Population coding and network dynamics during OFF responses in auditory cortex

Giulio Bondanelli [1], 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

Abstract

Across sensory systems, complex spatio-temporal patterns of neural activity arise following the onset (ON) and offset (OFF) of simple stimuli. While ON responses have been widely studied, the mechanisms generating OFF responses in cortical areas have so far not been fully elucidated. Recent studies have argued that OFF responses reflect strongly transient sensory coding at the population level, suggesting they may be generated by a collective, network mechanism. We examine here the hypothesis that OFF responses are single-cell signatures of network dynamics and propose a simple model that generates transient OFF responses through recurrent interactions. To test this hypothesis, we analyse the responses of large populations of neurons in auditory cortex recorded using two-photon calcium imaging in awake mice passively listening to auditory stimuli. Adopting a population approach, we find that the OFF responses evoked by individual stimuli define trajectories of activity that evolve in low-dimensional neural subspaces. A geometric analysis of the population OFF responses reveals that, across different stimuli, these neural subspaces lie on largely orthogonal dimensions that define low-dimensional transient coding channels. We show that a linear network model with recurrent interactions provides a simple mechanism for the generation of the strong OFF responses observed in auditory cortex, and accounts for the low-dimensionality of the transient channels and their global structure across stimuli. We finally compare the network model with a single-cell mechanism and show that the single-cell model, while explaining some prominent features of the data, does not account for the structure across stimuli 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, is able to evoke 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 offset-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 constitute from 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 activity patterns that have been hypothesised to serve as the basis for the generation of muscle activity (Churchland et al., 2012; Sussillo et al., 2015; Hennequin et al., 2014; Stroud et al., 2018). In the auditory cortex, this approach has suggested a central role of the neural dynamics across large population for the coding of different auditory features (Deneux et al., 2016; Saha et al., 2017). Yet, how a structured population code arises from the neural transient dynamics and how these dynamics are generated are still open questions.

Leveraging on 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), in this study we explore the hypothesis that transient OFF responses may be generated by network mechanisms. We analyse OFF responses evoked by multiple stimuli recorded using calcium imaging across a large population of neurons in the auditory cortex of mice passively listening to ramping auditory stimuli (Deneux et al., 2016). Across different stimuli, OFF responses were elicited in overlapping subsets of neurons, where individual neurons responded to more than one stimulus. Analyses of the population activity through dimensionality reduction techniques (Cunningham and Yu, 2014; Kobak et al., 2016; Bagur et al., 2018; Stringer et al., 2019) revealed that OFF responses encode different stimuli in orthogonal low-dimensional subspaces, consistent with the existence of a neural code for OFF responses at the ensemble level (Saha et al., 2017). We interpret these results by examining a linear network model with recurrent connectivity (Bondanelli and Ostojic, 2018), and show that it provides a simple mechanism for the generation of the strong single-cell OFF responses observed in data from auditory cortex. Crucially, we show that the proposed mechanism for OFF response generation can also explain the low-dimensional patterns of the population neural activity elicited in response to specific stimuli, as well as the global structure of the population OFF

responses across multiple stimuli. We finally compare the network model with a simplified single-cell model, and show that the single-cell model can reproduce population dynamics in response to individual stimuli, but cannot capture the structure across stimuli, in contrast to the network model.

Results

Amplified ON/OFF population responses in auditory cortex

We analyse the population responses of 2434 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 1 in *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 represents 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 its baseline level $\mathbf{r}_B$, 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 amplification of ON and OFF responses.

To examine what the transient amplification of ON and OFF responses implies in terms of stimulus decoding, we train a simple decoder to classify pairs of stimuli that differ 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).

Low-dimensional dynamics during OFF responses

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 350 ms after stimulus termination.

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 a few of the available 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). 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 is that the OFF response trajectories relative to stimuli with different frequency content span orthogonal low-dimensional subspaces.

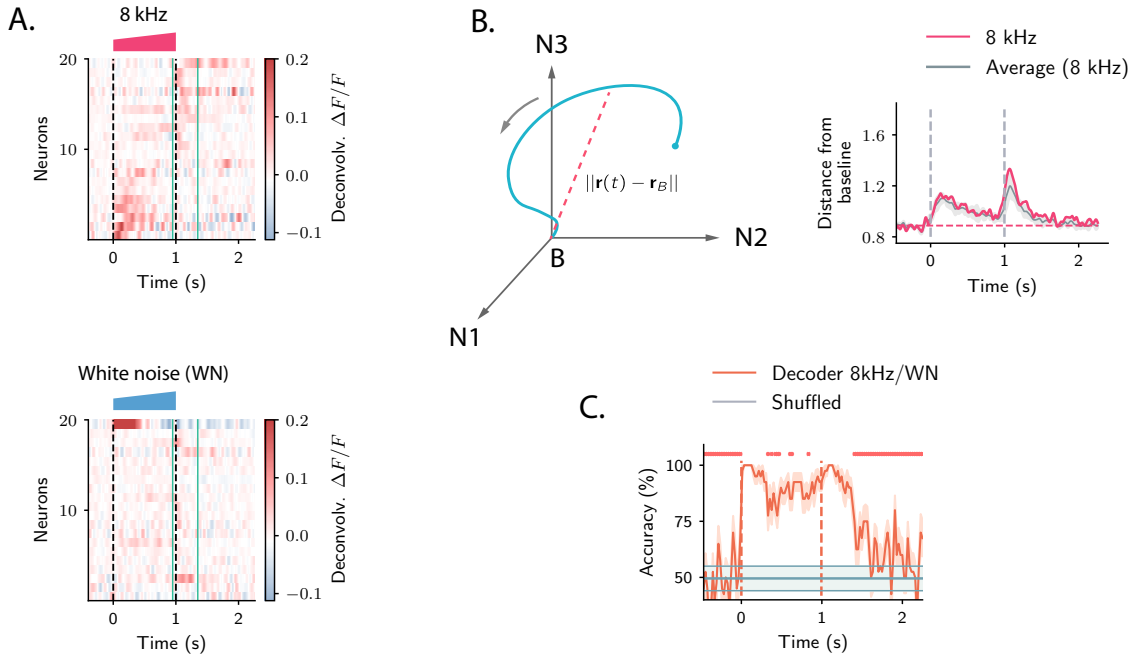

Figure 1: **Strong transient ON and OFF responses in auditory cortex of mice passively listening to different sounds and stimulus decoding. A.** *Top*: deconvolved calcium signals averaged over 20 trials showing the activity (estimated firing rate) of 20 out of 2343 neurons in response to a 8 kHz 1s 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 350 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 $||\mathbf{r}(t)||$ computed for the population response to the 8 kHz sound in **A**. The grey trace shows the distance from baseline averaged over all the 8 kHz 1 s sounds (4 stimuli). Gray shading represents ± 1 standard deviation. The dashed horizontal line shows the average level of the distance $||\mathbf{r}(t)||$ before stimulus presentation (even if baseline-subtracted responses are used, a value of the norm different from zero is expected because of the 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).

For instance, the response to the 8 kHz sound is poorly represented on the plane defined by the two principal component of the response to the white noise sound (Fig. 2B), and conversely, showing that they evolve 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 *Methods*).

This approach revealed an interesting structure between the OFF response subspaces for different stimuli (Fig. 2D). 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). This number is in fact close to the number of dominant components of the joint response to all 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.

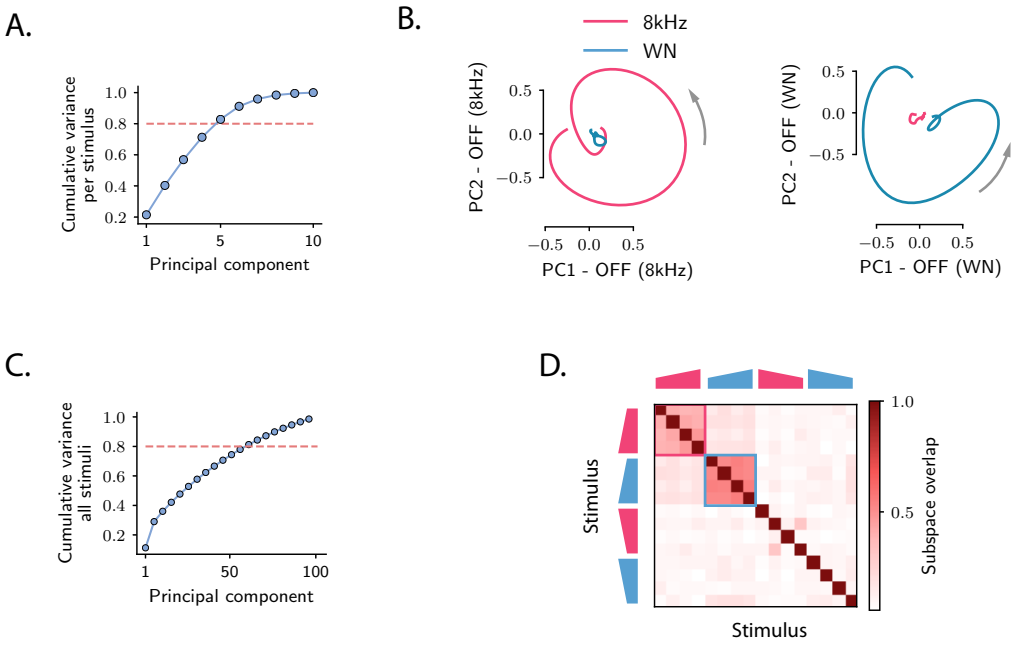

Figure 2: **Low-dimensional structure of population OFF responses. A.** Cumulative variance explained for OFF responses to individual stimuli as a function of the number of principal components. The trace shows the average cumulative variance over all 16 stimuli. Error bars are smaller than the symbol 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 350 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 to stimuli s_1 and s_2. The overlap is measured by the principal angle between these subspaces. 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: a dynamical system approach

The responses following the end of stimulus presentation in auditory cortex resemble qualitatively the neural activity in motor and premotor cortex following movement onset. At the single neuron level, this similarity is reflected in the strong transient responses present in both areas (Churchland and Shenoy, 2007; Churchland et al., 2010, 2012). At the population level, both OFF responses and motor activity display rotational-like structure Churchland et al. (2012). A prominent hypothesis has been that neural activity patterns during movement execution can be explained by a dynamical system mechanism (Hennequin et al., 2014; Stroud et al., 2018; Sussillo et al., 2015; Shenoy et al., 2011). Here we examine to which extent population OFF responses in auditory cortex can be accounted for using a similar mechanism based on recurrent network dynamics.

We consider a model network of N randomly and recurrently coupled linear rate units with dynamics

given by:

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

Such networks can be interpreted as describing the linearized dynamics of a system around an equilibrium state. In this picture, the quantity $r_i(t)$ represents the deviation of the activity of the unit i from its baseline firing activity, while J_{ij} denotes the effective strength of the connection from neuron j to neuron i, which can take positive or negative values (Fig. 3A). Without loss of generality, we assume that the baseline level is $\mathbf{r}_B = 0$.

We model the firing activity at the end of stimulus presentation as the initial condition of the network dynamics. If at the time of stimulus offset the firing rate of neuron i is equal to $r_{0,i}$, the initial condition of the network across all units is set to the activity vector $\mathbf{r}_0$. The evoked OFF responses can then be described as a trajectory $\mathbf{r}(t)$ in a high-dimensional state-space in which the i-th component represents the firing rate of neuron i at time t. We assume that the network receives no external input after stimulus offset. Within this simplifying assumption, OFF responses are the result of the internally generated dynamics following stimulus offset, and are therefore uniquely determined by the initial condition of the network $\mathbf{r}_0$.

