3)}$-$\mathbf{v}^{(4)}$, $\mathbf{u}^{(4)}$-$\mathbf{v}^{(3)}$, and so on. This structures gives rise to independent rank-2 channels grouping pairs of modes, of the form $\mathbf{J}_1 = \mathbf{u}^{(1)}\mathbf{v}^{(1)T} + \mathbf{u}^{(2)}\mathbf{v}^{(2)T}$. Moreover, within a single channel, the values of the correlation was high, and opposite in sign, for different pairs of vectors, i.e. $\hat{\mathbf{u}}^{(1)}\cdot\mathbf{v}^{(2)} \approx -\hat{\mathbf{u}}^{(2)}\cdot\mathbf{v}^{(1)} \approx 1$. As a results these channels had a strong anti-symmetric component that can be written as

$$\mathbf{J}_{1,\mathrm{antisymm}} = \Delta\mathbf{v}^{(2)}\mathbf{v}^{(1)T} - \Delta\mathbf{v}^{(1)}\mathbf{v}^{(2)T} \tag{63}$$

for one example channel, where we set

$$\begin{aligned}\mathbf{u}^{(1)} &= \Delta \mathbf{v}^{(2)} \\ \mathbf{u}^{(2)} &= -\Delta \mathbf{v}^{(1)}\end{aligned} \tag{64}$$

with $\Delta = ||\mathbf{u}^{(1)}|| = \mathbf{u}^{(2)}$. In the following we show that purely anti-symmetric rank-2 channels of the form given by Eq. (63) cannot alone produce amplified dynamics, thus requiring the strong symmetric component that was observed in the full connectivity spectrum (Fig. 4D) to exhibit amplification. To this end, we examine the norm of the population activity vector $||\mathbf{r}(t)||$ (with $\mathbf{r}(t)$ given by Eq. (35)) in response to an initial condition $\mathbf{r}_0$ with a connectivity matrix given by Eq. (63). We set the initial state to $\mathbf{r}_0 = \alpha_1 \mathbf{v}^{(1)} + \alpha_2 \mathbf{v}^{(2)}$. From Eq. (63) we can write

$$\mathbf{V}^T\mathbf{U} = \begin{pmatrix} 0 & -\Delta \\ +\Delta & 0 \end{pmatrix} \tag{65}$$

and therefore the term $\mathbf{a}(t) = \left(\mathbf{V}^T\mathbf{U}\right)^{-1}\left[\exp(t\mathbf{V}^T\mathbf{U}) - \mathbf{I}\right]\left(\mathbf{V}^T\mathbf{r}_0\right)$ in Eq. (35) can be written as:

$$\mathbf{a}(t) = \begin{pmatrix} 0 & +1/\Delta \\ -1/\Delta & 0 \end{pmatrix} \begin{pmatrix} \cos\Delta t - 1 & \sin\Delta t \\ -\sin\Delta t & \cos\Delta t - 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{66}$$

From Eq. (66), Eq. (35) and Eq. (64) we derive the expression for the dynamics and its norm as:

$$\begin{aligned}&\mathbf{r}(t) = e^{-t}(\alpha_1 \cos\Delta t + \alpha_2 \sin\Delta t)\,\mathbf{v}^{(1)} + e^{-t}(\alpha_2 \cos\Delta t - \alpha_1 \sin\Delta t)\,\mathbf{v}^{(2)} \\ &||\mathbf{r}(t)||^2 = e^{-2t}.\end{aligned} \tag{67}$$

We note that the temporal derivative of the norm takes on negative values at all times, and independently of the value of the initial state $\mathbf{r}_0$:

$$\frac{\mathrm{d}||\mathbf{r}(t)||}{\mathrm{d}t} = -e^{-t} < 0 \quad \forall t \geq 0. \tag{68}$$

We conclude that a rank-2 anti-symmetric channel of the form given by Eq. (63) cannot produce amplified dynamics.

3 Single-cell model for OFF response generation

We consider the single-cell model given by Eq. (5), where $r_{0,i}^{(s)}$ denote the initial state $r_i^{(s)}(0)$ of the response of unit i to stimulus s. Without loss of generality we assume that $L_i(0) = 1$. In fact if the responses were given by $r_i^{(s)}(t) = r_{0,i}^{(s)} K_i(t)$ such that $K_i(t) \neq 0$ we could always write the responses as $r_i^{(s)}(t) = \tilde{r}_{0,i}^{(s)} \tilde{L}_i(t)$, where $\tilde{r}_{0,i}^{(s)} = r_{0,i}^{(s)} K_i(0)$ and $\tilde{L}_i(t) = K_i(t)/K_i(0)$, formally equal to Eq. (5).

3.1 Fitting single-cell responses with basis functions

To fit the single-cell OFF responses $r_i^{(s)}(t)$, we use an approach based on basis functions (Pillow et al., 2008). The problem consists in finding the coefficients $a_{ij}^{(s)}$ that best approximate the equation

$$r_i^{(s)}(t) = \sum_{j=1}^{N_{\text{basis}}} a_{ij}^{(s)} f_j(t), \tag{69}$$

where the shape and the number of basis function N_{basis} is predetermined. We choose the functions $f_i(t)$ to be Gaussian functions centered around a certain value $\bar{t}_i$ and with a certain width w_i, i.e. $f_i(t) = \exp(-(t - \bar{t}_i)^2/2w^2)$. By dividing the left and right hand side of Eq. (69) by the initial condition $r_{0,i}^{(s)}$ we obtain:

$$\frac{r_i^{(s)}(t)}{r_{0,i}^{(s)}} = \sum_{j=1}^{N_{\text{basis}}} b_{ij}^{(s)} f_j(t), \quad b_{ij}^{(s)} = a_{ij}^{(s)}/r_{0,i}^{(s)}. \tag{70}$$

In general the coefficients $b_{ij}^{(s)}$ could be fitted independently for each stimulus. However, the single-cell model assumes that the coefficients b_{ij} do not change across stimuli (see Eq. (5)). Therefore, to find the

stimulus-independent coefficients b_{ij} that best approximate Eq. (70) across a given set of stimuli, we minimize the mean squared error given by:

$$MSE = \sum_{i,\{\text{stimuli}\}} \int \mathrm{d}t \left(\frac{r_i^{(s)}(t)}{r_{0,i}^{(s)}} - \sum_{j=1}^{N_{\text{basis}}} b_{ij} f_j(t) \right)^2 \tag{71}$$

The minimization of the mean squared error can be solved using linear regression techniques. Suppose we want to fit the population responses to C different stimuli simultaneously. Let $\mathbf{R}^{(C)}$ be the matrix of size $N \times TC$ obtained by concatenating the $N \times T$ matrices $(\mathbf{r}_{\text{norm}}^{(s)})_{it} = r_i^{(s)}(t)/r_{0,i}^{(s)}$ ($i = 1, ..., N$, $t = 1, ..., T$, $s = 1, ..., C$) corresponding to the normalized responses to the C stimuli. Let $\mathbf{F}^{(C)}$ be the $N_{\text{basis}} \times TC$ matrix obtained by concatenating C times the $N_{\text{basis}} \times T$ matrix $(\mathbf{f})_{it} = f_i(t)$. Let $\mathbf{B}$ be the $N \times N_{\text{basis}}$ matrix given by $\mathbf{B}_{ij} = b_{ij}$. Then Eq. (70) can be written as:

$$\mathbf{R}^{(C)} = \mathbf{B}\mathbf{F}^{(C)}, \tag{72}$$

which can be solved using linear regression.

