# Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons

Ulisse Ferrari,[1,*] Stéphane Deny,[2] Matthew Chalk,[1] Gašper Tkačik,[3] Olivier Marre,[1,4] and Thierry Mora[5,4]

[1]*Sorbonne Université, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, 75012 Paris, France.*
[2]*Neural Dynamics and Computation Lab, Stanford University, California*
[3]*Institute of Science and Technology, Klosterneuburg, Austria*
[4]*Equal contribution*
[5]*Laboratoire de physique statistique, CNRS, Sorbonne Université, Université Paris-Diderot and École normale supérieure (PSL), 24, rue Lhomond, 75005 Paris, France*

Correlations in sensory neural networks have both extrinsic and intrinsic origins. Extrinsic or stimulus correlations arise from shared inputs to the network, and thus depend strongly on the stimulus ensemble. Intrinsic or noise correlations reflect biophysical mechanisms of interactions between neurons, which are expected to be robust to changes of the stimulus ensemble. Despite the importance of this distinction for understanding how sensory networks encode information collectively, no method exists to reliably separate intrinsic interactions from extrinsic correlations in neural activity data, limiting our ability to build predictive models of the network response. In this paper we introduce a general strategy to infer population models of interacting neurons that collectively encode stimulus information. The key to disentangling intrinsic from extrinsic correlations is to infer the couplings between neurons separately from the encoding model, and to combine the two using corrections calculated in a mean-field approximation. We demonstrate the effectiveness of this approach on retinal recordings. The same coupling network is inferred from responses to radically different stimulus ensembles, showing that these couplings indeed reflect stimulus-independent interactions between neurons. The inferred model predicts accurately the collective response of retinal ganglion cell populations as a function of the stimulus.

## I. INTRODUCTION

An important goal in sensory neuroscience is to build network models to understand how sensory stimuli are encoded by the collective activity of neuronal populations. Pioneering work initiated in the retina [1–4] proposed to use disordered Ising models to characterize the joint activity of neurons in early sensory systems [5–9], sensory cortices [10] and beyond [11–14]. Motivated by the principle of Maximum Entropy [15], these models represent neurons as binary spins (spike or silence), biased by local fields and interacting through a network of pairwise couplings. The model parameters are fitted to reproduce the empirical mean neuron activities and pairwise correlations between them.

A major limitation of this approach in the context of sensory systems is that the inferred couplings are only effective and do not directly follow network circuitry. Instead, they reflect two different sources of correlations. On one hand, two neurons can be correlated because they receive correlated or common inputs. For example, in the retina, if the stimulus is correlated over space, nearby neurons will receive similar inputs and consequently will respond synchronously. This type of correlation has been termed 'signal correlation' and strongly depends on the actual stimulus and its statistics. On the other hand, neurons might be correlated because of actual interactions in the neural network, either because they are connected directly through gap junctions [16, 17] or indirect pathways [18], or because they receive the same noise sources from photoreceptors [19]. These types of correlations have been termed 'noise correlations' and results from the physological wiring of the network [18]. Similar network effects are present in all sensory systems [20, 21]. When fitting Ising models, signal and noise correlations are mixed together and difficult to disentangle. Consequently, the inferred couplings do not only reflect properties of the network circuitry, but also incidental properties of the actual stimulus.

The inability to separate extrinsic from intrinsic correlations limits the interpretability of these models and their capacity to generalize across different conditions. For example, if two Ising models are trained on the neural responses to two different types of stimuli, their interactions terms will be different [5]. The model thus cannot generalize and fails to predict the collective behaviour in response to a different type of stimulus. Interpreting a change in the interaction terms will also be difficult: it could trivially reflect changes in the stimulus statistics, or could correspond to changes in the network of couplings, and the way the network processes stimuli (adaptation). Modeling the influence of the stimulus is therefore crucial to understand the collective processing performed by sensory networks. Models with stimulus-dependent fields and couplings have been proposed to describe stimulus correlations and network effects [7, 22, 23]. However, here we argue that the strategy proposed to fit these models from neural recordings does not ensure a proper disentanglement between these two sources of correlated activity. Thus, a general strategy to model and accu-

*Correspondence should be sent to `ulisse.ferrari@gmail.com`.

rately separate stimulus and noise correlations in neural networks is still lacking.