We focus on two outstanding properties identified in the OFF responses to different stimuli. The first one is the strong amplification of OFF responses, as quantified by the transient increase of the distance from baseline $||\mathbf{r}(t)||$. The second important feature is the low-dimensionality of OFF response trajectories and their global structure across different stimuli. To account for these features we consider a minimal connectivity structure of the form:

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

where the $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ are N-dimensional vectors. Connectivity matrices of this form are called low-rank connectivities (Mastrogiuseppe and Ostojic, 2018; Hopfield, 1982). They belong to the broader class of non-normal matrices (Trefethen and Embree (2005); Murphy and Miller (2009); Goldman (2009); Hennequin et al. (2012); see *Methods*) and can produce amplified transient dynamics in responses to specific stimuli. However, the non-normality of the connectivity matrix alone does not guarantee the existence of amplified OFF responses, corresponding to an transient increase of the distance from baseline. Instead, the generation of amplified OFF responses requires one additional condition on the network connectivity, namely that the largest eigenvalue of the symmetric part defined by $\mathbf{J}_S = (\mathbf{J} + \mathbf{J}^T)/2$ is larger than unity (Bondanelli and Ostojic, 2018).

We first consider the case in which the connectivity is defined by one single term in Eq. (2) ($P = 1$), resulting in the unit-rank connectivity $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$, and focus on the case in which the vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ are orthogonal to each other. The dimensions defined by $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ essentially correspond to the first two principal components of the population responses following initial activity along $\mathbf{v}^{(1)}$ (see *Methods*). When the condition for amplification is met, initial activity along $\mathbf{v}^{(1)}$ elicits an amplified OFF response (Fig. 3B-C), and the evoked trajectory lies in the plane defined by $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$, therefore spanning two dimensions (Fig. 3D). A unit-rank connectivity can therefore support the encoding of a single stimulus (the one associated with $\mathbf{v}^{(1)}$) in the evoked transient OFF response dynamics. Crucially, the population activity vector at the peak of the transient dynamics lies along the vector $\mathbf{u}^{(1)}$ (see *Methods*). It follows that the unit-rank network generates amplified dynamics which decorrelates the population activity at the end of stimulus presentation (the initial condition $\mathbf{v}^{(1)}$) and the state at the peak of the OFF response (the vector $\mathbf{u}^{(1)}$; Fig. 3E).

When the connectivity defined in Eq. (2) includes P unit-rank terms, a number P of initial conditions, corresponding to the vectors $\mathbf{v}^{(i)}$'s, can evoke amplified OFF responses. An interesting scenario arises when the planes defined by each pair of vectors $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ are mutually orthogonal. In this case, the dimensions defined by the vector $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ correspond to the first $2P$ principal component of the joint population response elicited by activity along the $\mathbf{v}^{(i)}$'s. This structure implies that stimuli associated with orthogonal initial conditions along the $\mathbf{v}^{(i)}$'s evoke two-dimensional OFF response trajectories that lie in orthogonal subspaces, the $\mathbf{u}^{(i)}$-$\mathbf{v}^{(i)}$ planes (Fig. 3D). Each term in Eq. (2) encodes one single stimulus, corresponding to a particular initial condition $\mathbf{v}^{(i)}$, and can therefore be interpreted as a transient coding channel. It follows that the low-rank connectivity given by the sum of P unit-rank terms can support the encoding of P distinct stimuli through low-dimensional transient trajectories that span orthogonal subspaces. In this simplified set-up, the OFF responses to each stimulus span two dimensions. Nonetheless, low-rank connectivity structures of the form given by Eq. (2) can produce higher-dimensional transient dynamics if correlations between the single transient coding channels are introduced. This can be achieved for instance

by including correlations between pairs of vectors $\mathbf{v}^{(i+1)}$ and $\mathbf{u}^{(i)}$ (Sompolinsky and Kanter, 1986; Ganguli et al., 2008), preserving at the same time the orthogonality between channels where this correlation vanishes.

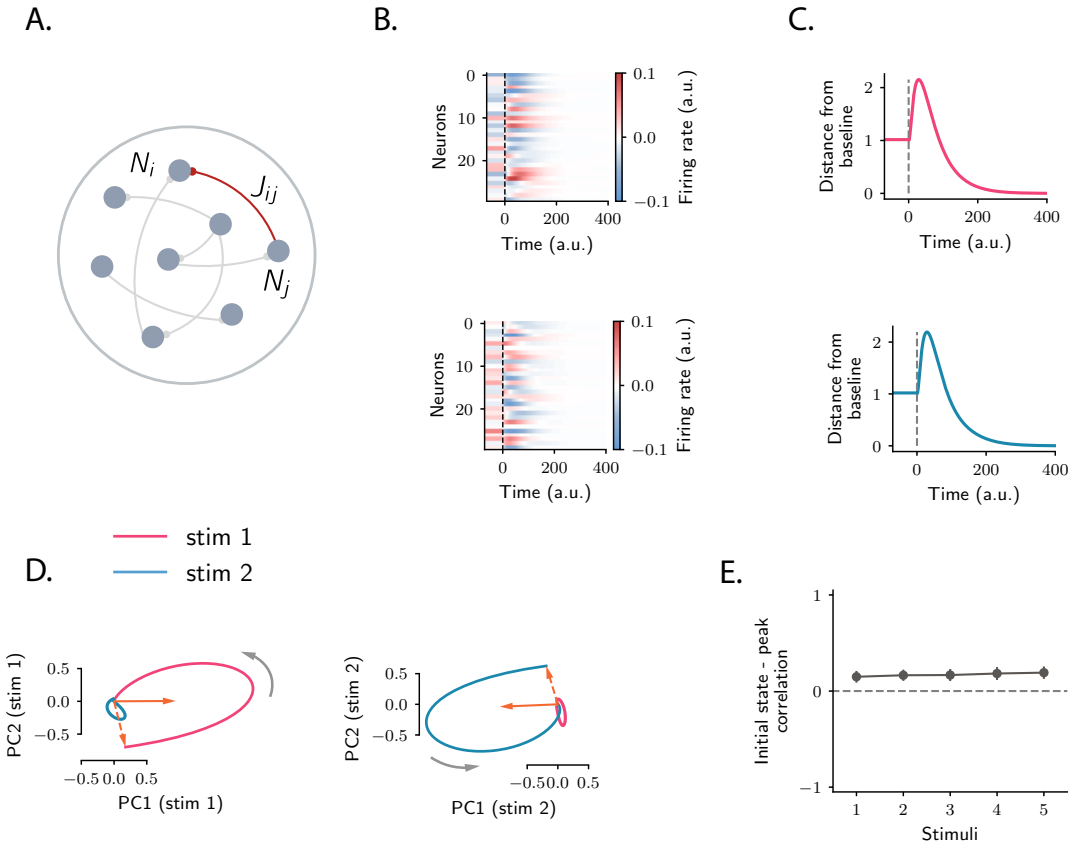

Figure 3: **Strong orthogonal population OFF responses in a network model with low-rank structure. A.** Illustration of the recurrent network model. The variables defining the state of the system are the firing rates of the units, denoted by r_i. The strength of the connection from unit j to unit i is denoted by J_{ij}. **B.** Single-unit OFF responses to two distinct stimuli generated using a low-rank network model. Different stimuli are modeled as different states at the end of stimulus presentation ($t = 0$), corresponding to the vectors $\mathbf{v}_i$ in Eq. (2). 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.** 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 represent the standard deviation computed over 100 bootstrap subsamplings of 10% of the units in the population. We simulated a network of 2000 units, with rank equal to 30 and $||\mathbf{u}_1|| = ||\mathbf{u}_2|| = 5$.

Testing the network mechanism on the data

In the following, we proceed to test the specific predictions of the proposed network mechanism on the data from auditory cortex. In the low-rank network given by Eq. (2), the recurrent dynamics transform orthogonal population activity states at the end of stimulus presentation into orthogonal population trajectories during the subsequent OFF responses. Conversely, correlated initial states at the end of stimulus presentation generate correlated dynamical trajectories. This implies that the geometric structure of the activity states at the end of stimulus presentation predicts the geometric structure of the dynamics during the OFF responses. To investigate this aspect, we computed the cross-correlations between the population states reached at the

end of stimulus presentation for all different stimuli, and found that this structure matched the pattern of correlations observed for the subspace overlap (Fig. 4A *left panel*). Therefore, the structure of the population activity at the end of stimulus presentation across stimuli was highly correlated with the structure of the dynamics during OFF responses (Fig. 4A *right panel*, S6), consistent with the interpretation that activity at the end of stimulus presentation provides an initial condition for the recurrent dynamics that generates the OFF responses.

This result could be alternatively explained by a simple geometrical model in which the network dynamics induce a one-dimensional linear scaling of the population activity vector (Bartho et al., 2009). By construction, this mechanism would predict that the population activity state at the peak of the transient OFF response is a scaled version of the state at the end of stimulus presentation, which would directly produce the observed pattern of correlations. In contrast, our network model predicts that, 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 near zero values of the correlation (Fig. 3E). Computing the correlation between the peak state and the initial state at the end of stimulus presentation we find that this correlation takes low values for all stimuli (Fig. 4B), consistent with the network model but not with the hypothesis of a one-dimensional scaling of the activity.

To further examine to which extent the population dynamics during OFF responses could be accounted for by a non-normal linear dynamical system, we next fitted our network model population OFF responses to all stimuli at once using reduced-rank regression (Fig. S1; see *Methods*). Qualitatively, we found that the fit reproduced well the low-dimensional population trajectories (Fig. 4C). We evaluated the goodness of the fitted trajectories 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. 4D). A concern that needs to be addressed is that this result may not reflect a dynamical system structure in the data, but may instead be due to correlations that do not involve a dynamical system mechanism. For instance, it could be a consequence of the temporal smoothness of the trajectories, i.e. its temporal correlations. To test for this possibility, we repeated the model fitting on control dataset which kept no additional structure than the one defined by correlations across time, neurons and stimuli of the original neural responses (Elsayed and Cunningham (2017), see *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 fitting the control datasets (Fig. 4D,S2,S3). This indicates that the recurrent network model captures structure in the data beyond its correlations across time, neurons and stimuli.

To test that the fitted network produces amplified transients through non-normal dynamics, we examined the spectra of the full connectivity $\mathbf{J}$ and of its symmetric part $\mathbf{J}_S$ (Bondanelli and Ostojic, 2018). The eigenvalues of the fitted connectivity matrix had real part smaller than one, indicating stable dynamics, and large imaginary part (Fig. 4E). However, the properties of the eigenvalues of $\mathbf{J}$ do not guarantee the existence of amplified OFF responses. Instead, 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. 4E), and could therefore produce amplified responses (Fig. 4C).

Finally we asked whether the fitted connectivity $\mathbf{J}$ consisted of independent low-rank coding channels, as postulated in our network model (Eq. (2)). To test for this aspect and identify the low-rank coding channels in the fitted connectivity, we proceeded in two steps. We first fitted the recurrent model to all OFF responses at the same time. We call $\mathbf{J}_{\text{Full}}$ the connectivity matrix that resulted from the fit. Next, we fitted the recurrent model to the OFF responses to each stimulus s independently, obtaining one matrix $\mathbf{J}^{(s)}$ for each stimulus (the rank parameter for each stimulus was set to $r = 8$). Each matrix $\mathbf{J}^{(s)}$ corresponds to one term in Eq. (2), with the difference that its rank is not unitary but equals the dimensionality of individual OFF responses. The matrix $\mathbf{J}^{(s)}$ represent therefore the transient coding channel for stimulus s. We computed the overlap between the subspaces spanned by the transient channels for each pair of stimuli s_1 and s_2, analogous to the overlap between the vectors $\mathbf{u}^{(s_1)}$ and $\mathbf{u}^{(s_2)}$ in Eq. (2) (see *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. 4F, right panel), consistent with the interpretation that individual OFF responses may be generated through transient coding channels of the form given by Eq. (2). To further examine if the matrix $\mathbf{J}$ indeed consisted of the sum of the low-rank channels given by each individual $\mathbf{J}^{(s)}$, we 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. 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 $\mathbf{J}^{(s)}$.