If population responses are generated through Eq. (5), then Eq. (69) implies that the performance of the fit Eq. (72) is independent of the number C of stimuli considered. In the case of OFF responses generated through a recurrent network mechanisms, Eq. (69) does not hold true, and the performance of the fit decreases as the number of stimuli C is increased. For convenience in Fig. 5D-F instead of normalizing the responses by their initial condition, we normalized them by the firing rate range, i.e. we choose $r_{0,i}^{(s)} = |\max_t r_i^{(s)}(t) - \min_t r_i^{(S)}(t))| = \alpha_i r_{0,i}$ (this normalization is equivalent to fitting the filters $\tilde{L}_i(t) \equiv L_i(t)/\alpha_i$ instead of $L_i(t)$). In order to avoid normalizing for very small values, we fit only the most responding neurons in the population, as quantified by their firing rate range.

3.2 Structure of single-trial responses in the recurrent model

In this section we examine the structure of single-trial population OFF responses generated by a low-rank recurrent model with rank R. Specifically, we compute the covariance matrix of the neural population at fixed stimulus s and fixed time t. For simplicity we assume that the variability in the single-trial responses is fully due to the variability in the initial state at stimulus offset, $\mathbf{r}_0^{(s)}$, but these result can be extended to the case where external input noise is fed into the network, without affecting the key results. We also assume that Eq. (41) is satisfied, so that we can write the dynamics of the network using Eq. (42). We consider single-trial OFF responses to stimulus s, denoted by $\mathbf{r}^{(s;\nu)}(t)$, where the response corresponding to the ν-th trial has initial state given by $\mathbf{r}_0^{(s;\nu)} = \mathbf{r}_0^{(s)} + \boldsymbol{\eta}^{(\nu)}$, so that the OFF response can be written as

$$\mathbf{r}^{(s;\nu)}(t) = \left[e^{-t} + te^{-t}\,\mathbf{U}\mathbf{V}^T\right](\mathbf{r}_0^{(s)} + \boldsymbol{\eta}^{(\nu)}), \tag{73}$$

where the variability in the initial condition $\boldsymbol{\eta}^{(\nu)}$ is independent for each unit, and has zero mean and variance equal to σ^2 ($\langle \eta_i^{(\nu)} \rangle = 0$ and $\langle \eta_i^{(\nu)} \eta_j^{(\nu)} \rangle = \sigma^2 \delta_{ij}$, with $\langle \cdot \rangle$ denoting the average over trials). From Eq. (73) we can write the covariance matrix of the single-trial population activity at time t as:

$$\mathbf{C}(t) = \left\langle \left(\mathbf{r}^{(s;\nu)}(t) - \langle \mathbf{r}^{(s;\nu)}(t) \rangle\right) \left(\mathbf{r}^{(s;\nu)}(t) - \langle \mathbf{r}^{(s;\nu)}(t) \rangle\right)^T \right\rangle, \tag{74}$$

where $\langle \mathbf{r}^{(s;\nu)}(t) \rangle$ denotes the trial-averaged response. Using Eq. (73) and Eq. (42) we can write

$$\begin{aligned} \mathbf{C}(t) &= \left\langle \left(e^{-t} + te^{-t}\,\mathbf{U}\mathbf{V}^T\right) \boldsymbol{\eta}^{(\nu)} \boldsymbol{\eta}^{(\nu)T} \left(e^{-t} + te^{-t}\,\mathbf{U}\mathbf{V}^T\right)^T \right\rangle \\ &= \sigma^2 \left(e^{-t} + te^{-t}\,\mathbf{U}\mathbf{V}^T\right) \left(e^{-t} + te^{-t}\,\mathbf{U}\mathbf{V}^T\right)^T, \end{aligned} \tag{75}$$

where in the second equality we used the independence of the variability across neurons. Therefore from Eq. (75) we directly obtain the expression for $\mathbf{C}(t)$ as:

$$\mathbf{C}(t) = \sigma^2 e^{-2t} \Big[\mathbf{I} + t\big(\mathbf{V}\mathbf{U}^T + \mathbf{U}\mathbf{V}^T\big) + t^2 \mathbf{U}\mathbf{U}^T \Big]. \tag{76}$$

In the following we examine the covariance matrix at two particular times, i.e. at time $t = 0$ and at the peak time. We study the case where the initial state is strongly aligned with the vectors $\mathbf{v}^{(r)}$ (see Eq. (47)) and the dynamics is strongly amplified, $||\mathbf{u}^{(r)}|| \gg 2$, so that the peak time can be approximated by $t \approx 1$. In

particular, we will consider the variability along the amplified direction $\mathbf{r}_{\text{ampl}} \equiv \mathbf{r}(t \approx 1)$ and along random directions $\mathbf{r}_{\text{rand}}$ in the state-space, where the variability along a given direction $\mathbf{z}_0$ with $||\mathbf{z}_0|| = 1$ at any time t is given by

$$\text{var}(\mathbf{z}_0; t) = \mathbf{z}_0^T \mathbf{C}(t) \mathbf{z}_0. \tag{77}$$

We approximate the amplified direction by $\mathbf{r}_{\text{ampl}} \propto \mathbf{U}\mathbf{\Lambda}$, where $\mathbf{\Lambda} = (\alpha_1, ..., \alpha_r)^T$ represents the components of the initial condition on $\mathbf{V}$, i.e. $\mathbf{\Lambda} = \mathbf{V}^T \mathbf{r}_0^{(s)}$. Using Eq. (76) we can write:

$$\begin{aligned} \mathbf{C}(0) &= \sigma^2 \mathbf{I} \\ \mathbf{C}(1) &= \frac{\sigma^2}{e^2} \Big[\mathbf{I} + \big(\mathbf{V}\mathbf{U}^T + \mathbf{U}\mathbf{V}^T \big) + \mathbf{U}\mathbf{U}^T \Big] \end{aligned} \tag{78}$$

From Eq. (78) we obtain the expressions for the variability along the amplified and random directions, at the initial and peak times:

$$\begin{aligned} &\text{var}(\mathbf{r}_{\text{rand}}; 0) = \sigma^2, \quad \text{var}(\mathbf{r}_{\text{rand}}; 1) = \sigma^2 \\ &\text{var}(\mathbf{r}_{\text{ampl}}; 0) = \sigma^2/e^2, \quad \text{var}(\mathbf{r}_{\text{ampl}}; 1) = \frac{\sigma^2}{e^2} \sum_{r=1}^{R} \alpha_r^2 ||\mathbf{u}^{(r)}||^2 + O(\{||\mathbf{u}^{(r)}||\}) \end{aligned} \tag{79}$$

Equation (79) indicates that in a rank-R recurrent network, while the variability along a random direction does not change as a function of time, the variability along the amplified direction increases from the initial time to the peak time. In Fig. 6 we quantified this observation by computing the variability amplification (VA), defined as the ratio between the variability at peak time and the variability at the initial time

$$\text{VA}(\mathbf{z}_0) = \frac{\text{var}(\mathbf{z}_0; 1)}{\text{var}(\mathbf{z}_0; 0)} \tag{80}$$

so that we have:

$$\begin{aligned} &\text{VA}(\mathbf{r}_{\text{rand}}) = 1 \\ &\text{VA}(\mathbf{r}_{\text{ampl}}) \approx \sum_{r=1}^{R} \alpha_r^2 ||\mathbf{u}^{(r)}||^2 \gg 1. \end{aligned} \tag{81}$$

Shuffled responses In the following we compare the variability amplification of the original responses with the variability amplification obtained from the responses where the trial labels have been shuffled independently for each cell. The covariance matrix $\mathbf{C}^{(\text{shuffled})}(t)$ of the shuffled responses is obtained by retaining only the diagonal elements of the real covariance matrix:

$$\mathbf{C}_{ij}^{(\text{shuffled})}(t) = \mathbf{C}_{ij}(t) \delta_{ij}. \tag{82}$$