Here we propose a general method to achieve this task. We define a population encoding model where each neuron's spiking probability is governed by its couplings to other neurons, and by an external, time-dependent field encoding the effect of the stimulus. We describe a working strategy for learning the parameters of this model from neural recordings. First, we infer the coupling matrix from population responses to repetitions of short films. Second, we model how each neuron's firing rate depends on the stimulus, with no regard for noise correlations. Third, we use a mean field (Thoughless-Anderson-Palmer) approximation to calculate the value of the fluctuating field as a function the stimulus from those predicted rates, corrected by the influence of the network.

We apply our model to describe the responses of retinal neurons to a visual stimulus. We quantify the importance of noise correlations in this system, and we fit the corresponding coupling matrix from recordings of responses to repeated films. We then combine this coupling matrix with a previously proposed model of stimulus encoding fitted on single cells [24] to obtain a complete model that reproduces the population response. We show that this strategy can be used to obtain accurate predictions across different stimulus statistics. We have therefore found a way to design and train models of population responses that can generalize across different stimulus ensembles. The method can be applied to any system where a single-cell encoding model is available.

## II. A GENERAL POPULATION MODEL

### A. Model definition

We start by introducing a general model of the activity of a population of correlated neurons, labeled by $i = 1, \ldots, N$, in response to a stimulus. This probabilistic model accounts for both arbitrary single-cell dependences on the stimulus, and direct interactions between cells. Let us denote by $n_i$ the number of spikes emitted by cell $i$ during a short time bin. The probability distribution of spiking patterns $\mathbf{n} = (n_1, \ldots, n_N)$ in response to a time-dependent stimulus $S$ is given by:

$$P_{\text{pop}}(\mathbf{n}|t) = \frac{1}{Z} \exp\left[-H_1(\mathbf{n}, t) - H_2(\mathbf{n})\right], \qquad (1)$$

where $Z$ is a normalization constant and

$$H_1(\mathbf{n}, t) = -\sum_{i=1}^{N} \left[h_i(t) n_i - \gamma\, n_i^2 - \delta\, n_i^3 - \ln n_i!\right] \qquad (2)$$

$$H_2(\mathbf{n}) = -\sum_{i \leq j} J_{ij} n_i n_j. \qquad (3)$$

$H_1$ and $H_2$ encode extrinsic and intrinsic sources of correlations in the population, respectively. The first term $H_1$ accounts for the behaviour of single cells in response to the stimulus. $h_i(t) = \hat{h}_i[S_t]$ corresponds to a time-dependent external field applied to neuron $i$, which reflects the influence of the past stimulus $S_t$ at time $t$. The functional form and parametrization of $\hat{h}_i$ as a function of the stimulus will be prescribed later, and depend on the particular sensory system and stimulus of interest. The quadratic, cubic, and factorial terms in $n_i(t)$ in $H_1$ correspond to a correction to the Poisson distribution of spikes allowing for general dependencies between the mean and variance of $n_i$. These corrections have been shown to be essential for describing single neurons [25]. The second interaction term $H_2$ is parametrized by a matrix of couplings between neurons, $\mathbf{J} = (J_{ij})$.

Given recordings of the activity of a neural population presented with a known sensory stimulation, the goal is to infer the parameters of the model to best predict the collective response to arbitrary stimuli.

### B. Why direct likelihood maximization cannot be used

Our goal is to infer couplings that solely reflect noise correlations, for two reasons. First, this makes the values of the different parameters easier to interpret. Second, while the stimulus correlation will systematically change with the stimulus statistics, noise correlations may reflect some intrinsic network properties, and the corresponding couplings should be robust to the stimulus statistics. By separating the two types of correlations, we want to develop models that can generalize and predict responses to stimulus ensembles that are radically different from the ones they were trained on.