In summary we have shown that, across stimuli, the structure of the initial states at the end of stimulus presentation is preserved during the OFF transient dynamics. A simple one-dimensional scaling of the population activity does not account for this finding, since the OFF responses trajectories evolved along dimensions orthogonal to the initial states. Instead, this structure is consistent with a mechanism based on recurrent network dynamics, which generates amplified OFF responses through non-normal connectivity. We found that a specific class of low-rank non-normal connectivities is able to account for the low-dimensionality of OFF responses and the global structure of the corresponding transient channels across stimuli.

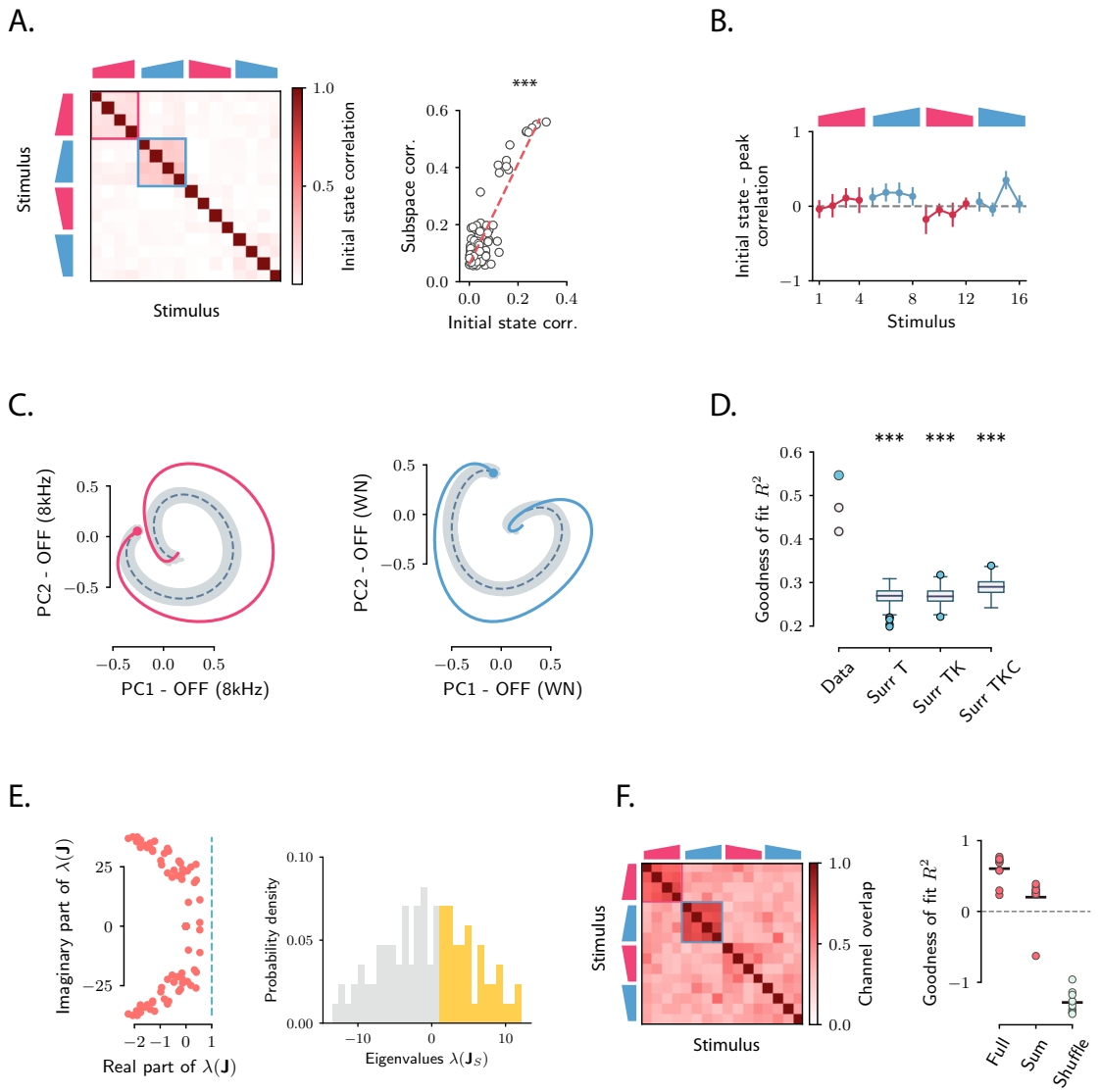

Figure 4

Figure 4 *(previous page)*: **Testing the predictions of the network model on auditory cortical data. A.** *Left*: correlation between the population activity vectors at the end of stimulus presentation (50 ms before offset) for each pair of stimuli. *Right*: linear correlation between the overlap between dynamical subspaces (Fig. 2D) and the overlap between initial conditions for each stimulus pair. **B.** 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 100 bootstrap subsamplings of 10% of the neurons in the population (2343 neurons). **C.** 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 350 ms with respect to stimulus offset. Light grey traces: projections of multiple trajectories generated by the network model using the fitted connectivity matrix. 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. **D.** 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 across multiple surrogates for mouse 1 are shown. 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. **E.** 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 1 determine strongly non-normal dynamics and are highlighted in yellow. The parameters used for reduced-rank ridge regression are the same as in **C. F.** Left: overlap between the transient channels relative to each stimulus, as quantified by the overlap between the left singular values $\mathbf{L}^{(s)}$ of the connectivities $\mathbf{J}^{(s)}$ fitted for each individual stimulus (see *Methods*). *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 (10-fold). In all panels, the fit is computed using reduced-rank ridge regression (see *Methods*), where the best ridge and rank hyperparameters are found using cross-validation (see *Methods* and Fig. S1). The number of principal component is set to 100.

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 linear model for OFF response generation that mimics single-cell activation after stimulus offset (Fig. 2 in Kopp-Scheinpflug et al. (2018)). We assume that neuron i in the population responds after stimulus offset as specified by a characteristic response filter $L_i(t)$ (Meyer et al., 2017), which describes the cell's intrinsic response generated by intracellular or synaptic mechanisms. While the shape of the temporal response of a neuron is assumed to be set by intrinsic properties and is therefore the same across different stimuli, each stimulus modulates the response of each neuron in a multiplicative way, so that the OFF response of neuron i to stimulus s can be written as:

$$r_i(t) = r_{0,i}^{(s)} \, L_i(t) \tag{3}$$

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 different from zero, so that we can interpret the vector of modulation factors $\mathbf{r}_0^{(s)}$ as the vector of initial conditions.

We choose the shape of the temporal responses $L_i(t)$'s to be non-monotonic functions with a single peak at a specific latency $t_{L,i}$ (Fig. 5A; see *Methods*). To account for the OFF response patterns observed in the data, the latencies $t_{L,i}$ are distributed across the population. If the response latencies were the same for all neurons, the dynamics would lie on a single dimension, corresponding to the vector of initial conditions

$\mathbf{r}_0^{(s)}$ (see *Methods*), and the OFF response dynamics would simply scale the population state at the end of stimulus presentation, a scenario we ruled out earlier.

We first show that this model is able to account for some of the prominent features observed in the data, namely the heterogeneity of responses across neurons, non-monotonic dynamics of the distance from baseline, two-dimensional neural trajectories and orthogonal responses for distinct stimuli. For a particular stimulus, randomly choosing the values of the initial conditions $r_{0,i}^{(s)}$ results in some unit increasing, some decreasing their firing rates during the OFF response (Fig. 5B). Random initial conditions also imply that there are units that respond to multiple stimuli. Because the single-cell responses are non-monotonic (due to non-monotonic $L_i(t)$'s), the distance from baseline of the population activity vector shows amplified dynamics, as observed in the calcium activity data (Fig. 5C). At the population level, the trajectories generated by the single-cell model span at least two dimensions (Fig. 5C) and resemble qualitatively the trajectories generated by the recurrent model (Fig. 3D). When the vectors $\mathbf{r}_0^{(s)}$ of the initial conditions are chosen to be orthogonal to each other for distinct stimuli, the trajectories evoked in response to those stimuli lie on orthogonal subspaces (Fig. 5D), therefore accounting for the orthogonal structure observed in auditory cortical OFF responses (Fig. 2D,4A).

Although the single-cell mechanism is consistent with a number of observations on single-cell and population structure during OFF responses in auditory cortex, we found that it cannot account for other features in the data. The first discrepancy concerns the decorrelation between the state at the end of stimulus presentation and the state at the peak of the transient phase. The single-cell model predicts that these two states are only weakly decorrelated and essentially lie along a single dimension (Fig. 5E and S7), which is inconsistent with experimental observations.

A second important difference between the two models is that in the single-cell model the temporal shape of the response of each neuron is the same across all stimuli, while for the recurrent model this is not necessarily the case. To illustrate the consequences of this difference, we simulated OFF response trajectories from both models and fitted the two models to each set of data. For the recurrent model we fit the dynamical system, as already described; for the single-cell model we fit basis functions to the responses of all the neurons subject to prior normalized by the initial conditions $r_{0,i}^{(s)}$ (see *Methods*). For both models the fit is performed on the population responses progressively increasing the number of stimuli. For data generated by the single-cell model, we find that fitting basis functions yields equivalent performance irrespective of the number of stimuli considered. Importantly, the goodness of fit (R^2) using basis functions is considerably higher than the goodness of fit for the recurrent model (Fig. 5G), indicating that the recurrent model is not able to capture the precise shape of the temporal responses of the neurons, while the single-cell model can in principle capture arbitrary shapes of the individual responses. For data generated by the recurrent model, we instead find that the goodness of fit using the basis functions method decreases as more stimuli are included in the fit, because the shape of the temporal responses of the neurons changes with the stimulus. Conversely, fitting the recurrent model yields higher and stable performance even when the fit is performed on responses to multiple stimuli at once (Fig. 5H).

Fitting the two models to the calcium activity data, we found that the goodness of fit for the single-cell model decreases as a function of the number of stimuli included, while for the recurrent model it maintains a higher value even when multiple stimuli are included. This indicates that the temporal response of individual neurons differs in general across stimuli (Fig. 5F), a key feature captured by the recurrent network model and not by the single-cell model. Computing the goodness of fit by taking into account the number of parameters of the recurrent and single-cell models leads to the same conclusion (Fig. S8).