It is straightforward to verify that the first three terms of Eqs. (79) computed using $\mathbf{C}^{(\text{shuffled})}$ instead of $\mathbf{C}$ do not change. This in turns implies that $\text{VA}(\mathbf{r}_{\text{ampl}}; \text{shuffled}) = \text{VA}(\mathbf{r}_{\text{ampl}}) = 1$. Assuming that the elements of the right connectivity vectors $\mathbf{u}^{(r)}$ are randomly and independently distributed with mean zero and variance equal to $||\mathbf{u}^{(r)}||^2/N$, we compute the expected value of the term $\text{var}(\mathbf{r}_{\text{ampl}}; 1; \text{shuffled})$ over all possible realization of $\mathbf{U}$, and obtain:

$$\begin{aligned} [\text{var}(\mathbf{r}_{\text{ampl}}; 1; \text{shuffled})]_{\mathbf{U}} &= \sigma^2 + \sigma^2 \left[\sum_{i,n=1}^{R} \sum_{j,m=1}^{N} \alpha_i \, \mathbf{U}_{ij}^T \, (\mathbf{U}\mathbf{U}^T)_{jm} \delta_{jm} \, \mathbf{U}_{mn} \, \alpha_n \right]_{\mathbf{U}} \\ &= \sigma^2 + \sigma^2 \left[\sum_{i,n=1}^{R} \sum_{j=1}^{N} \alpha_i \, \mathbf{u}_j^{(i)} \left(\sum_{r=1}^{R} \mathbf{u}_j^{(r)2} \right) \mathbf{u}_j^{(n)} \, \alpha_n \right]_{\mathbf{U}} \\ &= \sigma^2 + \sigma^2 \sum_{i=1}^{R} \sum_{j=1}^{N} \alpha_i^2 \frac{||\mathbf{u}^{(i)}||^2}{N} \sum_{r=1}^{R} \frac{||\mathbf{u}^{(r)}||^2}{N} \\ &= \sigma^2 + \sigma^2 \sum_{i=1}^{R} \alpha_i^2 ||\mathbf{u}^{(i)}||^2 \left(\sum_{r=1}^{R} \frac{||\mathbf{u}^{(r)}||^2}{N} \right), \end{aligned} \tag{83}$$

where in the third equality the only non-zero contribution comes from the terms with $i = n$. From Eq. (83) we note that, for fixed rank R, when $N \to \infty$ we have $[\mathrm{var}(\mathbf{r}_{\mathrm{ampl}}; 1; \mathrm{shuffled})]_{\mathbf{U}} = \sigma^2$, so that $\mathrm{VA}(\mathbf{r}_{\mathrm{ampl}}; \mathrm{shuffled}) = 1$. More generally, if $\sum_{r=1}^{R} ||\mathbf{u}^{(r)}||^2/N < 1$, we have $[\mathrm{var}(\mathbf{r}_{\mathrm{ampl}}; 1; \mathrm{shuffled})]_{\mathbf{U}} < \mathrm{var}(\mathbf{r}_{\mathrm{ampl}}; 1)$, so that $\mathrm{VA}(\mathbf{r}_{\mathrm{ampl}}; \mathrm{shuffled}) < \mathrm{VA}(\mathbf{r}_{\mathrm{ampl}})$, implying a weaker modulation of the variability amplification between random and amplified direction in the case of shuffled responses with respect to the original responses. This in turn shows that the difference in the variability amplification between the amplified and random directions is a result of the recurrent interactions within the network, reflected in the off-diagonal elements of $\mathbf{C}(t)$, and it cannot be explained solely by the structure of the variability of single units.

3.3 Structure of single-trial responses in the single-cell model

Here we examine the structure of the single-trial population responses generated by the single-cell model (Eq. (5)). We consider single-trial responses $\mathbf{r}^{(s;\nu)}(t)$ generated by perturbing the initial state of the trial averaged responses $\mathbf{r}_0^{(s)}$, yielding

$$\begin{aligned} r_i^{(s;\nu)}(t) &= r_{0,i}^{(s;\nu)} L_i(t) \\ &= (r_{0,i}^{(s)} + \eta_i^{(\nu)}) L_i(t), \end{aligned} \tag{84}$$

where $\eta^{(\nu)}$ has zero mean and variance equal to σ^2 ($\langle \eta_i^{(\nu)} \rangle = 0$ and $\langle \eta_i^{(\nu)} \eta_j^{(\nu)} \rangle = \sigma^2 \delta_{ij}$). From Eq. (84) we can compute the covariance matrix of the population response for the single-cell model as:

$$\begin{aligned} \mathbf{C}_{ij}(t) &= \left\langle \left(\mathbf{r}_i^{(s;\nu)}(t) - \langle \mathbf{r}_i^{(s;\nu)}(t) \rangle\right) \left(\mathbf{r}_j^{(s;\nu)}(t) - \langle \mathbf{r}_j^{(s;\nu)}(t) \rangle\right)^T \right\rangle \\ &= L_i(t) \langle \eta_i^{(\nu)} \eta_j^{(\nu)} \rangle L_j(t) \\ &= \sigma^2 \, L_i^2(t) \, \delta_{ij}. \end{aligned} \tag{85}$$

We can obtain the variability along a given direction $\mathbf{z}_0$ as:

$$\mathrm{var}(\mathbf{z}_0; t) = \mathbf{z}_0^T \mathbf{C}(t) \mathbf{z}_0 = \sigma^2 \sum_{i=1}^{N} \mathbf{z}_{0,i}^2 L_i^2(t). \tag{86}$$

Notably, the term on the right hand side is independent of the initial condition of the dynamics, and is proportional to the norm of the dynamics that would be evoked by the initial condition $\mathbf{z}_0$. In contrast to the recurrent model, where variability is maximally amplified along the amplified direction, in the single-cell model the variability is maximally amplified along the initial condition leading to highest amplification. Nonetheless, the particular relationship between the modulation of the variability along the amplified dimension and a random direction explored in the previous section strictly depends on the specific shape of the temporal filters $L_i(t)$, and no general statement can be made in this regard. We can however contrast the single-trial structure generated by the single-cell model with the structure generated by the recurrect model by noting that the covariance matrix given by Eq. (85) is diagonal, and is therefore equal to the covariance matrix of the shuffled responses, $\mathbf{C}_{ij}(t) = \mathbf{C}_{ij}^{\mathrm{shuffled}}(t)$. This implies that no difference is observed when computing the variability amplification for the amplified and random directions for the real and shuffled responses (i.e. $\mathrm{VA}(\mathbf{r}_{\mathrm{rand}}) = \mathrm{VA}(\mathbf{r}_{\mathrm{rand}}; \mathrm{shuffled})$ and $\mathrm{VA}(\mathbf{r}_{\mathrm{ampl}}) = \mathrm{VA}(\mathbf{r}_{\mathrm{ampl}}; \mathrm{shuffled})$), indicating that the structure of the trial-to-trial variability can be fully explained in terms of the variability of single units.

Software

Numerical simulations and data analyses were done using Python 3.6 (using numpy and scipy, and the scikit-learn package (Pedregosa et al., 2011)). Code is available at https://github.com/gbondanelli/OffResponses. Surrogate datasets for model validation (see Materials and methods and Fig. 4C, Fig. S5, Fig. S6) were generated using the Matlab code provided by Elsayed and Cunningham (2017) available at https://github.com/gamaleldin/TME.