One strategy to infer the population model (1) could be to estimate all the parameters by maximizing the likelihood on the complete dataset. However, this approach does not explicitly separate noise and signal correlations in the data, and is expected to misestimate the coupling matrix in the inference procedure. This effect comes from the fact that the stimulus-encoding model is never perfect. As a result, when the stimulus-dependent fields $h_i(t)$ fail to perfectly reproduce single neuron activities, the interaction field $\sum_j J_{ij}\sigma_j$ may try to compensate this error by drawing additional information about the stimulus from the activity of other neurons, instead of reserving these couplings for accounting for noise correlations. Previous work [22] combining a simple (i.e. linear) stimulus-encoding model with neuron-neuron couplings, and using maximum likelihood as an inference method, shows a clear example of this effect: the addition of couplings improves the prediction of single-cell firing rates (Fig. 2b of [22]). An extreme version of this phenomenon is also demonstrated in [9], Fig. 15, where the response of one neuron to a natural scene is predicted from the responses of the other neurons without any use of the stimulus.

It should also be noted that the full maximum-likelihood task can be computationally hard. The in-

ference of precise single-cell encoding models such as described in [24] and the inference of complete interaction networks are each computationally costly, and combining both in a single maximum-likelihood maximization would require developing new methods and algorithms.

To demonstrate the effect described above in a computationally tractable case, we inferred a Generalized Linear Model (GLM) to our retinal data (see Sec. III A for a description of the data). The GLM is arguably the most popular model to account for some dependence on the stimulus and for noise correlations using linear filters [23]. Note that the GLM is not expected to accurately describe retinal activity because of its linear assumptions [24], but it can still be used to make our point that its inference mixes signal and noise correlations. As shown in App. A, straightforward likelihood maximization does not guarantee that coupling terms only reflect noise correlations; rather, they reflect an uncontrolled combination or mixture of signal and noise correlations, which have very different biological interpretations.

In summary, from previous observations and the analysis on our own data with the GLM, we conclude that direct maximum-likelihood optimization of the model's parameters is not appropriate and will give biased estimates of the model parameters, and in particular of the coupling network.

### C. Inference of neuronal couplings

In order to obtain the parameters values $J_{ij}$ that account solely for noise correlations, we need a model that reproduces perfectly the time course of the firing rate of each neuron, thereby setting aside the question of its dependence on the stimulus. To construct such a model, we use repetitions of the same stimulation to estimate the empirical firing rate of cell $\lambda_i(t)$ (the average number of spikes in each time bin across repetitions). We then let the time-dependent fields $h_i(t)$ in the population model (1) be inferred from the responses to the repeated stimulus along with the couplings $J_{ij}$, by maximizing the likelihood following [22]. In the maximum-likelihood fit, the fields $h_i(t)$ act as Lagrange multipliers (or chemical potentials if the spike count is viewed as a particle number) enforcing the value of firing rate of each neuron in each time bin, $\langle n_i(t)\rangle = \lambda_i(t)$, while the couplings $J_{ij}$ enforce the noise correlations averaged over the repeated stimulus: $(1/T)\sum_{t=1}^{T}\langle(n_i(t)-\lambda_i(t))(n_j(t)-\lambda_j(t))\rangle$. Thanks to these constraints, the time courses of the firing rates are exactly reproduced by construction, and so are the resulting stimulus correlations. The inference of the couplings thus only reflects noise correlations.

Note that the model inferred in this way is not a stimulus encoding model. It cannot predict the spiking activity in response to a different stimulus sequence than the one used in the repetition. To do this, we need to learn how $h_i(t)$ depends on the stimulus, which is the object of the next step of the inference.

### D. Conditionally independent model

Let us now assume that a model can be built to predict the firing rate $\lambda_i(t)$ of each neuron as a function of the presented stimulus, $\lambda_i(t) = \hat{\lambda}_i[S_t]$. This step depends on the specific sensory system studied, as well as on the stimulus ensemble. We will see an explicit example of such as model in the case of the retina in Sec. III D.

From these firing rate predictions, we derive a non-interacting model of neurons in which the interaction term $H_2$ has been removed:

$$P_{\rm CI}(\mathbf{n}|t) = \frac{1}{Z}\exp\left[-H_1(\mathbf{n},t)\right], \tag{4}$$

where the fields are set as a function of the stimulus, $h_i(t) = \hat{h}_i^0[S_t]$, to enforce the constraint $\langle n_i(t)\rangle_{\rm CI} = \hat{\lambda}_i[S_t]$. This model is conditionally independent, meaning that neurons respond independently of each other, when conditioned on a given stimulus.