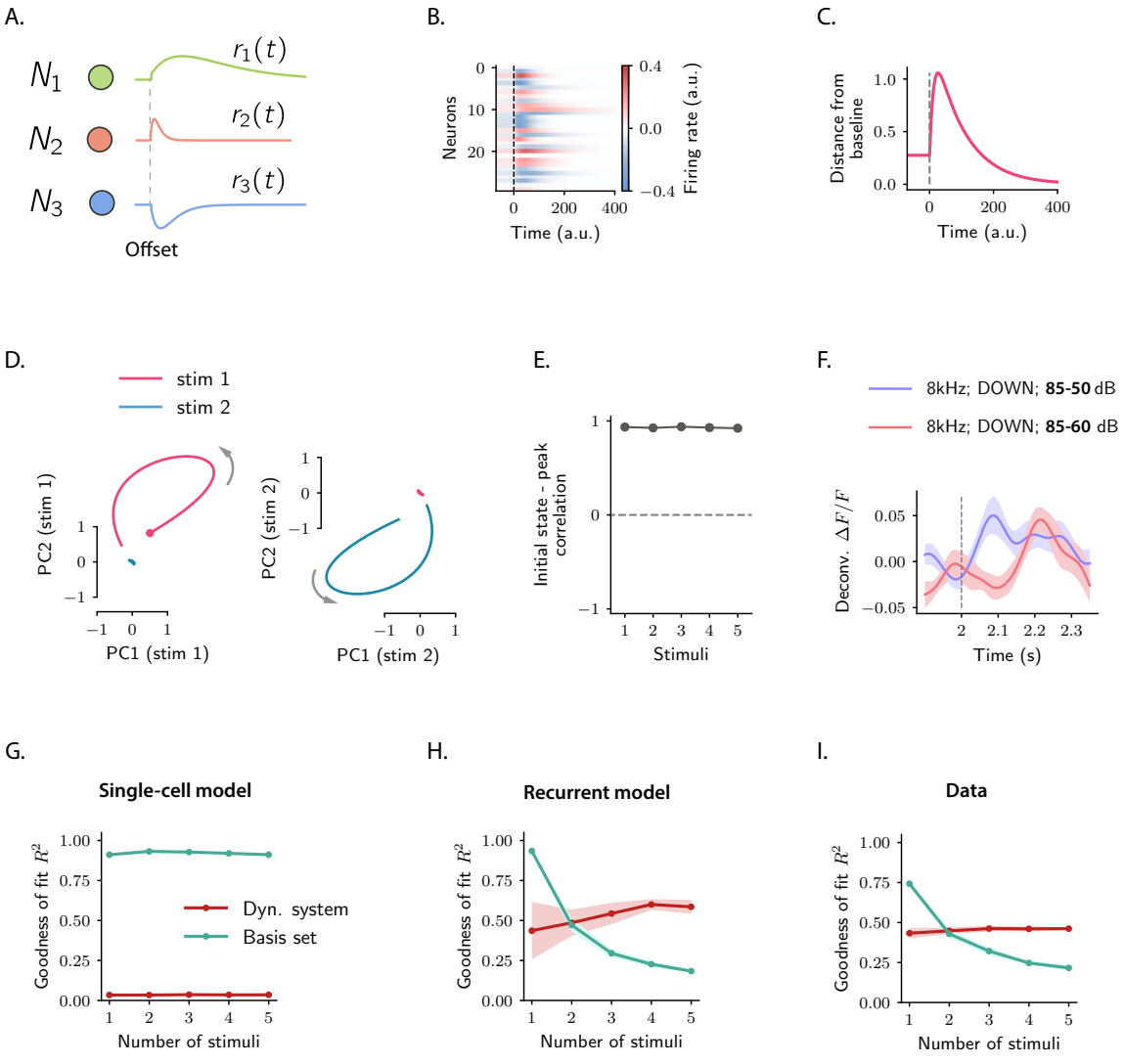

Figure 5

Figure 5 *(previous page)*: **Comparison between single-cell and network models for OFF response generation. A.** Cartoon illustrating the single-cell model defined by Eq. (3). The OFF responses of three example units to one stimulus are shown. The activity of each unit is described by a characteristic response filer $L_i(t)$ (shape of the colored traces) with a specific response latency, defined as the peak of its OFF response. Across stimuli, only the relative activation of the units changes, but the shape of the responses of each unit $L_i(t)$, including its latency, remains constant. **B.** Single-unit OFF responses generated by the single-cell model defined by Eq. (3). **C.** Distance from baseline of the population activity vector during the OFF response to the stimulus in the upper panel. Dashed line indicate the time of stimulus offset. **D.** Projection of the population OFF responses to two distinct stimuli on the two principal components of either stimuli. Distinct stimuli are modeled choosing a different relative activation vectors $\mathbf{r}_0^{(s)}$ for each stimulus s. When the relative activation vectors $\mathbf{r}_0^{(1)}$ and $\mathbf{r}_0^{(2)}$ are orthogonal the population OFF responses for the two stimuli lie on orthogonal subspaces. **E.** Correlation between the state at the end of stimulus presentation and the state at the peak of the OFF responses, defined as the time of maximum distance from baseline, for 5 example stimuli. Error bars (smaller than symbol size) represent the standard deviation computed over 100 bootstrap subsamplings of 10% of the units in the population. **F.** Example of a auditory cortical neuron that responds at the offset of two distinct stimuli with different response profiles. In particular, the overall response range is comparable across the two stimuli, but the response latency is different. **G.-I.** Relative variance explained, as quantified by R^2, when fitting basis functions (green trace) and a linear dynamical system (red trace) to data generated from the single-cell model (**G.**), from the recurrent network model (**H.**), and to the AC calcium activity data (**I.**), as a function of the number of stimuli. 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. In panels **G.** and **H.** $N = 500$, the number of stimuli is set to 10. In **G.-I.** 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 *Methods*). The basis function fit is performed using 10 Gaussian basis functions. In **H.** the linear dynamical system does not explain 100% of the variance of the data generated by the recurrent model due to noise in the input and subsampling of units.

Discussion

Adopting a population approach, we showed that strong OFF responses observed in auditory cortical neurons form transient 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 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.

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. S9 A-E). 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 between 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 reconcile 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 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 and 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 asymmetric part of the connectivity to isolate rotational dynamics, in our analysis we instead pointed out the role of the symmetric part of the connectivity in generating strongly amplified transient activity. Moreover we focused on the structure of low-dimensional dynamics across stimuli.

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). 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.

Methods

Contents

Data analysis

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. We analysed the trial-averaged activity of 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 have 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 1.

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

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 test the obtained decoder on the remaining trial. The classification accuracy is 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:

$$\mathbf{w}_t = \mathbf{c}_{1t} - \mathbf{c}_{2t}, \qquad \mathbf{b}_t = \frac{\mathbf{c}_{1t} + \mathbf{c}_{2t}}{2} \tag{4}$$

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

$$\begin{cases} \text{if } y(\mathbf{x}) > 0 \text{ then } \mathbf{x} \text{ is classified as stimulus } s_1 \\ \text{if } y(\mathbf{x}) < 0 \text{ then } \mathbf{x} \text{ is classified as stimulus } s_2 \end{cases} \tag{5}$$

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.

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. $\mathbf{X}$ contains 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:

$$\mathsf{VAR}(i) = \lambda_i \Big/ \sum_{j=1}^{N} \lambda_j \tag{6}$$

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

$$\mathsf{CUMVAR}(M) = \sum_{j=1}^{M} \lambda_j \Big/ \sum_{j=1}^{N} \lambda_j \tag{7}$$

In Fig. 1A, for each stimulus s we consider the matrix $\mathbf{X}^{(s)} \in \mathbb{R}^{N \times T}$ containing the populaition 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.

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)

$$\cos\theta_P = \max_{\mathbf{u}\in U, \mathbf{v}\in V} \mathbf{u}^T\mathbf{v} \tag{8}$$

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 organize them in a $N \times K$ matrix $\mathbf{Q}(s_1)$. We repeated this for stimulus s_2, which yields 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):

$$\cos\theta_P(s_1, s_2) = \sigma_1\Big(\mathbf{Q}(s_1)^T\mathbf{Q}(s_2)\Big) \tag{9}$$

i.e. 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).

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

$$\dot{\mathbf{X}} = \frac{\mathbf{X}(t_{i+1}) - \mathbf{X}(t_i)}{t_{i+1} - t_i}. \tag{10}$$

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:

$$(\mathbf{J} - \mathbf{I})^*_\lambda = \underset{\mathbf{J}-\mathbf{I}}{\operatorname{argmin}} ||\dot{\mathbf{X}}_\lambda - \mathbf{X}_\lambda(\mathbf{J} - \mathbf{I})||^2 \tag{11}$$

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):

$$(\mathbf{J} - \mathbf{I})^*_{r,\lambda} = \underset{\operatorname{rank}\mathbf{J} \leq r}{\operatorname{argmin}} ||\dot{\mathbf{X}}_\lambda - \mathbf{X}_\lambda(\mathbf{J} - \mathbf{I})||^2 \tag{12}$$

To solve Eq. (12) we used the solution to Eq. (11) given by:

$$\mathbf{J}^*_\lambda = \mathbf{I} + (\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T\dot{\mathbf{X}}_\lambda. \tag{13}$$

Supposing that the matrix $\mathbf{X}_\lambda\mathbf{J}^*_\lambda$ has singular value decomposition given by $\mathbf{X}_\lambda\mathbf{J}^*_\lambda = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, then it can be shown that the solution to the reduced rank ridge regression problem Eq. (12) can be written as:

$$\mathbf{J}^*_{r,\lambda} = \mathbf{J}^*_\lambda \sum_{i=1}^{r} \mathbf{V}_i\mathbf{V}_i^T \tag{14}$$

We note that each term in the sum of Eq. (14) has unit rank, so that the resulting matrix $\mathbf{J}^*_{r,\lambda}$ has rank equal to r.

Selection of hyperparameters To find the hyperparameters r and λ we fitted the linear dynamical system to the data with hyperparameters (r, λ) and computed the goodness of the fit $R^2(r, \lambda)$ using cross-validation. We repeat the process for a range of values of (r, λ). We observe that, independently of the value of λ, the function $R^2(r, \cdot)$ saturates at a particular value of the rank r^*, but it does not exhibit a clear maximum. We took the value r^* as the rank hyperparameter, while we defined the best ridge parameter as $\lambda^* = \operatorname{argmax}_\lambda R^2(\lambda, r^*)$ (Fig. S1).

We used K-fold cross-validation. 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.

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

$$\dot{\mathbf{Y}} = \mathbf{Y}(\mathbf{M} - \mathbf{I}), \tag{15}$$

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:

$$\mathbf{X}_{:,0:K} = \mathbf{Y}\mathbf{Q}_{:,0:K} \tag{16}$$

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:

$$\dot{\mathbf{X}}_{:,0:K} = \mathbf{X}_{:,0:K}(\mathbf{J} - \mathbf{I}) \tag{17}$$

In the following, we derive the relation between the connectivity matrix $\mathbf{J}$ resulting from fitting Eq. (17) and the full connectivity $\mathbf{M}$. We first find the expression of the velocity of the population trajectory given by $\mathbf{X}$. Using Eq. (16) and Eq. (15) and the orthogonality of $\mathbf{Q}$, we write:

$$\dot{\mathbf{X}} = \mathbf{X}\mathbf{Q}^T(\mathbf{M} - \mathbf{I})\mathbf{Q} \tag{18}$$

As in Eq. (17), we want to express $\dot{\mathbf{X}}_{:,0:K}$ as a function of $\mathbf{X}_{:,0:K}$. Using Eq. (18) we obtain:

$$\dot{\mathbf{X}}_{:,0:K} = \mathbf{X}_{:,0:K}\,(\mathbf{Q}^T_{0:K,:}\,\mathbf{M}\,\mathbf{Q}_{:,0:K} - \mathbf{I}) + \boldsymbol{\epsilon} \tag{19}$$

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. (17)) corresponds to the projection of the original connectivity $\mathbf{M}$ onto the first K principal components (Fig. S5):

$$\mathbf{J} \simeq \mathbf{J}_{\mathrm{PC}} = \mathbf{Q}^T_{0:K,:}\,\mathbf{M}\,\mathbf{Q}_{:,0:K} \tag{20}$$

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 its 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:

$$\sum_{k=1}^{K}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tkc} = 0, \quad \sum_{t=1}^{T}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tkc} = 0, \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\bar{\mathbf{Z}}_{tkc} = 0 \tag{21}$$

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

$$\begin{cases} \mathbf{\Sigma}^T_{ij} = \displaystyle\sum_{k=1}^{K}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{ikc}\bar{\mathbf{Z}}_{jkc} \\ \mathbf{\Sigma}^K_{ij} = \displaystyle\sum_{t=1}^{T}\sum_{c=1}^{C}\bar{\mathbf{Z}}_{tic}\bar{\mathbf{Z}}_{tjc} \\ \mathbf{\Sigma}^C_{ij} = \displaystyle\sum_{t=1}^{T}\sum_{k=1}^{K}\bar{\mathbf{Z}}_{tki}\bar{\mathbf{Z}}_{tkj} \end{cases} \tag{22}$$

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 $\boldsymbol{\mu}$, $\mathbf{\Lambda}^T$, $\mathbf{\Lambda}^K$ and $\mathbf{\Lambda}^C$ 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

$$P(\mathbf{S}) = \underset{y(\mathbf{S})}{\mathrm{argmax}}\Big[-\int_{\mathbf{S}} y(\mathbf{S})\log y(\mathbf{S})\mathrm{d}\mathbf{S}\Big], \qquad \text{with} \int_{\mathbf{S}} P(\mathbf{S})\mathrm{d}\mathbf{S} = 1 \tag{23}$$

subject to the constraints

$$\mathbb{E}_P[\boldsymbol{\mu}] = \mathbf{M}, \quad \mathbb{E}_P[\Lambda^T] = \mathbf{\Sigma}^T, \quad \mathbb{E}_P[\Lambda^K] = \mathbf{\Sigma}^K, \quad \mathbb{E}_P[\Lambda^C] = \mathbf{\Sigma}^C, \tag{24}$$

where $\mathbb{E}_P[\cdot]$ denotes the expectation over the probability density P. We use three types of surrogate datasets, denoted as T, TK and TKC. All the three types of surrogates obey to the first constraint in Eq. (24) 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.