Supplementary figures

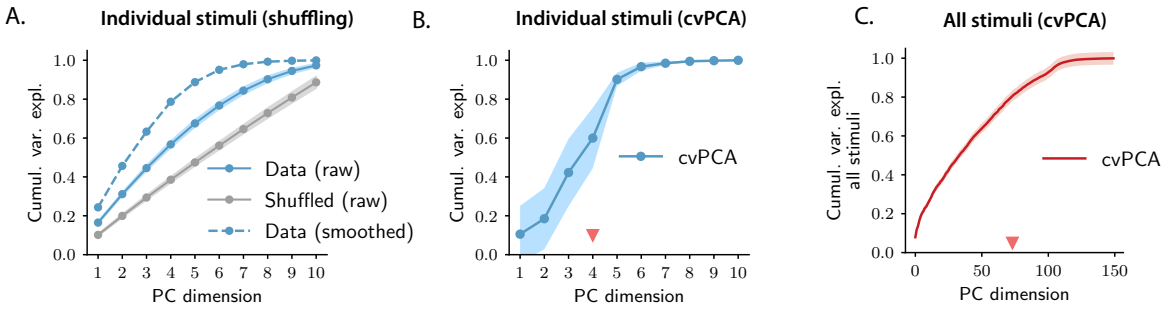

Figure S1: **Controls for PCA of OFF responses to individual stimuli and across stimuli A.** Solid blue trace: cumulative variance explained, evaluated using PCA applied on OFF responses to individual stimuli using raw calcium traces (13 data points). Grey trace: cumulative variance explained after shuffling cell indices independently at each time step, so as to remove temporal correlations between time points. Responses where the temporal correlations have been removed exhibit higher dimensionality with respect to the original responses. For comparison, the cumulative variance explained computed on the temporally smoothed responses (see Materials and methods) is shown (dashed blue trace). **B.** Cumulative variance explained evaluated using the cross-validated PCA (cvPCA) method (Stringer et al., 2019) applied on the OFF responses to individual stimuli. For each stimulus we evaluated the cvPCA spectrum averaging over all pairs of trials extracted from 20 total trials (see Materials and methods). The solid trace and color shading represent the mean and standard deviation of the cumulative variance over stimuli. **C.** Cumulative variance explained evaluated using the cvPCA method applied on the OFF responses to all stimuli at once. Here the solid trace and color shading represent the mean and standard error of the cumulative variance computed over all pairs of trials. In both plots the triangular red markers indicates the cvPCA dimensionalities that explains at least 80% of the response variance.

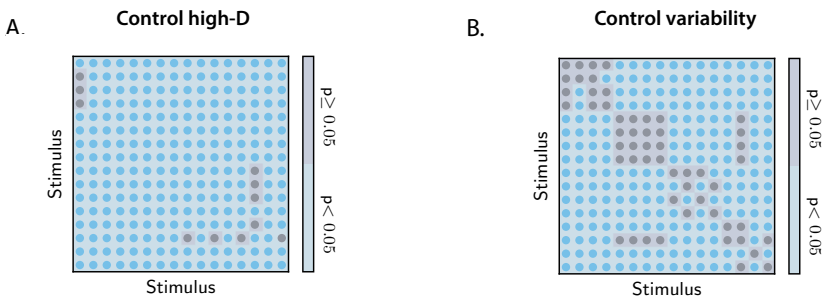

Figure S2: **Controls for the orthogonality between OFF response subspaces A.** Control testing for the hypothesis that small subspace overlaps are due to the high dimensionality of neural responses. Blue and grey indicate stimulus pairs for which the p-value for the difference between the data and the controls is respectively smaller and larger or equal than 0.05 (lower tail test). Blue regions therefore indicate the stimulus pairs for which low values of the subspace overlaps are not attributable to the high dimensionality of the state-space. The number of shuffling is set to 200 (see Materials and methods). **B.** Control testing for the hypothesis that small subspace overlaps are due to the trial-to-trial variability of neural responses. Blue and grey indicate stimulus pairs for which the p-value for the difference between the data and the controls is respectively smaller and larger or equal than 0.05 (two-tailed independent t-test). Blue regions therefore indicate the stimulus pairs for which low values of the subspace overlaps are not attributable to the trial-to-trial variability of neural activity. The number of trial subsamplings permutations is set to 200 (see Materials and methods).

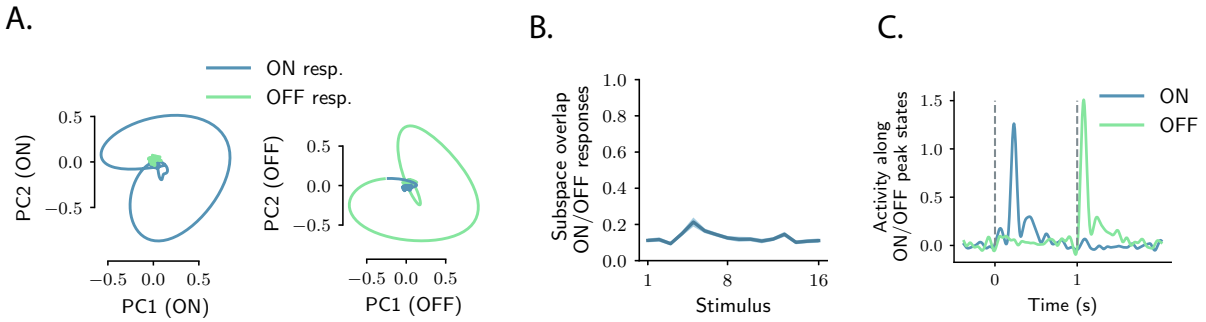

Figure S3: **Relation between ON and OFF responses in the auditory cortex. A.** *Left*: projection of the population ON and OFF responses to a 8 kHz pure tone (1s, UP-ramp, with intensity range 60-85dB) on the first two principal components computed on the ON response. *Right*: projection of both ON and OFF responses to the same stimulus on the first two principal components computed on the OFF response. PCA was performed on the period from -50 ms to 300 ms with respect to stimulus onset and offset. **B.** Subspace overlap between the subspaces spanned by the ON and OFF population responses to individual stimuli. The subspace overlap is computed as the cosine of the principal angle between the ON and OFF subspaces, defined by the first 5 principal components of the ON and OFF population responses respectively. Error bars correspond to ± 1 standard deviations computed over 100 subsamplings of the 80% of the units. **C.** Projection of the population response to the stimulus in **A.** along the population states at the peak of the transient ON and OFF responses (corresponding to maximum distance from baseline).