### E. Putting it together: mean-field correction to network effects

Now that we have inferred the parameters of the interaction network ($J_{ij}$) and of the conditionally independent model (the functions $\hat{h}_i^0$), the last step is to combine them to obtain the complete population encoding model (1). In doing so, we must be careful to correct for the effect of the network on the activity of each neuron.

Because of the interaction term $H_2$, each cell receives an additional field $\sum_j J_{ij}n_j$ from the rest of the network. This field is stochastic and explains noise correlations. However, it also generates a mean contribution $\Delta\hat{h}_i[S_t]$ that affects the cell firing rate. This contribution thus needs to be removed from the stimulus-dependent field: $\hat{h}_i[S_t] = \hat{h}_i^0[S_t] - \Delta\hat{h}_i[S_t]$.

To estimate this correction, we compute the Thouless, Anderson and Palmer (TAP) free energy formalism of the model (1) [26] and use it to derive an approximation for $\Delta\hat{h}_i$. We followed [27] and [28] and apply a second-order Plefka expansion [29] (see App. B for more details). The result is:

$$\begin{aligned}\Delta\hat{h}_i \;\approx\; & \sum_{j\neq i} J_{ij}\hat{\lambda}_j + J_{ii}\Big(V'(\hat{\lambda}_i) + 2\hat{\lambda}_i\Big) \\ & + \frac{1}{2}V'(\hat{\lambda}_i)\sum_{j\neq i} J_{ij}^2 V(\hat{\lambda}_j) + \frac{1}{2}J_{ii}^2 W'(\hat{\lambda}_i)\end{aligned} \tag{5}$$

where the first two terms are the mean-field contributions of network and self couplings, whereas the lasts two are their Onsager [30] reaction terms. In (5) we have introduced the functions:

$$V(\lambda) \;\equiv\; \langle n^2\rangle_\lambda - \langle n\rangle_\lambda^2 \tag{6}$$

$$W(\lambda) \;\equiv\; \langle n^4\rangle_\lambda - \langle n^2\rangle_\lambda^2 - \frac{\big(\langle n^3\rangle_\lambda - \langle n^2\rangle_\lambda\langle n\rangle_\lambda\big)^2}{V(\lambda)} \tag{7}$$

where $\langle\cdot\rangle_\lambda$ denote averages according to the distribution:

$$P_\lambda(n) = \frac{1}{Z}\exp\left[h(\lambda)n - \gamma n^2 - \delta n^3 - \ln(n!)\right] \quad (8)$$

where $h(\lambda)$ is set to enforce $\langle n\rangle = \lambda$. Note that the correction (5) only depends on the stimulus through the predicted firing rates $\hat{\lambda}[S_t]$, and not on the spike counts $n_i(t)$, as desired.

This computation completes the procedure. The population model (1), now endowed with stimulus-dependent field functions $\hat{h}_i[S_t]$, spike count parameters $\gamma$ and $\delta$, and network couplings $J_{ij}$, can be used to predict the population response to any stimulus $S$. Next we apply the procedure to the retinal network.

## III. APPLICATION TO THE RETINA

### A. Description of the data

We reanalyzed a dataset of *ex-vivo* recording of the activity of retina ganglion cells (RGCs) in Long-Evans adult rats using a multi-electrode array [24] (see Fig. 1A, top for an example response). The stimulus was a movie of two parallel bars, whose trajectories followed the statistics of two independent overdamped stochastic oscillators. Additionally, a white noise stimulus (random binary checkerboard) was projected for one hour to allow for receptive field estimation and RGC type identification. Raw voltage traces were stored and spike-sorted off-line through a custom spike sorting algorithm [31]. We applied a standard clustering analysis based on the cell response to various stimuli to isolate a population of OFF ganglion cells of the same type. Their receptive fields tiled the visual field to form a mosaic (Fig. 1B).

The ganglion cell spike times were binned in (1/60)s time windows (locked to the stimulus frame rate) to estimate the empirical spike counts, $n_i(t)$ for cell $i$ in time bin $t$. The stimulus alternates between non-repeated sequences of random bar trajectories, and a repeated sequence of randomly moving bars, displayed 54 times. Non-repeated sequences were divided into training (two thirds of the total) and testing (one third) sets. Repeated sequences were equally divided into training and testing sets by splitting the repeated trajectory in two different halves.