Controls on simulated data We tested the controls generated as described above on simulated OFF responses generated by the recurrent model (Eq. (2)) and by the single-cell model (Eq. (3); Fig. S4). For the recurrent model the connectivity had a low-rank component and an unstructured component: $\mathbf{J} = \sum_{i=1}^{P} \mathbf{u}^{(i)}\mathbf{v}^{(i)T} + g\boldsymbol{\chi}$. Multiple OFF responses were generated by simulating the dynamics in response to initial conditions along the vectors $\mathbf{v}^{(i)}$'s. An equal number of OFF responses was generated using the single-cell model Eq. (3). For both datasets we then generated control datasets following the tensor maximum entropy method. Fitting a dynamical system to the simulated OFF responses and to the controls yielded significantly different goodness of fit when OFF responses were generated through the recurrent model. For OFF responses generated by the single-cell model this difference was significant only when a large number of principal components was kept, but was nonetheless small compared to the difference found for the recurrent model. The same differences in the fit of recurrent and single-cell OFF responses were observed even when the fitting was performed on smaller sets of units (50% and 20% of the units; new control datasets were generated for each subsampling of units). Thus, these results confirm that the tensor maximum entropy method provides suitable controls that allow us to identify dynamical structure in population responses datasets.

Analysis of the transient channels

Let $\mathbf{J}$ be the connectivity matrix resulting from fitting the responses to all stimuli at once, and $\mathbf{J}^{(s)}$ the connectivity obtained from fitting the response to stimulus s only. In the fitting procedure, we imposed a different value of the rank for the two matrices (for the plots in Fig. 4 we set $r = 80$ for $\mathbf{J}$ and $r = 8$ for $\mathbf{J}^{(s)}$). Using the singular value decomposition, we can write the matrices $\mathbf{J}^{(s)}$ as:

$$\mathbf{J}^{(s)} = \mathbf{L}^{(s)}\boldsymbol{\Sigma}^{(s)}\mathbf{R}^{(s)T}, \tag{25}$$

where $\mathbf{L}^{(s)}$ and $\mathbf{R}^{(s)}$ containing respectively the r left and right singular vectors as columns, while $\boldsymbol{\Sigma}^{(s)}$ has the singular values of $\mathbf{J}^{(s)}$ in the diagonal. The left and right singular vectors for stimulus s play the same role as the vectors $\mathbf{u}^{(s)}$ (normalized) and $\mathbf{v}^{(s)}$ in the model connectivity defined by Eq. (2). If the connectivity consists of unit-rank transient channels as in Eq. (2), the transient dynamics elicited by stimulus s has a strong component along a single dimension specified by $\mathbf{u}^{(s)}$. If instead the connectivity consists of rank-r transient channels of the form given by Eq. (25), then the transient dynamics elicited by stimulus s evolves in the subspace spanned by the column vectors of $\mathbf{L}^{(s)}$. We therefore define the overlap between the transient channels corresponding to pairs of stimuli showed in Fig. 4F as the principal angle between the subspaces specified by the respective $\mathbf{L}^{(s)}$ (see Eqs. (8)-(9)).

In analogy with equation Eq. (2), we define the matrix $\mathbf{J}_{\text{Sum}}$ as the sum of the individual transient channels for all stimuli:

$$\mathbf{J}_{\text{Sum}} = \mathbf{L}^{(1)}\boldsymbol{\Sigma}^{(1)}\mathbf{R}^{(1)T} + \mathbf{L}^{(P)}\boldsymbol{\Sigma}^{(P)}\mathbf{R}^{(P)T} + ... + \mathbf{L}^{(P)}\boldsymbol{\Sigma}^{(P)}\mathbf{R}^{(P)T} \tag{26}$$

We then compare goodness of the fit of the population OFF responses using the matrix $\mathbf{J}$ and the matrix $\mathbf{J}_{\text{Sum}}$.

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:

$$\tau\dot{r}_i = -r_i + \sum_{j=1}^{N} J_{ij}r_j, \tag{27}$$

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 $\mathbf{r}_0$ is therefore the OFF response to the stimulus associated with $\mathbf{r}_0$. We assume that during the OFF response the network receives no external input.

Normal and non-normal connectivity matrices

We define strong 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 a strong OFF response to the stimulus associated with the initial condition $\mathbf{r}_0$ when the value of $||\mathbf{r}(t)||$ 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 transiently increasing temporal dynamics. Importantly, the transient behaviour of $||\mathbf{r}(t)||$ depends on the stimulus, i.e. $\mathbf{r}_0$, 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 strong 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 strong OFF responses. However, the non-normality of the network constitutes only a necessary but not a sufficient condition for the generation of strong OFF responses.

Criterion for strong OFF responses

To find the necessary and sufficient condition for the generation of strong OFF responses in recurrent networks, we write the differential equation governing the dynamics of the distance from baseline as (Neubert and Caswell, 1997; Bondanelli and Ostojic, 2018):

$$\frac{1}{||\mathbf{r}||}\frac{\mathrm{d}||\mathbf{r}||}{\mathrm{d}t} = \frac{\mathbf{r}^T(\mathbf{J}_S - \mathbf{I})\mathbf{r}}{||\mathbf{r}||^2}, \qquad \mathbf{J}_S = \frac{\mathbf{J} + \mathbf{J}^T}{2}, \tag{28}$$

where we denote by $\mathbf{J}_S$ the symmetric part of the connectivity $\mathbf{J}$. The linear recurrent network exhibits strong OFF 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. (28) 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. (28) it is possible to show that the necessary and sufficient condition for the generation of strong OFF responses in recurrent networks is given by

$$\lambda_{\max}(\mathbf{J}_S) > 1 \tag{29}$$

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

Dynamics in a rank-1 network

We consider a network with unit-rank connectivity $\mathbf{J} = \mathbf{u}^{(1)}\mathbf{v}^{(1)T}$. For simplicity, we take the norm of the vectors $\mathbf{v}^{(i)}$ to be unitary ($||\mathbf{v}^{(i)}|| = 1$), while the norm of the vectors $\mathbf{u}^{(i)}$ can vary. For these networks, the condition for the generation of strong OFF responses given by Eq. (29) can be determined by computing the eigenvalues of the symmetric part $\mathbf{J}_S = (\mathbf{u}^{(1)}\mathbf{v}^{(1)T} + \mathbf{v}^{(1)}\mathbf{u}^{(1)T})/2$. The matrix $\mathbf{J}_S$ has two non-trivial eigenvalues, given by

$$\lambda_{\max,\min}(\mathbf{J}_S) = ||\mathbf{u}^{(1)}||\frac{1 \pm \cos\theta}{2}, \tag{30}$$

where θ is the angle between the two vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)T}$. Therefore the condition for strong OFF responses is given in terms of the length of the vector $\mathbf{u}^{(1)}$ and reads:

$$||\mathbf{u}^{(1)}|| > \frac{2}{1 \pm \cos\theta} \tag{31}$$

Here we focus on the case where the two vectors are orthogonal to each other, i.e. $\cos\theta = 0$. When the number of units N is large, this can be achieved by randomly drawing the vector elements from a given probability distribution (e.g. Gaussian). In this case, the condition for strong OFF responses reduces to $||\mathbf{u}^{(1)}|| > 2$. In this regime, strong OFF responses are evoked when the initial condition is closely aligned to the vector $\mathbf{v}^{(1)}$. If the initial condition ($t = 0$) is aligned with $\mathbf{v}^{(1)}$, we can explicitly write the OFF response dynamics as:

$$\mathbf{r}(t) = e^{-t}\mathbf{v}^{(1)} + te^{-t}\mathbf{u}^{(1)} \tag{32}$$

From Eq. (32), we can directly notice that the OFF response $\mathbf{r}(t)$ evolves in the two-dimensional plane defined by the vectors $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$. At time $t = 0$ the population activity vector is aligned with the vector $\mathbf{v}^{(1)}$. Then, at the time when the distance from baseline is the largest, the component of the activity along the direction $\mathbf{v}^{(1)}$ has decayed exponentially, while the component of the activity along the direction given by $\mathbf{u}^{(1)}$ has increased proportionally to the value of $||\mathbf{u}^{(1)}||$. Eventually, the activity decays to the baseline level given by $\mathbf{r}_B = \mathbf{r}(t \to \infty) = 0$. Therefore, the population activity vector at the peak of the transient OFF responses, aligned with $\mathbf{u}^{(1)}$, is essentially orthogonal to the population activity vector at time $t = 0$, which is along $\mathbf{v}^{(1)}$.

Dynamics in a rank-P network

Here we examine the case where the connectivity consists of P orthogonal unit-rank transient channels, as given by Eq. (2). In this scenario, all the two-dimensional subspaces defined by the vectors $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$ are mutually orthogonal, meaning that a vector lying on the subspace defined by $\mathbf{u}^{(s_1)}$ and $\mathbf{v}^{(s_1)}$ has zero component on the subspace defined by $\mathbf{u}^{(s_2)}$ and $\mathbf{v}^{(s_2)}$. When the number of units $N \gg P$ is large, this can be achieved by randomly drawing the elements of all the vectors $\mathbf{u}^{(i)}$'s and $\mathbf{v}^{(i)}$'s from a given probability distribution (e.g. Gaussian). Assuming that all the vectors $\mathbf{u}^{(i)}$ has the same length equal to $||\mathbf{u}||$, the condition for having amplified OFF responses reduces to $||\mathbf{u}|| > 2$. If this condition is met, the network can produce amplified dynamics in responses to a number P of stimuli, corresponding to the P initial conditions given by the vectors $\mathbf{v}^{(i)}$. If the transient channels are mutually uncorrelated, the OFF response to stimulus s can be written as in Eq. (32):

$$\mathbf{r}^{(s)}(t) = e^{-t}\mathbf{v}^{(s)} + te^{-t}\mathbf{u}^{(s)} \tag{33}$$

The OFF responses to different stimuli follow therefore the same dynamics, but evolve on orthogonal subspaces.

Since the dynamics is linear, any correlation between pairs of initial condition is reflected in the correlation between the subspaces spanned by the evoked trajectories. We illustrate this point using our network model. Suppose we have a pair of initial conditions given by $\mathbf{v}^{(s_1)}$ and $\alpha\mathbf{v}^{(s_1)} + \beta\mathbf{v}^{(s_2)}$ (with $\alpha^2 + \beta^2 = 1$). Since the vectors $\mathbf{v}^{(s_1)}$ and $\mathbf{v}^{(s_2)}$ are orthogonal to each other, the correlation between these initial condition is equal to α. While the OFF response with initial condition $\mathbf{v}^{(s_1)}$ can be written as in Eq. (33), the OFF response to the second stimulus reads:

$$\mathbf{r}^{(s)}(t) = e^{-t}(\alpha\mathbf{v}^{(s_1)} + \beta\mathbf{v}^{(s_2)}) + te^{-t}(\alpha\mathbf{u}^{(s_1)} + \beta\mathbf{u}^{(s_2)}) \tag{34}$$

At the time of peak distance from baseline, activity vectors for the two OFF responses are given respectively by $\mathbf{u}^{(s_1)}$ and $\alpha\mathbf{u}^{(s_1)} + \beta\mathbf{u}^{(s_2)}$. Since $\mathbf{u}^{(s_1)}$ and $\mathbf{u}^{(s_2)}$ are orthogonal to each other, their correlation is equal to α. Therefore, the correlation between the initial conditions is preserved in the correlation between the states explore during the evoked transient dynamics.