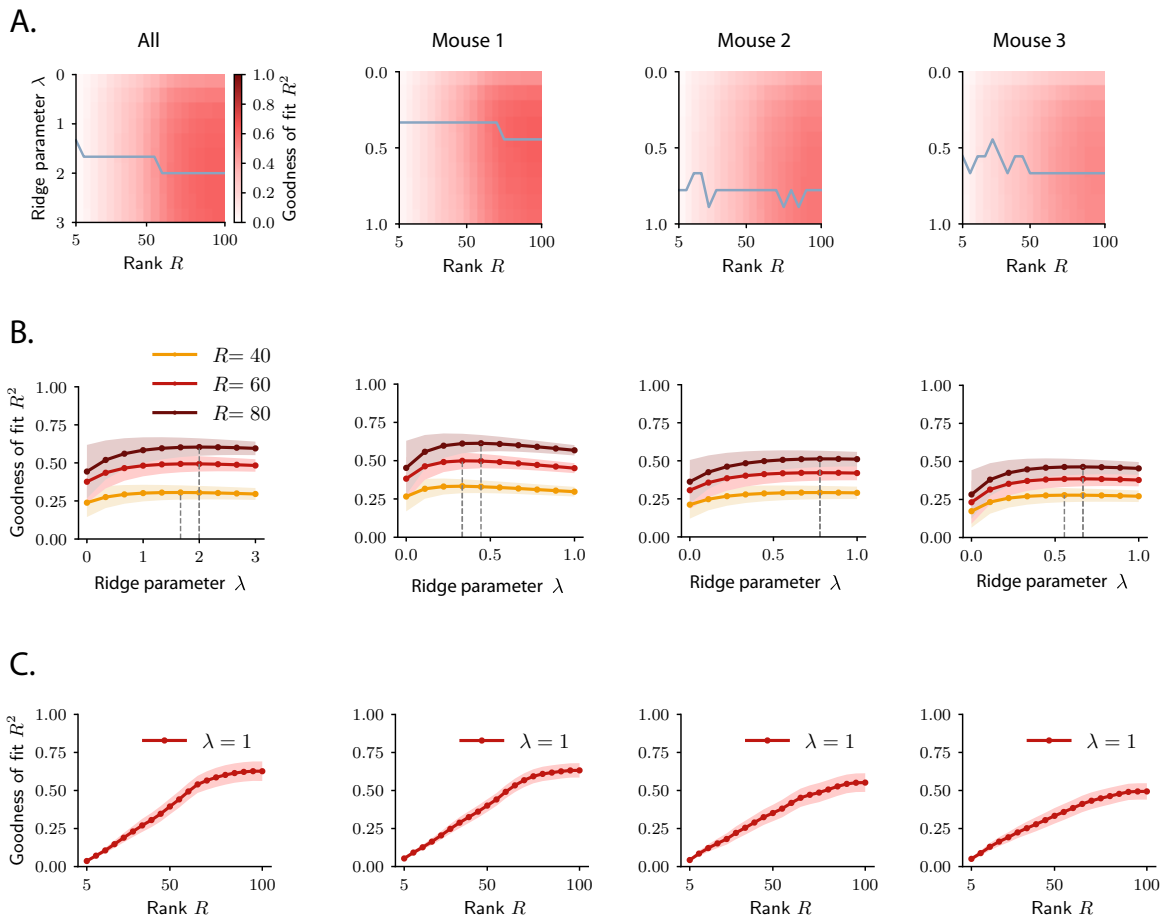

Figure S4: **Selection of hyperparameters in rank-reduced ridge regression.** The ridge and rank hyperparameters λ and R are selected using 10-fold cross-validation in the time domain, when fitting the activity of the whole pseudopopulation (All) or the activity of individual animals. **A.** Goodness of fit R^2 as a function of the hyperparameters λ and R. The gray trace shows the value of λ for which the R^2 is maximum for each value of the rank parameter R. **B.** Goodness of fit R^2 as a function of the hyperparameters λ for three different choices of the rank R. The value of λ for which R^2 is maximum is indicated by the dashed line. **C.** Goodness of fit R^2 as a function of the rank R for $\lambda = 1$. The value of the R^2 as a function of the rank R does not exhibit a clear maximum, but rather tends to saturate at a certain value R^*.

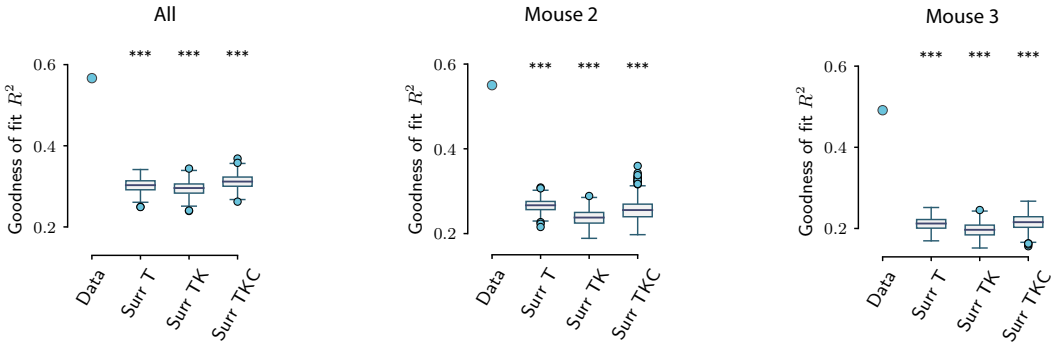

Figure S5: **Goodness of fit for the original and surrogate datasets.** Value of R^2 for the original and 500 surrogate datasets computed using the activity of the whole pseudopopulation (All) and the activity of individual animals (mouse 1 is shown in fig4D). The parameters for the reduced-rank ridge fitting for the three plots are respectively: $\lambda_{\text{All}} = 2$, $R_{\text{All}} = 70$; $\lambda_2 = 0.8$, $R_2 = 90$; $\lambda_3 = 0.6$, $R = 90$.

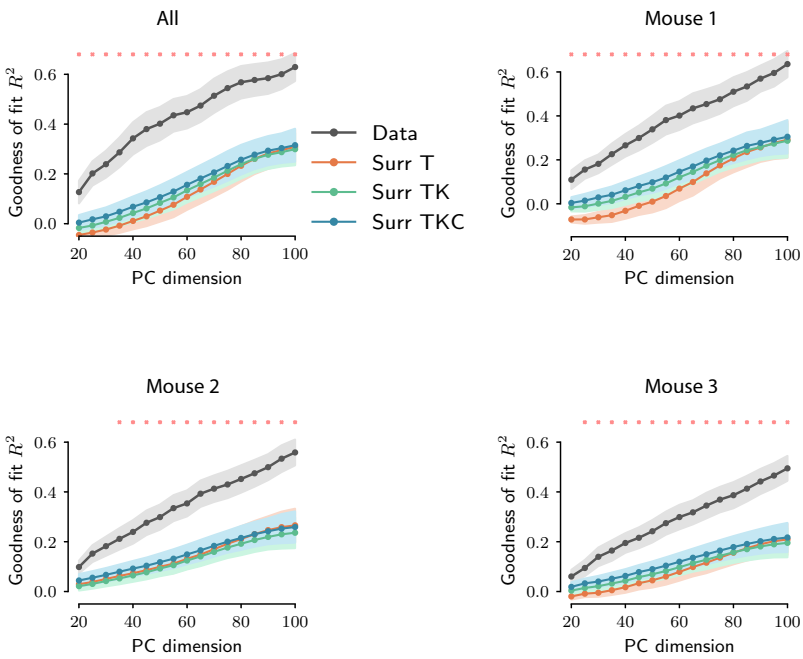

Figure S6: **Goodness of fit as a function of PC dimensionality for original and surrogate datasets.** Value of R^2 as a function of PC dimensionality for the original dataset (black trace) and for the surrogate datasets (colored traces). The value of R^2 was computed using ridge regression using 10-fold cross-validation. The ridge parameter is optimized for each choice of the PC dimensionality. For the original dataset, error bars represent the standard error over the 10 cross-validation folds; for the surrogate datasets, error bars represent the average cross-validation standard error over all the surrogates. Here the number of surrogates is set to 100. For each choice of the PC dimensionality, a red cross indicates that the difference in R^2 between the original dataset and the most constraining surrogate (TKC) is statistically significant ($P < 0.01$; upper tail test using the mean value of R^2 over the cross-validations folds for each dataset).