### B. Stimulus and noise correlations

Before describing the inference of the model, we first briefly characterize the amount and properties of the stimulus and noise correlations. To this end, we estimate the mean firing rate $\lambda_i(t)$ as a function of time in response to the repeated stimulus sequence as the empirical mean of $n_i(t)$ across repetitions (Fig. 1A, bottom), and its temporal average as $\lambda_i \equiv (1/T)\sum_{t=1}^{T}\lambda_i(t)$, where $t = 1,\ldots,T$ spans the duration of the repeated stimulus. We measure the covariance between pairs of cells (example in Fig. 1C) computed from the repeated dataset between two cells. The total pairwise covariance, represented in green in Fig. 1C, is the sum of stimulus and noise covariances,

$$c_{ij}^{\text{tot}}(\tau) = c_{ij}^{S}(\tau) + c_{ij}^{N}(\tau), \quad (9)$$

with

$$c_{ij}^{\text{tot}}(\tau) = \frac{1}{T}\sum_{t=1}^{T}\langle(n_i(t)-\lambda_i)(n_j(t+\tau)-\lambda_j)\rangle_{\text{rep}}, \quad (10)$$

$$c_{ij}^{S}(\tau) = \frac{1}{T}\sum_{t=1}^{T}(\lambda_i(t)-\lambda_i)(\lambda_j(t+\tau)-\lambda_j), \quad (11)$$

$$c_{ij}^{N}(\tau) = \frac{1}{T}\sum_{t=1}^{T}\left\langle(n_i(t)-\lambda_i(t))(n_j(t+\tau)-\lambda_j(t+\tau))\right\rangle_{\text{rep}}, \quad (12)$$

where $\langle\cdot\rangle_{\text{rep}}$ represents averages over stimulus repetitions. The stimulus covariance, shown in black in Fig. 1C, can be calculated from the empirical firing rates $\lambda_i(t)$. The noise covariance corresponds to the difference between the total and stimulus covariance. Fig. 1C shows that this difference is significantly different from zero only at zero lag ($\tau = 0$), meaning noise correlations happen at a short time scale. This suggests they may be due to gap-junctions [18, 32]. Only cells that are physically close (as measured by the distance between the neurons' receptive fields) have large noise correlations (Fig. 1D), and their values strongly decreased with distance.

### C. Coupling network

We applied the procedure described in Sec. II C to the responses to the repeated stimulus to learn the coupling matrix $\mathbf{J}$. The inference was performed by maximizing the log-likelihood (see for example [33]). We added a small L2 penalty (with coefficient $\eta_{\text{L2}} \sim 2\cdot 10^{-6}$) on the fields $h_i(t)$ to avoid divergences to $-\infty$ when $\lambda_i(t) = 0$. In order to avoid spurious non-zero values, we also add an L1 penalty (with coefficient $\eta_{\text{L1}} = 0.04$). We further imposed that $J_{ii}$ be independent of $i$, for consistency with the single-cell model (see below).

Fig. 2 shows the results of the inference. To evaluate the robustness of the inference with respect to a change of stimulus realization, in Fig. 2A we plot the interactions inferred from the training set against those inferred from another training set of the same size, where the bars followed a different trajectory. The comparison shows that inferred networks are robust against a change of stimulus.

To check the validity of this approach, in Fig. 2B we compare empirical noise correlations obtained with the test dataset with those predicted by the model. To obtain this prediction, we freeze the $\mathbf{J}$ matrix obtained from