Principal component analysis of low-rank dynamics

Rank-1 connectivity Here we show the link between the vectors $\mathbf{u}$ and $\mathbf{v}$ defining a single transient channel and the principal components of the OFF response with initial condition $\mathbf{v}$ for a unit-rank connectivity given by $\mathbf{J} = \mathbf{u}\mathbf{v}^T$. We examine the case in which the two vectors are orthogonal to each other, i.e. $\cos\theta = 0$. In this case, we can write the OFF response in Eq. (32) as:

$$\mathbf{r}(t) = f_1(t)\mathbf{v} + f_2(t)\mathbf{u}, \tag{35}$$

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:

$$\mathbf{C} = \int_0^{\tau_0} \big(\mathbf{r}(t) - \boldsymbol{\mu}\big)\big(\mathbf{r}(t) - \boldsymbol{\mu}\big)^T \, \mathrm{d}t, \qquad \boldsymbol{\mu} = \frac{1}{\tau_0}\int_0^{\tau_0} \mathbf{r}(t)\, \mathrm{d}t \tag{36}$$

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. (35) we can re-write the covariance matrix as:

$$\mathbf{C} = \alpha\, \mathbf{v}\mathbf{v}^T + \beta\, (\mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T) + \gamma\, \mathbf{u}\mathbf{u}^T \tag{37}$$

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, $||\mathbf{u}|| \ll 1$, and strongly amplified dynamics $||\mathbf{u}|| \gg 1$. In the case of weakly amplified dynamics we can write the covariance matrix as:

$$\mathbf{C} = \alpha\, \mathbf{v}\mathbf{v}^T + O(||\mathbf{u}||/\sqrt{N}) \tag{38}$$

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

$$\mathbf{C} = \gamma\, ||\mathbf{u}||^2 \hat{\mathbf{u}}\hat{\mathbf{u}}^T + O(||\mathbf{u}||/\sqrt{N}) \tag{39}$$

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

Rank-P connectivity Here we show the relationship between the vectors $\mathbf{u}^{(i)}$, $i = 1, ..., P$ and the first principal components of the joint dynamics with initial conditions $\mathbf{v}^{(i)}$, $i = 1, ..., P$ when the connectivity matrix is of the form given by Eq. (2): $\mathbf{J} = \sum_{i=1}^{P} \mathbf{u}^{(i)}\mathbf{v}^{(i)T}$. For simplicity we assume that the number of transient channels is smaller than the number of units, $P \ll N$. Under this condition the dynamics with initial condition $\mathbf{v}^{(i)}$ can be written as (Bondanelli and Ostojic, 2018):

$$\mathbf{r}^{(i)}(t) = f_1(t)\mathbf{v}^{(i)} + f_2(t)\mathbf{u}^{(i)} \tag{40}$$

Therefore the correlation matrix of the joint dynamics reads:

$$\mathbf{C} = \sum_{i=1}^{P} \int_0^{\tau_0} \big(\mathbf{r}^{(i)}(t) - \boldsymbol{\mu}\big)\big(\mathbf{r}^{(i)}(t) - \boldsymbol{\mu}\big)^T \, \mathrm{d}t, \qquad \boldsymbol{\mu} = \frac{1}{\tau_0 P}\sum_{i=1}^{P} \int_0^{\tau_0} \mathbf{r}^{(i)}(t)\, \mathrm{d}t \tag{41}$$

We examine the case of strongly amplified dynamics, corresponding to $||\mathbf{u}^{(i)}|| = ||\mathbf{u}|| \gg 1$, $i = 1, ..., P$ (the case of weakly amplified dynamics $||\mathbf{u}|| \ll 1$ is analogous). If $||\mathbf{u}^{(i)}|| \gg 1$ for all i, the correlation matrix $\mathbf{C}$ is dominated by quadratic terms in $\{\mathbf{u}^{(i)}\}$:

$$\mathbf{C} = \alpha \sum_{i=1}^{P} \mathbf{u}^{(i)}\mathbf{u}^{(i)T} - \frac{1}{P}\beta \sum_{i,j=1}^{P} \mathbf{u}^{(i)}\mathbf{u}^{(j)T} + O(||\mathbf{u}||/\sqrt{N}) \tag{42}$$

where $\alpha = \int f_2^2$ and $\beta = \tau_0^{-1}(\int f_2)^2$. If the number of transient channels P is large (but still much smaller than the number of units N), the vectors $\mathbf{u}^{(i)}$'s are strongly correlated with the eigenvectors of $\mathbf{C}$. In fact, multiplying the correlation matrix by $\hat{\mathbf{u}}^{(l)} = \mathbf{u}^{(l)}/||\mathbf{u}^{(l)}||$ yields:

$$\mathbf{C}\hat{\mathbf{u}}^{(l)} = \alpha||\mathbf{u}||^2\hat{\mathbf{u}}^{(l)} - \frac{1}{P}\beta||\mathbf{u}||^2 \sum_{i=1}^{P} \hat{\mathbf{u}}^{(i)} + O(||\mathbf{u}||/\sqrt{N}) \tag{43}$$

While the first term on the right hand side of Eq. (43) is of order $1/\sqrt{N}$, the second term is of order $1/\sqrt{PN}$. This implies that when the number of transient channels P is large but smaller than N ($1 \ll P \ll N$), the vectors $\mathbf{u}^{(i)}$'s are strongly correlated with the top eigenvector of $\mathbf{C}$, corresponding to the first principal components of the joint dynamics $\mathbf{r}^{(i)}(t)$, $i = 1, ..., P$.

Single-cell model for OFF response generation

The model

The temporal filters $L_i(t)$ in Eq. (3) are chosen to be simple exponential functions (for illustration see Fig. 5A):

$$L_i(t) = (1 + \alpha_i t)\, e^{-\alpha_i t/\tau}. \tag{44}$$

Note that $L_i(0) = 1$, so that the modulation factors $r_{0,i}^{(s)}$ can be interpreted as the initial condition of the response of neuron i to stimulus s. The parameters α_i control both the peak latency and the effective timescale of the response of neuron i. In fact, we have that the effective timescale is given by $\tau_{\text{eff,i}} = \tau/\alpha_i$, while the peak latency is $t_{L,i} = (\tau - 1)/\alpha_i$. Neurons with bigger response latencies have also longer decay times. The functions $L_i(t)$ have a single peak. The peak amplitude does not depend on α_i and it is given by $A(\tau) = \tau \exp(-1 + 1/\tau)$. Note that the timescale of the distance from baseline $||\mathbf{r}(t)||$ is controlled by the neuron with the largest latency, i.e. $||\mathbf{r}(t)|| \sim \exp(-\min_i(\alpha_i)t/\tau)$.

Principal component analysis of the single-cell model

Here we show that in the case where all the neurons have the same response latency $\alpha_i = \alpha$, the population response evolves along only one dimension. Using Eq. (3) we can write the covariance matrix goven by Eq. (36) as:

$$\mathbf{C}_{ij} = r_{0,i} r_{0,j} \xi_{ij}, \quad \xi_{ij} = \int_0^{\tau_0} L_i L_j \,\mathrm{d}t - \tfrac{1}{\tau_0} \int_0^{\tau_0} L_i \,\mathrm{d}t \int_0^{\tau_0} L_j \,\mathrm{d}t \tag{45}$$

If the response latency is the same for every neuron, then $\xi_{ij} = \xi$ for all i, j and $\mathbf{C} = \xi \mathbf{r}_0 \mathbf{r}_0^T$. In this case, the covariance matrix has only one non-trivial eigenvector equal to $\mathbf{r}_0$, which corresponds to the dimension on which the population OFF response evolves.

Fitting 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 coefficient $a_{ij}^{(s)}$ that best approximate the equation

$$r_i^{(s)}(t) = \sum_{j=1}^{N_{\text{basis}}} a_{ij}^{(s)} f_j(t), \tag{46}$$

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. (46) by the initial condition $r_{0,i}^{(s)}$ we obtain:

$$\frac{r_i^{(s)}(t)}{r_{0,i}^{(s)}} = \sum_{j=1}^{N_{\text{basis}}} b_{ij}^{(s)} f_j(t), \quad b_{ij}^{(s)} = a_{ij}^{(s)}/r_{0,i}^{(s)}. \tag{47}$$

In general the coefficients $b_{ij}^{(s)}$ could be fitted independently for each stimulus. However, our single-cell model Eq. (3) assumes that the coefficients b_{ij} do not change across stimuli. Under this assumption, Eq. (46) can be written as:

$$L_i(t) = \sum_{j=1}^{N_{\text{basis}}} b_{ij} f_j(t) \tag{48}$$

If the coefficients b_{ij} are independent on the stimulus s, the problem of fitting Eq. (46) can be reduced to a linear regression problem. 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 $N_{\text{basis}} \times TC$ obtained by concatenating C times the $N_{\text{basis}} \times T$ matrix $(\mathbf{f})_{it} = f_i(t)$. Let $\mathbf{B}$ be the $N \times N_{\text{basis}}$ given by $\mathbf{B}_{ij} = b_{ij}$. Then Eq. (47) can be written as:

$$\mathbf{R}^{(C)} = \mathbf{B}\mathbf{F}^{(C)}, \tag{49}$$

which can be solved using linear regression techniques.

If population responses are generated through Eq. (3), then Eq. (46) implies that the performance of the fit Eq. (49) is independent of the number C of stimuli considered. In the case of OFF responses generated through a recurrent network mechanisms, Eq. (46) does not hold true, and the performance of the fit decreases as the number of stimuli C is increased. In Fig. 5**D.-F.** instead of normalizing the responses by their initial condition, we normalized them by the firing rate range (up to a sign), i.e. we choose $r_{0,i} = \pm(\max_t r_i(t) - \min_t r_i(t))$, where the sign corresponds to $\mathrm{sgn}(r_i(t^*))$, with $t^* = \mathrm{argmax}_t |r_i(t)|$. Note that the two normalizations are equivalent for the single cell model Eq. (3). 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.