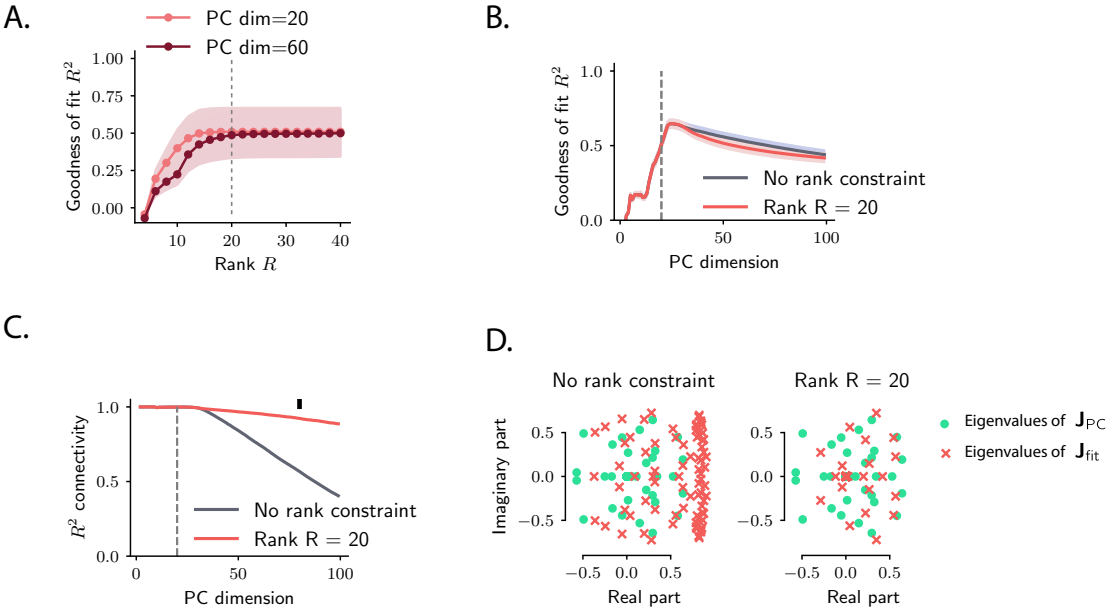

Figure S7: **Model recovery for simulated OFF responses using a low-rank network model.** Model recovery for OFF responses generated through a low-rank network mechanism. The number of units was set to $N = 500$ and the rank of the connectivity was fixed to 30. We generated 10 OFF responses by setting the initial state of the network to 10 orthogonal left connectivity vectors $\mathbf{v}^{(r)}$. Input noise was injected into the system. Thus, the dynamics across stimuli spanned at least 20 dimensions. **A.** Goodness of fit R^2 as a function of the rank parameter R in the reduced-rank regression fit, for two values of the PC dimensionality. 10-fold cross validation is used; error bars represent the standard error across the 10 cross-validation folds. The value of R^2 increases as a function of the rank R and stays approximately constant for rank bigger or equal to 20 (dashed line). **B.** Goodness of fit R^2 as a function of the PC dimensionality computed using ordinary least square regression (with no rank constraint, grey trace) and reduced-rank regression (with $R = 20$, red trace). The dashed line indicated the number of dimensions spanned by the dynamics for all stimuli in case of zero input noise (equal to 20 in this case). For a value of the PC dimensionality larger than the number of dimensions spanned by the dynamics, the value of R^2 decreases due to overfitting. Cross-validation is performed as in **A.** **C.** Variance explained, as quantified by the R^2, between the elements of connectivity resulting from the fitting procedure $\mathbf{J}_{\mathrm{fit}}$ and the projection of the real connectivity $\mathbf{J}$ on the top principal components, $\mathbf{J}_{\mathrm{PC}}$, as a function of the number of principal components considered (see Materials and methods). The gray trace corresponds to ordinary least square regression, while the red trace to reduced-rank regression with rank $R = 20$. **D.** Spectra of the matrices $\mathbf{J}_{\mathrm{fit}}$ (green dots) and $\mathbf{J}_{\mathrm{PC}}$ (red crosses) using ordinary least square regression (left) and reduced-rank regression (right) for fixed dimensionality (equal to 80, black marker in panel **C.**). In all panels, the ridge parameter is set to $\lambda = 0.1$.

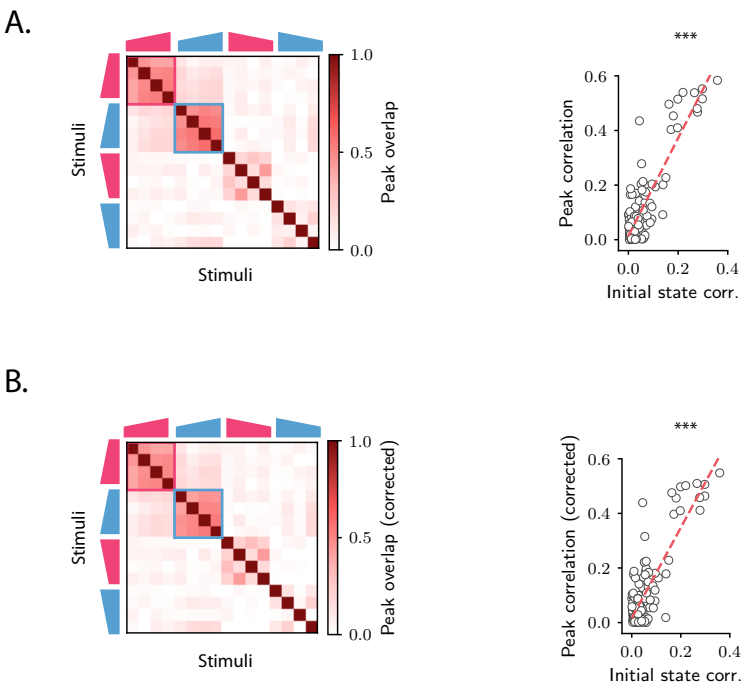

Figure S8: **Overlap between the state at the peak of the transient OFF responses. A.** *Left*: overlap between the states at the peak of the OFF responses for each pair of stimuli. The peak time is defined as the time at which the distance from baseline of the population vector is maximum. *Right*: linear correlation between the overlap between peak states and the overlap between initial conditions for each stimulus pair. **B.** Same as **A.** except that, for each stimulus, the component of the peak state along the corresponding initial condition has been subtracted.

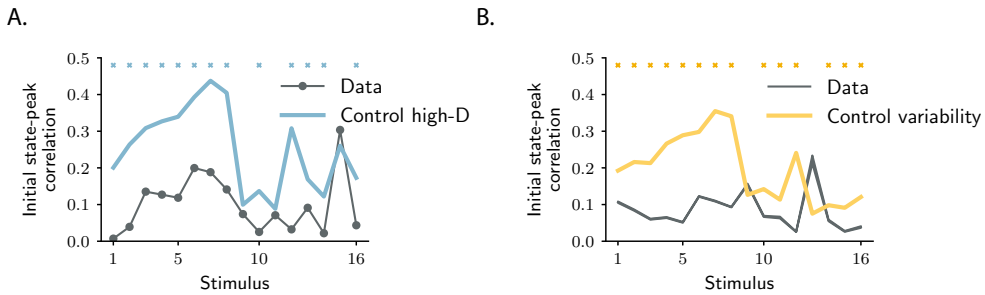

Figure S9: **Controls for the orthogonality between initial and peak state** Control testing for the hypothesis that small correlations are due to the high dimensionality of neural responses. The dark trace represents the absolute value of the correlation between initial and peak state for individual stimuli. The blue trace represents the correlations between initial and peak state after shuffling "initial state" and "peak state" labels across trials (mean and standard error across shuffling is shown). Blue markers indicates stimulus indices for which the p-value for the difference between the data and the controls is smaller than 0.05 (lower tail test), where a low value of the correlation is not attributable to the high dimensionality of the state-space (see Materials and methods). The number of shuffles is set to 1000. **B.** Control testing for the hypothesis that small correlations are due to the trial-to-trial variability of neural responses. The dark trace represents absolute value of the correlation between initial and peak state for individual stimuli (mean and standard error over 1000 subsampling of 10 trials are shown). The yellow trace represents the correlations between responses at the same time points but averaged over two different sets of trials (mean and standard error over 1000 permutations of trials are shown). Yellow markers indicates stimulus indices for which the p-value for the difference between the data and the controls is smaller than 0.05 (two-tailed independent t-test), where a low value of the correlation is not attributable to the trial-to-trial variability of neural responses (see Materials and methods).

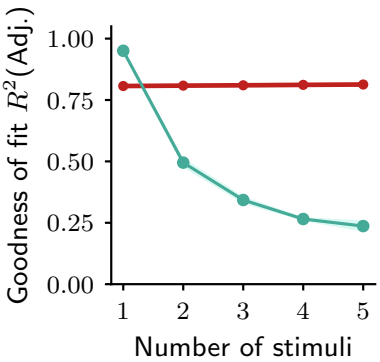

Figure S10: **Fit of population OFF responses using the recurrent and single-cell model adjusted for the number of parameters.** Relative variance explained, as quantified by the adjusted index R^2_{Adj}, when fitting the single-cell (green trace) and recurrent model (red trace) to calcium activity data, as a function of the number of simultaneously fitted stimuli. Error bars represent the standard deviation over multiple samples of the stimuli. Both fitting procedures are performed on the first 40% most responding units, to avoid normalizing the responses by very low values when fitting basis functions (see Materials and methods). The single-cell fit is performed using 10 Gaussian basis functions.