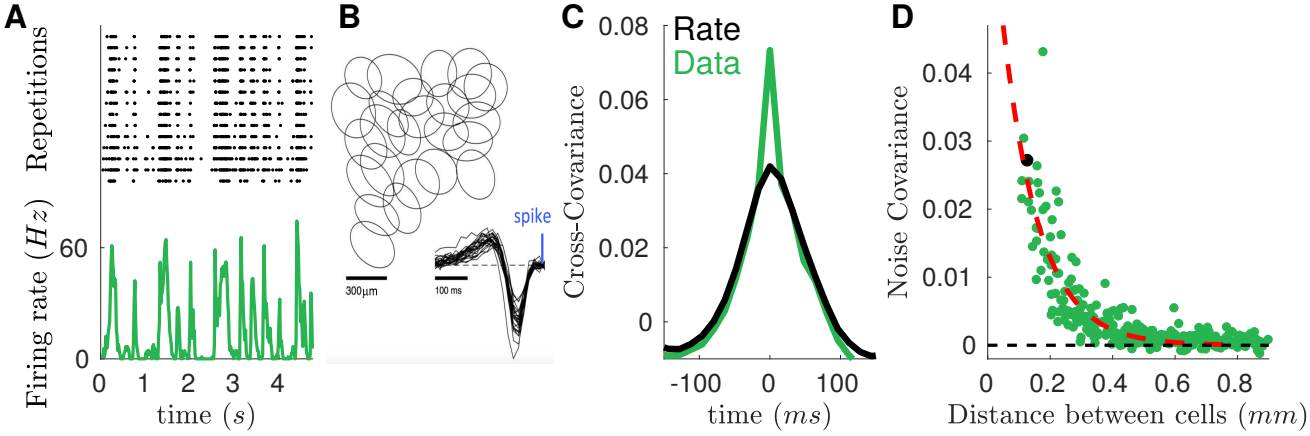


FIG. 1: ***Ex-vivo* recording of retinal ganglion cell shows short time-scale noise correlations.** A) Top: activity of one ganglion cell over repeated presentations of the same stimulus. Bottom: firing rate of the cell across time. B) Receptive field mosaic of the isolated OFF ganglion cells C) Empirical cross-covariance between two example cells (green trace) superimposed on the covariance between their average firing rate (black trace). The difference, which is large only at short time-scales, is the noise correlation. D) Zero-lag noise correlation as a function of cell pair distance. Black point refers to the pair of neurons plotted in panel C. Red dashed line shows an exponential fit with spatial scale of $0.11mm$.

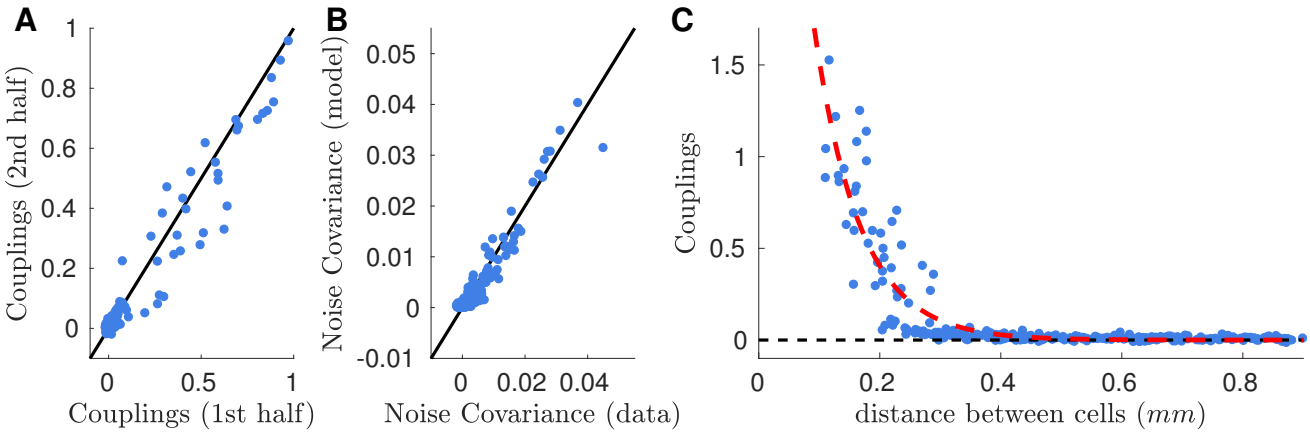


FIG. 2: **Time-dependent inference provides good estimates of the interaction network.** A) Comparison between the inferred interaction from first (later used as training) and second (later used as testing) halves of the repeated dataset. Black line is the identity. B) Comparison of the predicted noise correlations when the interaction matrix $\boldsymbol{J}$ learned using the training set is applied to the testing set. Black line is the identity. C) The behavior with distance of the inferred interactions scales similarly to that of noise correlations (see Fig. 1D), although it goes to zero at shorter distances. Red dashed line is an exponential fit with spatial constant $0.08mm$.

the inference on the training set and we re-infer the $h_i(t)$ to match the firing rates of the testing set. The inferred coupling matrix is able to well predict the noise correlation on a part of the recording that had not been used to learn them. Fig. 2C shows the behavior of the interaction parameters as a function of the distance between the two neurons' receptive fields. $J_{ij}$ decreases with distance slightly faster than noise correlation (see Fig. 1 for comparison).