Supplementary figures

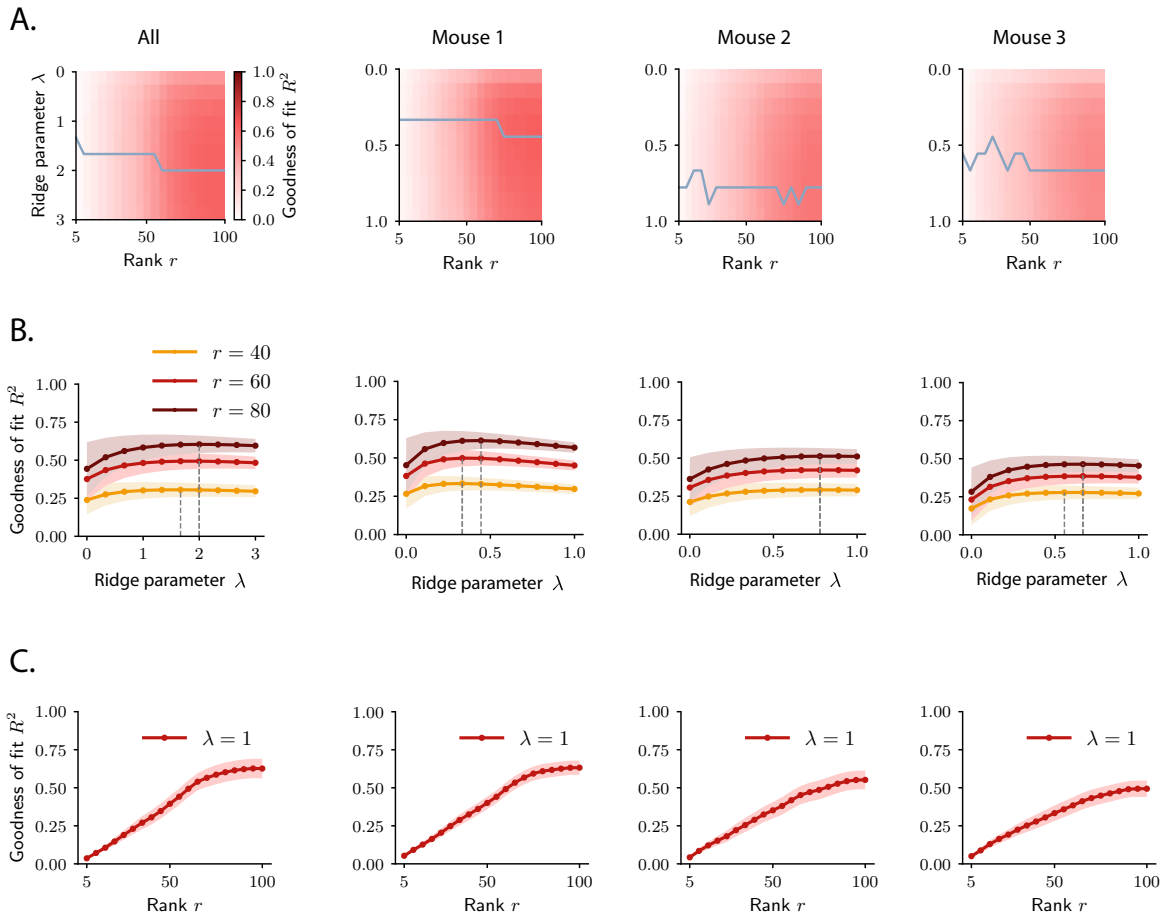

Figure S1: **Selection of hyperparameters in rank-reduced ridge regression.** The ridge and rank hyperparameters λ and r are selected using 10-fold cross-validation in the time domain, when fitting the activity of the whole pseudopopulation (All) or the activity of individual animals. **A.** Goodness 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^*.

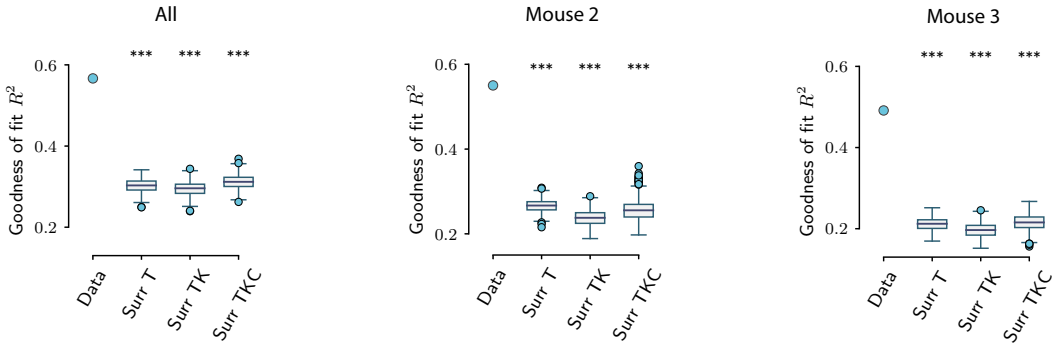

Figure S2: **Goodness of fit for the original and surrogate datasets.** Value of R^2 for the original and 500 surrogate datasets computed using the activity of the whole pseudopopulation (All) and the activity of individual animals (mouse 1 is shown in Fig. 4D). The parameters for the reduced-rank ridge fitting for the three plots are respectively: $\lambda_{\text{All}} = 2$, $r_{\text{All}} = 70$; $\lambda_2 = 0.8$, $r_2 = 90$; $\lambda_3 = 0.6$, $r = 90$.

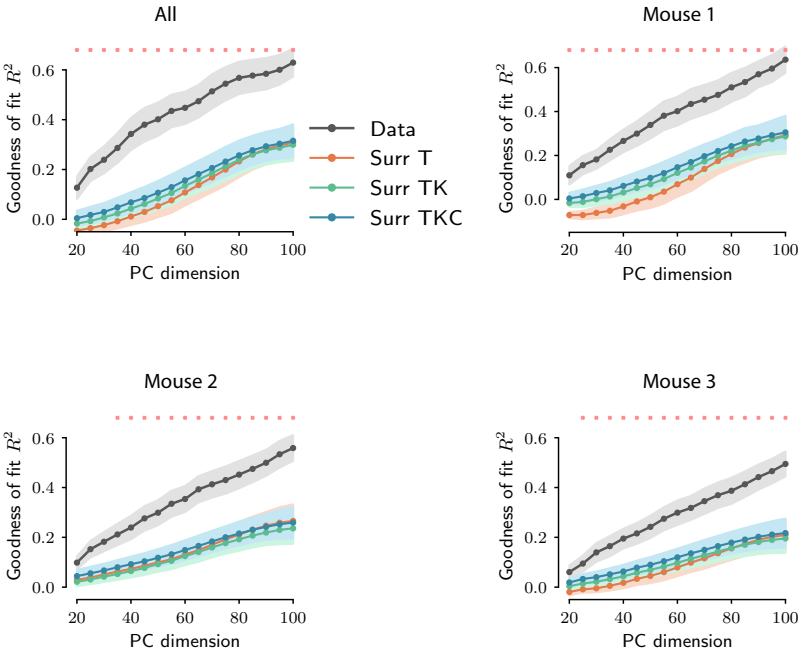

Figure S3: **Goodness of fit as a function of PC dimensionality for original and surrogate datasets.** Value of R^2 as a function of PC dimensionality for the original dataset (black trace) and for the surrogate datasets (colored traces). The value of R^2 was computed using ridge regression using 10-fold 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 (at least $P < 0.01$; upper tail test using the mean value of R^2 over the cross-validations folds for each dataset).

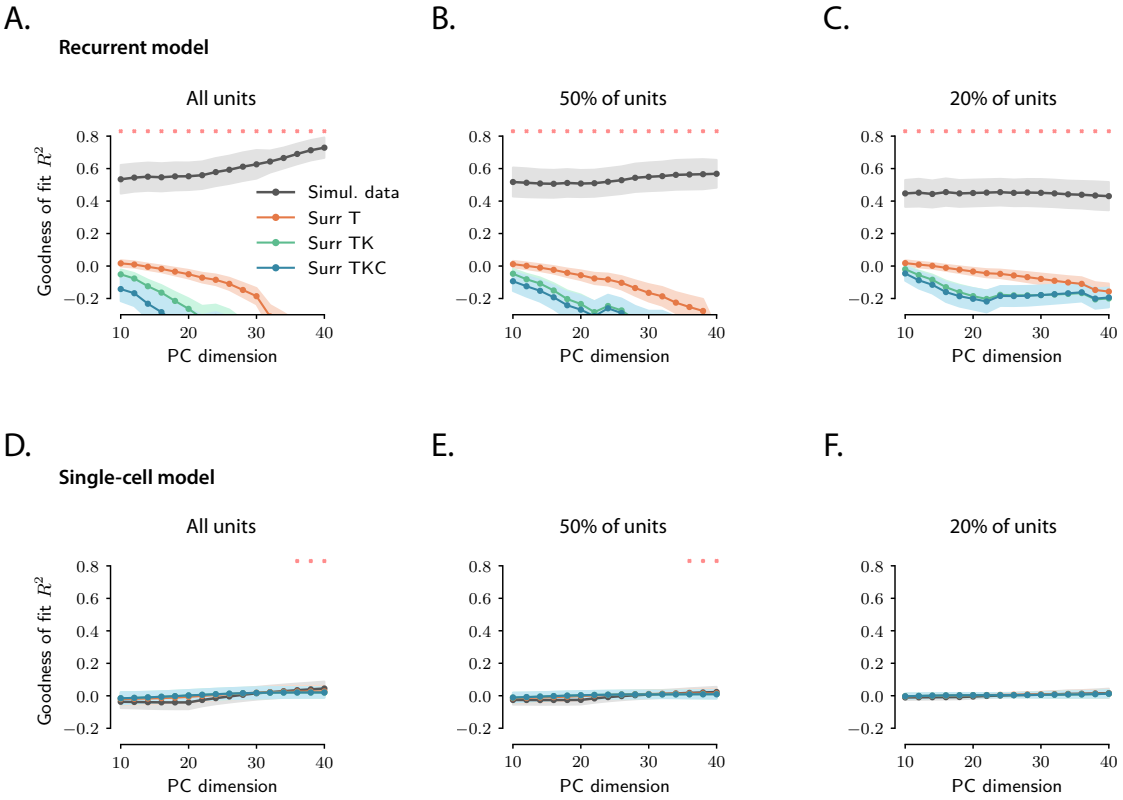

Figure S4: **Hypothesis testing for simulated OFF responses.** We fit a linear dynamical system to the OFF responses generated through a recurrent network mechanism with low-rank connectivity (Eq. (2), **A.-C.**) or through the single-cell mechanism (Eq. (3), **D.-F.**) The goodness of the fit R^2 for the original simulated data is then compared with the three types of surrogate datasets. For the recurrent model, we find that there is a substantial difference between the value of R^2 for the original datasets and for the surrogates, and this difference is statistically significant over the whole range of PC dimensionality considered. This feature is preserved even when the fitting is performed on a smaller subset of units (subsampling 50% and 20% of the units). In **A.-C.** we used a recurrent connectivity $\mathbf{J} = \mathbf{J}_{\text{low-rank}} + g\boldsymbol{\chi}$, given by the sum of a low-rank connectivity (with rank 30) and a Gaussian connectivity with mean zero and standard deviation $g/\sqrt{N}$ (with $g = 0.2$), which acts as connectivity noise (Mastrogiuseppe and Ostojic, 2018; Bondanelli and Ostojic, 2018). 20 OFF responses were generated by setting the state before stimulus offset along the first 20 amplified initial conditions. Here $N = 500$. Input noise was included in the simulation. **D.-F.** When OFF responses are generated through the single-cell mechanism, fitting a dynamical system yields very poor performance. The difference in the values of R^2 between the original and surrogate datasets is relatively small and it becomes significant only when a large number of principal components is considered. We simulated 20 OFF responses generated by the single-cell model using 500 units. In all panels, we used ridge regression and 10-fold cross validation. The ridge parameter is optimized for each choice of the PC dimensionality. For the simulated 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 (computed as $\langle \text{SEM} \rangle = \sqrt{\sum_{i=1}^{N_{\text{surr}}} \text{SEM}_i^2 / N_{\text{surr}}}$). 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 (at least $P < 0.01$; upper tail test using the mean value of R^2 over the cross-validations folds for each dataset).