References

R. K. Alluri, G. J. Rose, J. L. Hanson, C. J. Leary, G. A. Vasquez-Opazo, J. A. Graham, and J. Wilkerson. Phasic, suprathreshold excitation and sustained inhibition underlie neuronal selectivity for short-duration sounds. *Proceedings of the National Academy of Sciences*, 113(13):1927–1935, 2016.

B. Aubie, S. Becker, and P. A. Faure. Computational models of millisecond level duration tuning in neural circuits. *Journal of Neuroscience*, 29(29):9255–9270, 2009.

D. L. Barbour and E. M. Callaway. Excitatory local connections of superficial neurons in rat auditory cortex. *Journal of Neuroscience*, 28(44):11174–11185, 2008.

C. M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, 2006.

J. K. Bizley, V. M. Bajo, F. R. Nodal, and A. J. King. Cortico-cortical connectivity within ferret auditory cortex. *Journal of Comparative Neurology*, 523(15):2187–2210, 2015.

A. Bjorck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

G. Bondanelli and S. Ostojic. Coding with transient trajectories in recurrent neural networks. *PLOS Computational Biology*, 16(2):1–36, 2020.

A. Brand, R. Urban, and B. Grothe. Duration tuning in the mouse auditory midbrain. *Journal of Neurophysiology*, 84(4):1790–1799, 2000.

B. M. Broome, V. Jayaraman, and G. Laurent. Encoding and decoding of overlapping odor sequences. *Neuron*, 51(4):467–482, 2006.

D. V. Buonomano and W. Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113–125, 2009.

B. M. Calhoun and C. E. Schreiner. Spectral envelope coding in cat primary auditory cortex: linear and non-linear effects of stimulus characteristics. *European Journal of Neuroscience*, 10(3):926–940, 1998.

M. M. Churchland and K. V. Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of Neurophysiology*, 97(6):4235–4257, 2007.

M. M. Churchland, B. M. Yu, S. I. Ryu, G. Santhanam, and K. V. Shenoy. Neural variability in premotor cortex provides a signature of motor preparation. *Journal of Neuroscience*, 26(14):3697–3712, 2006.

M. M. Churchland, J. P. Cunningham, M. T. Kaufman, S. I. Ryu, and K. V. Shenoy. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68(3):387–400, 2010.

M. M. Churchland, J. P. Cunningham, M. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487:51–56, 2012.

J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

P. T. Davies and M. K.-S. Tso. Procedures for reduced-rank regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):244–255, 1982.

T. Deneux, A. Kempf, A. Daret, E. Ponsot, and B. Bathellier. Temporal asymmetries in auditory coding and perception reflect multi-layered nonlinearities. *Nature Communications*, 7:12682, 2016.

J. Duysens, S. J. Schaasfma, and G. Orban. Cortical off response tuning for stimulus duration. *Vision Research*, 36(20):3243 – 3251, 1996.

G. F. Elsayed and J. P. Cunningham. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience*, 20:1310–1318, 2017.

R. W. Friedrich and G. Laurent. Dynamic optimization of odor representations by slow temporal patterning of mitral cell activity. *Science*, 291(5505):889–894, 2001.

Z.-Y. Fu, J. Tang, P. J. Hung-Sun, and Q.-C. Chen. The auditory response properties of single-on and double-on responders in the inferior colliculus of the leaf-nosed bat, hipposideros armiger. *Brain Research*, 1306:39 – 52, 2010.

P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262.

M. S. Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear Algebra for Signal Processing*, pages 27–49. Stanford University, 1992.

B. Grothe, M. Vater, J. H. Casseday, and E. Covey. Monaural interaction of excitation and inhibition in the medial superior olive of the mustached bat: an adaptation for biosonar. *Proceedings of the National Academy of Sciences of the United States of America*, 89(11):5108–5112, 1992.

Y. Guo and R. Burkard. Onset and offset responses from inferior colliculus and auditory cortex to paired noisebursts: inner hair cell loss. *Hearing Research*, 171(1):158 – 166, 2002.

K. E. Hancock and H. F. Voigt. Wideband inhibition of dorsal cochlear nucleus type iv units in cat: a computational model. *Annals of Biomedical Engineering*, 27(1):73–87, 1999.

D. E. H. Hartley, J. C. Dahmen, A. J. King, and J. W. H. Schnupp. Binaural sensitivity changes between cortical on and off responses. *Journal of Neurophysiology*, 106(1):30–43, 2011.

J. He. Off responses in the auditory thalamus of the guinea pig. *Journal of Neurophysiology*, 88(5):2377–2386, 2002.

J. He. Corticofugal modulation on both on and off responses in the nonlemniscal auditory thalamus of the guinea pig. *Journal of Neurophysiology*, 89(1):367–381, 2003.

P. Heil. Auditory cortical onset responses revisited. i. first-spike timing. *Journal of Neurophysiology*, 77(5): 2616–2641, 1997a.

P. Heil. Auditory cortical onset responses revisited. ii. response strength. *Journal of Neurophysiology*, 77(5): 2642–2660, 1997b.

G. Hennequin, T. P. Vogels, and W. Gerstner. Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86:011909, 2012.

G. Hennequin, T. P. Vogels, and W. Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394 – 1406, 2014.

K. R. Henry. Tuning of the auditory brainstem off responses is complementary to tuning of the auditory brainstem on response. *Hearing Research*, 19(2):115 – 125, 1985.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

T. Ito and M. S. Malmierca. Neurons, connections, and microcircuits of the inferior colliculus. In *The Mammalian Auditory Pathways: Synaptic Organization and Microcircuits*, pages 127–167. Springer International Publishing, 2018.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248 – 264, 1975.

B. Joachimsthaler, M. Uhlmann, F. Miller, G. Ehret, and S. Kurt. Quantitative analysis of neuronal response properties in primary and higher-order auditory cortical fields of awake house mice (mus musculus). *European Journal of Neuroscience*, 39(6):904–918, 2014.

M. Kasai, M. Ono, and H. Ohmori. Distinct neural firing mechanisms to tonal stimuli offset in the inferior colliculus of mice in vivo. *Neuroscience Research*, 73(3):224 – 237, 2012.

C. H. Keller, K. Kaylegian, and M. Wehr. Gap encoding by parvalbumin-expressing interneurons in auditory cortex. *Journal of Neurophysiology*, 120(1):105–114, 2018.

A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2009–2041, 2002.

C. Kopp-Scheinpflug, A. J. Tozer, S. W. Robinson, B. L. Tempel, M. H. Hennig, and I. D. Forsythe. The sound of silence: ionic mechanisms encoding sound termination. *Neuron*, 71(5):911 – 925, 2011.

C. Kopp-Scheinpflug, J. L. Sinclair, and J. F. Linden. When sound stops: offset responses in the auditory system. *Trends in Neurosciences*, 41(10):712–728, 2018.

S. Kuwada and R. Batra. Coding of sound envelopes by inhibitory rebound in neurons of the superior olivary complex in the unanesthetized rabbit. *Journal of Neuroscience*, 19(6):2273–2287, 1999.

A. H. Lara, J. P. Cunningham, and M. M. Churchland. Different population dynamics in the supplementary motor area and motor cortex during reaching. *Nature Communications*, 9(1):2754, 2018.

C. C. Lee, A. U. Kishan, and J. A. Winer. Wiring of divergent networks in the central auditory system. *Frontiers in neuroanatomy*, 5:46–46, 2011.

L. Y. Li, Y. T. Li, M. Zhou, H. W. Tao, and L. I. Zhang. Intracortical multiplication of thalamocortical signals in mouse auditory cortex. *Nature Neuroscience*, 16:1179–1181, 2013.

Y. Lim, R. Lagoy, B. G. Shinn-Cunningham, and T. J. Gardner. Transformation of temporal sequences in the zebra finch auditory system. *eLife*, 5:e18205, 2016.