### D. A feed-forward single-cell model

To apply the procedure of Sec. II D, we need a model for the encoding of the stimulus by single neurons. We use a previously proposed feed-forward model [24] that was specifically developed to predict responses to two-bar stimuli. The stimulus $S(x,t)$, representing the time behavior of each pixel, is first convolved with a Gaussian and biphasic factorized kernel $K_{\rm BP}(x,t)$ and then passed through rectified quadratic units with the two possible

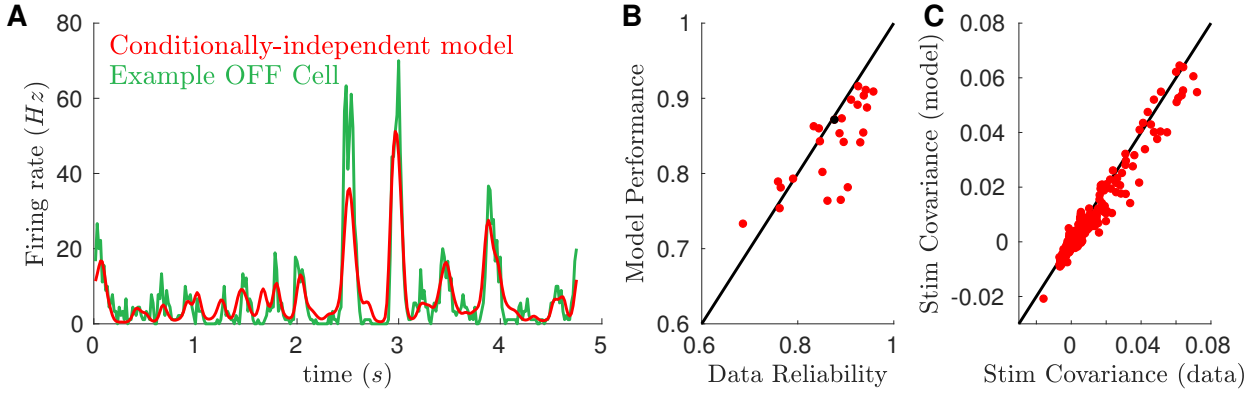


FIG. 3: **Predictions of the conditionally independent feedforward model.** A) For an example cell, the model-predicted firing rate (red trace) is super-imposed to the empirical one (green trace). By computing the correlation between the two traces ($\rho = 0.87$) we can estimate the model performance. B) Scatter-plot of the model performance against the cell reliability, computed as the correlation between two halves of the repeated dataset. Black point refers to the cell of panel A. C) Model-predicted stimulus covariances account for more than 84% of stimulus covariance (slope of a linear fit), but systematically under-estimate their empirical value. Black line is the identity.

polarities:

$$\Theta^{\pm}(x,t) = \left[\int dx'\, dt'\, K_{\rm BP}(x-x', t-t')S(x',t')\right]^2_{\pm} \tag{13}$$

where $[y]_+ = \max(y,0)$ and $[y]_- = \min(y,0)$ . The intermediate variable $\Theta(x,t)$ is then fed into a second non-linear stage: for each cell $i$, it is first convolved with a receptive field $K_i^{\pm}(x,t)$ and then passed through a non-linear function:

$$\begin{aligned}\hat{\lambda}_i[S_t] =& f_i\left(\int dx\, dt'\, K_i^{+}(x,t-t')\Theta^{+}(x,t)\right.\\ &\left.+K_i^{-}(x,t-t')\Theta^{-}(x,t)\right),\end{aligned} \tag{14}$$

with $f_i(y) = a_i \ln[1+\exp\left(b_i(y+c_i)\right)]$.

To infer the parameters of the model we follow a simplified version of [34], where only the second non-linear stage is learned. We keep fixed the first stage on a setting that has been shown to work efficiently [24]. For the second stage we apply an iterative procedure where we maximize the log-likelihood of the data under the model given by Eq. 4, penalized by a weighted sum of the L1 and L2 norms of the parameter vector. We used the non-repeated training set to compute the log-likelihood gradient. To avoid overfitting we early-stopped the iterative procedure when the log-likelihood computed on the non-repeated testing set stopped to increase. L1 and L2 penalties were optimized by maximizing the performance on the repeated dataset that later we will use for training the population model.