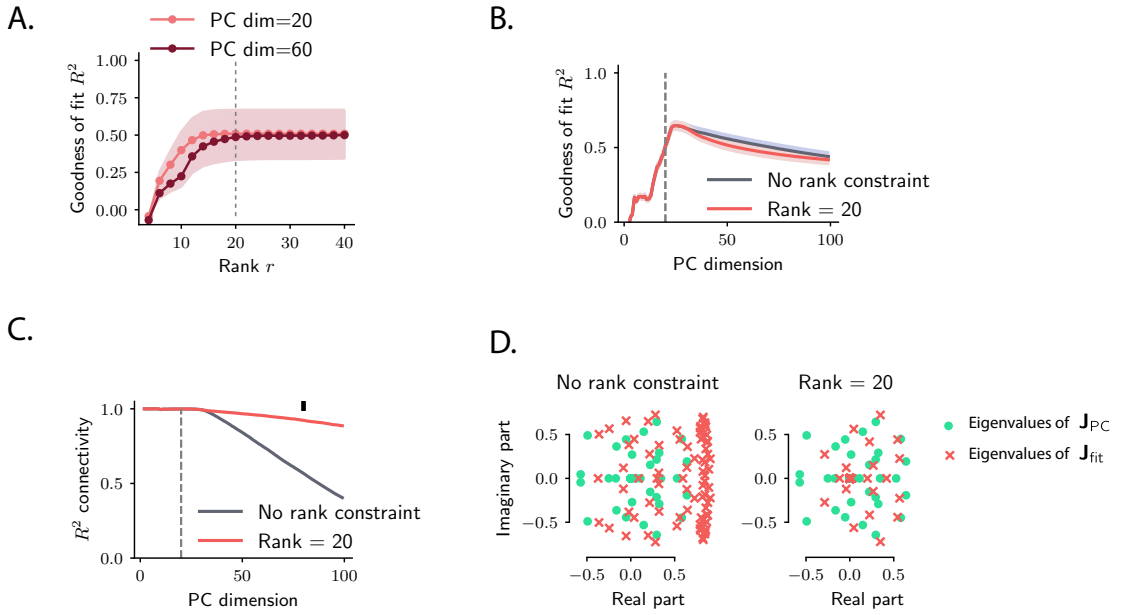

Figure S5: **Model recovery for simulated OFF responses using a low-rank network model.** Model recovery for OFF responses generated through a low-rank network mechanism. The number of units was set to $N = 500$ and the rank of the connectivity was fixed to 30. We generated 10 OFF responses by setting the initial state of the network to the first 10 amplified initial conditions. Input noise was injected into the system. Thus, the dynamics across stimuli spans 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 Eq. (20) in *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 $\lambda = 0.1$.

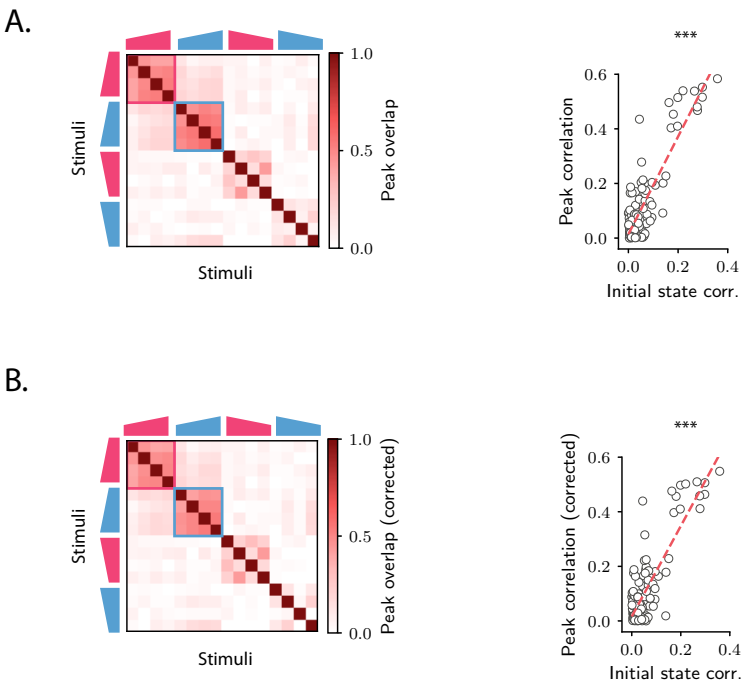

Figure S6: **Overlap between the state at the peak of the transient OFF responses. A.** *Left*: overlap between the states at the peak of the OFF responses for each pair of stimuli. The peak time is defined as the time at which the distance from baseline of the population vector is maximum. *Right*: 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.

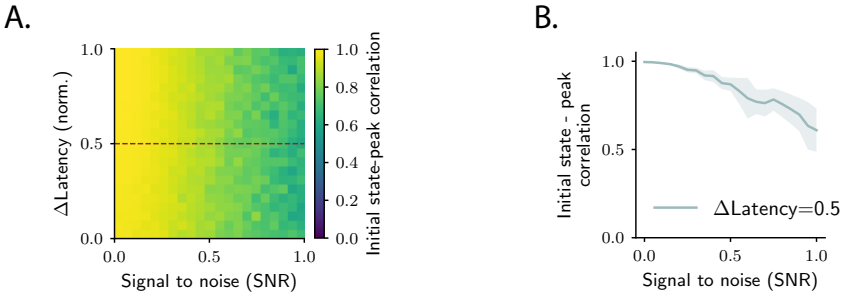

Figure S7: **Correlation between initial state and peak state for the single-cell model as a function of latency span and signal-to-noise ratio. A.** Average correlation between the state at the end of stimulus presentation and the state at the peak of the OFF responses as a function of the latency span, defined as the difference between the maximum and the minimum value of the latency across all units (ΔLatency(norm.) = (Latency$_{\max}$ − Latency$_{\min}$)/Latency$_{\max}$), and the signal-to-noise ratio (SNR), defined as the ratio between the standard deviation of the noise and the peak amplitude of the temporal responses $L_i(t)$. Average is computed over 1000 bootstrap subsamplings of 10% of the units in the population. **B.** Correlation between the state at the end of stimulus presentation and the state at the peak of the OFF responses as a function of the SNR for ΔLatency(norm.) = 0.5 (red dashed line in **A.**). Error bars represent the standard deviation computed over 1000 bootstrap subsamplings of 10% of the units in the population. Here $N = 500$, $\tau = 10$ and $\alpha_{\min} = 1$.

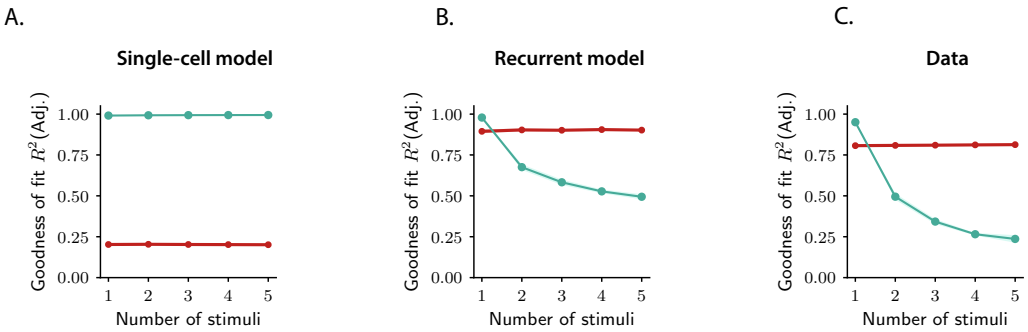

Figure S8: **Fit of population OFF responses using the recurrent and single-cell model adjusted for the number of parameters. A.-C.** Relative variance explained, as quantified by the adjusted index R^2_{Adj}, when fitting basis functions (green trace) and a linear dynamical system (red trace) to data generated from the single-cell model (**A.**), from the recurrent network model (**B.**), and to the AC calcium activity data (**C.**), as a function of the number of stimuli. Error bars represent the standard deviation over multiple samples of the stimuli. In panels **A.** and **B.** $N = 500$, the number of stimuli is set to 10. 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 *Methods*). The basis function fit is performed using 10 Gaussian basis functions. In **B.** the linear dynamical system does not explain 100% of the variance of the data generated by the recurrent model due to noise in the input and subsampling of units.

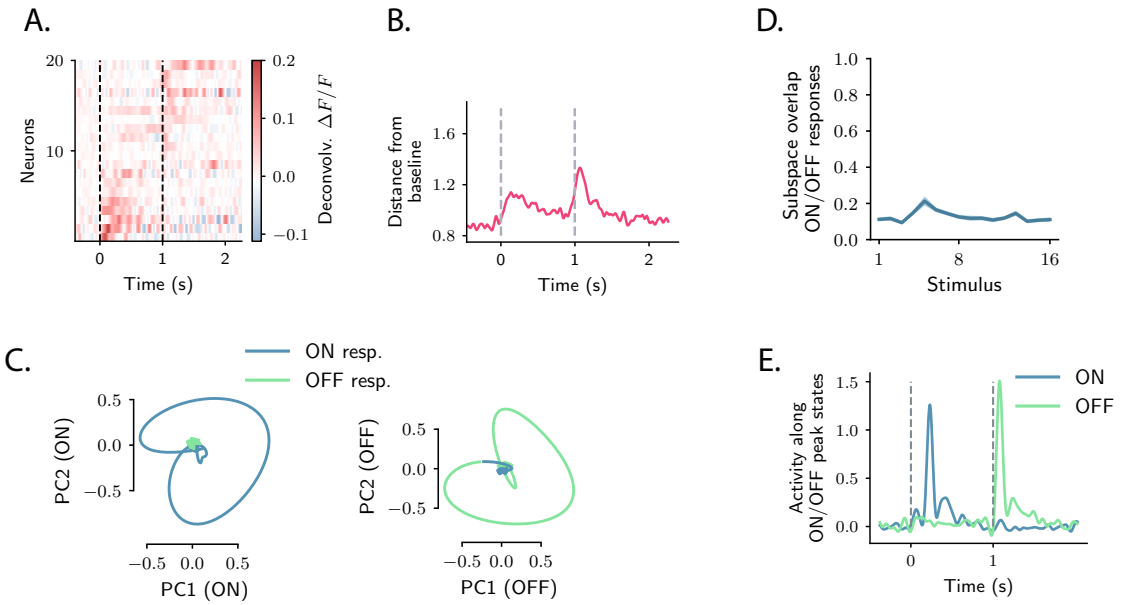

Figure S9: **Population ON and OFF responses in auditory cortical activity and in a nonlinear recurrent network model. A.** Trial-averaged deconvolved calcium signals showing the activity of 20 out of 2343 neurons in response to a 8 kHz 1s UP-ramp with intensity range 60-85 dB (see Fig. 1A). We selected neurons with high signal-to-noise ratios and ordered them according to the difference between peak activity during ON and OFF response epochs. **B.** Distance from baseline $||\mathbf{r}(t)||$ computed for the population response to the 8 kHz sound in **A** (see Fig. 1A). **C.** *Left*: projection of the population ON and OFF responses to the stimulus in **A.** on the first two principal components computed for the ON response. *Right*: projection of both ON and OFF responses on the first two principal components computed for the OFF response. PCA was performed on the period from -50 ms to 350 ms with respect to stimulus onset and offset. **D.** Subspace overlap between the subspaces spanned by the ON and OFF population responses to individual stimuli. The subspace overlap is computed as the principal angle between the ON and OFF subspaces defined by the first 5 principal components of the ON and OFF population responses. Error bars correspond to ± 1 standard deviations computed over 100 subsamplings of the 80% of the units. **E.** Projection of the population response to the stimulus in **A.** along the population states at the peak (corresponding to maximum distance from baseline) of the ON and OFF responses.

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):E1927–E1935, 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.

S. Bagur, M. Averseng, D. Elgueda, S. David, J. Fritz, P. Yin, S. Shamma, Y. Boubenec, and S. Ostojic. Go/no-go task engagement enhances population representation of target stimuli in primary auditory cortex. *Nature Communications*, 9(1):2529, 2018.

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.

P. Bartho, C. Curto, A. Luczak, S. L. Marguet, and K. D. Harris. Population coding of tone stimuli in auditory cortex: dynamic rate vector analysis. *European Journal of Neuroscience*, 30(9):1767–1778, 2009.

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

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. *arXiv*, 2018.

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, 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.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008.

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. 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.

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.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

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*, 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.

D. Kobak, W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X.-L. Qi, R. Romo, N. Uchida, and C. K. Machens. Demixed principal component analysis of neural population data. *eLife*, 5: e10989, 2016.

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.

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.e7, 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.e29, 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.

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.

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.

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.

H. Sompolinsky and I. Kanter. Temporal association in asymmetric neural networks. *Physical Review Letters*, 57:2861–2864, 1986.

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.