J. F. Linden and C. E. Schreiner. Columnar transformations in auditory cortex? A comparison to visual and somatosensory cortices. *Cerebral Cortex*, 13(1):83–89, 2003.

J. Liu, M. R. Whiteway, A. Sheikhattar, D. A. Butts, B. Babadi, and P. O. Kanold. Parallel processing of sound dynamics across mouse auditory cortex via spatially patterned thalamic inputs and distinct areal intracortical circuits. *Cell Reports*, 27(3):872 – 885, 2019a.

X. Liu, O. Zhang, J. Qi, A. Chen, K. Hu, and J. Yan. The onset and post-onset auditory responses of cochlear nucleus neurons are modulated differently by cortical activation. *Hearing Research*, 373:96 – 102, 2019b.

F. Luo, W. Metzner, F. J. Wu, S. Y. Zhang, and Q. C. Chen. Duration-sensitive neurons in the inferior colliculus of horseshoe bats: adaptations for using cf-fm echolocation pulses. *Journal of Neurophysiology*, 99(1):284–296, 2008.

C. K. Machens, M. S. Wehr, and A. M. Zador. Linearity of cortical receptive fields measured with natural sounds. *Journal of Neuroscience*, 24(5):1089–1100, 2004.

F. Mastrogiuseppe and S. Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609 – 623, 2018.

O. Mazor and G. Laurent. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron*, 48(4):661–673, 2005.

A. F. Meyer, R. S. Williamson, J. F. Linden, and M. Sahani. Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. *Frontiers in systems neuroscience*, 10:109–109, 2017.

A. Mukherjee, K. Chen, N. Wang, and J. Zhu. On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477, 2015.

B. K. Murphy and K. D. Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

I. Nelken, Y. Rotman, and O. B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715):154–157, 1999.

M. G. Neubert and H. Caswell. Alternatives to resilience for measuring the responses of ecological systems to perturbations. 78(3):653–665, 1997.

A.-M. M. Oswald and A. D. Reyes. Maturation of intrinsic and synaptic properties of layer 2/3 pyramidal neurons in mouse auditory cortex. *Journal of Neurophysiology*, 99(6):2998–3008, 2008.

A.-M. M. Oswald, B. Doiron, J. Rinzel, and A. D. Reyes. Spatial profile and differential recruitment of gaba b modulate oscillatory activity in auditory cortex. *Journal of Neuroscience*, 29(33):10321–10334, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

D. P. Phillips, S. E. Hall, and S. E. Boehnke. Central auditory onset responses, and temporal asymmetries in auditory perception. *Hearing Research*, 167(1):192 – 205, 2002.

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

G. D. Pollak and R. D. Bodenhamer. Specialized characteristics of single units in inferior colliculus of mustache bat: frequency representation, tuning, and discharge patterns. *Journal of Neurophysiology*, 46 (3):605–620, 1981.

L. Qin, S. Chimoto, M. Sakai, J. Wang, and Y. Sato. Comparison between offset and onset responses of primary auditory cortex on–off neurons in awake cats. *Journal of Neurophysiology*, 97(5):3421–3431, 2007.

E. D. Remington, S. W. Egger, D. Narain, J. Wang, and M. Jazayeri. A dynamical systems perspective on flexible motor timing. *Trends in Cognitive Sciences*, 22(10):938 – 952, 2018.

G. Rothschild, I. Nelken, and A. Mizrahi. Functional organization and population dynamics in the mouse primary auditory cortex. *Nature Neuroscience*, 13(3):353–360, 2010.

Y. Rotman, O. Bar-Yosef, and I. Nelken. Relating cluster and population responses to natural sounds and tonal stimuli in cat primary auditory cortex. *Hearing Research*, 152(1):110 – 127, 2001.

D. Saha, W. Sun, C. Li, S. Nizampatnam, W. Padovano, Z. Chen, A. Chen, E. Altan, R. Lo, D. L. Barbour, and B. Raman. Engaging and disengaging recurrent inhibition coincides with sensing and unsensing of a sensory stimulus. *Nature Communications*, 8:15413, 2017.

M. Sahani and J. F. Linden. How linear are auditory cortical responses? In *Advances in Neural Information Processing Systems 15*, pages 125–132. 2003.

C. Savin and G. Tkačik. Maximum entropy models as a tool for building precise neural controls. *Current Opinion in Neurobiology*, 46:120 – 126, 2017.

S. Saxena and J. P. Cunningham. Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55:103 – 111, 2019.

B. Scholl, X. Gao, and M. Wehr. Nonoverlapping sets of synapses drive on responses and off responses in auditory cortex. *Neuron*, 65(3):412–421, 2010.

F. Schuessler, A. Dubreuil, F. Mastrogiuseppe, S. Ostojic, and O. Barak. Dynamics of random recurrent networks with correlated low-rank structure. *Phys. Rev. Research*, 2:013111, 2020.

J. S. Seely, M. T. Kaufman, S. I. Ryu, K. V. Shenoy, J. P. Cunningham, and M. M. Churchland. Tensor analysis reveals distinct population structure that parallels the different computational roles of areas m1 and v1. *PLOS Computational Biology*, 12(11):1–34, 2016.

K. V. Shenoy, M. T. Kaufman, M. Sahani, and M. M. Churchland. A dynamical systems view of motor preparation: implications for neural prosthetic system design. *Progress in Brain Research*, 192:33 – 58, 2011.

R. L. Smith and M. L. Brachman. Operating range and maximum response of single auditory nerve fibers. *Brain Research*, 184(2):499 – 505, 1980.

R. L. Smith and M. L. Brachman. Adaptation in auditory-nerve fibers: A revised model. *Biological Cybernetics*, 44(2):107–120, 1982.

J. Sollini, G. A. Chapuis, C. Clopath, and P. Chadderton. On-off receptive fields in auditory cortex diverge during development and contribute to directional sweep selectivity. *Nature Communications*, 9(1):2084, 2018.

M. Stopfer, V. Jayaraman, and G. Laurent. Intensity versus identity coding in an olfactory system. *Neuron*, 39(6):991–1004, 2003.

C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

J. P. Stroud, M. A. Porter, G. Hennequin, and T. P. Vogels. Motor primitives in space and time via targeted gain modulation in cortical networks. *Nature Neuroscience*, 21(12):1774–1783, 2018.

N. Suga. Single unit activity in cochlear nucleus and inferior colliculus of echo-locating bats. *The Journal of Physiology*, 172(3):449–474, 1964.

D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18:1025–1033, 2015.

L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

V. Uurtio, J. a. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu. A tutorial on canonical correlation methods. *ACM Comput. Surv.*, 50(6):1–33, 2017.

M. Vater, H. Habbicht, M. Kössl, and B. Grothe. The functional role of gaba and glycine in monaural and binaural processing in the inferior colliculus of horseshoe bats. *Journal of Comparative Physiology A*, 171 (4):541–553, 1992.

X. Wang. Neural coding strategies in auditory cortex. *Hearing Research*, 229(1):81 – 93, 2007.

X. Wang, T. Lu, R. K. Snider, and L. Liang. Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, 435(7040):341–346, 2005.

R. S. Williamson, M. B. Ahrens, J. F. Linden, and M. Sahani. Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. *Neuron*, 91(2):467–481, 2016.

J. A. Winer, D. T. Larue, J. J. Diehl, and B. J. Hefti. Auditory cortical projections to the cat inferior colliculus. *Journal of Comparative Neurology*, 400(2):147–174, 1998.

N. Xu, Z.-Y. Fu, and Q.-C. Chen. The function of offset neurons in auditory information processing. *Translational Neuroscience*, 5(4):275–285, 2014.

Y.-Q. Yu, Y. Xiong, Y.-S. Chan, and J. He. In vivo intracellular responses of the medial geniculate neurones to acoustic stimuli in anaesthetized guinea pigs. *The Journal of physiology*, 560:191–205, 2004.