In Fig. 3A we compare the time course of the empirical firing rate, $\lambda_i(t)$, with the prediction of the inferred model for an example cell. By computing the Pearson correlation among these two temporal traces ($\rho = 0.87$), we can estimate the performance of the model for each cell. In Fig. 3B we compare this performance with the reliability of the retinal spike activity, estimated as the correlation between two disjoint subsets of responses to the repeated stimulus and found that they were comparable. In Fig. 3C we show how the model predicts the empirical stimulus covariances. Even if a small under-estimation is present, the model accounts for more than 84% of the empirical value (slope of a linear fit).

### E. Complete population model

The final step is to combine the single cell model and the interaction network, as explained in Sec. II E. In Fig. 4 we compare the performance of the population (1) and conditionally-independent (4) models on a testing set that was not used for learning. In panels A and B, we check that inferring the population model still preserves the quality of the prediction of single cell activity obtained with the independent model. We compare the performance of the two models in reproducing the firing rate of the recorded cells (same quantity as Fig. 3B). Firing rate is a stimulus dependent quantity and accordingly the two models have similar performance. The fact that the population model's performance is not degraded compared to the single-cell model validates the approximations made to calculate the corrections $\Delta\hat{h}[S_t]$ within the TAP approach. In addition, the fact that this performance is not improved is also a positive sign: it implies that the couplings do not try to compensate for failings of the encoding model, and only reflects noise correlations. Panel B shows that the two models have similar performance also for spike count variance.

However, the population model largely outperforms the conditionally independent model in predicting the population joint activity. Fig. 4C shows how the popula-

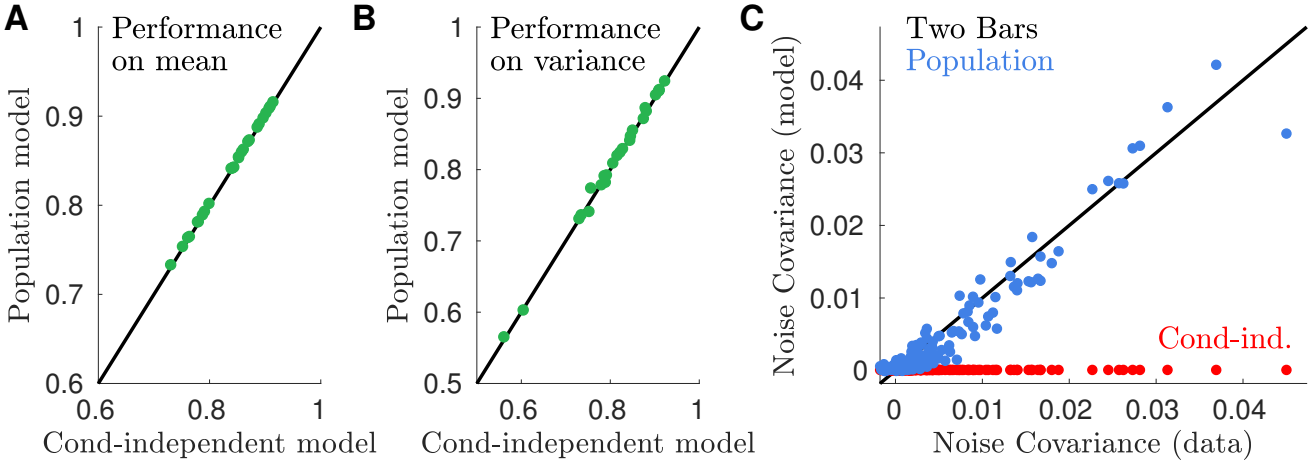


FIG. 4: **The population model predicts empirical noise covariances and performs as well as the conditional-independent model on stimulus-dependent quantities.** A) Performance in estimating the cell firing rate (same as Fig. 3B) of the population model is equivalent to that of the conditionally independent model. B) Both models also have the same performance in predicting the spike count variance. C) The population model predicts empirical noise covariances when applied on a testing set (blue points), while the conditionally independent model predicts zero noise covariances (red).

tion model accounts well for noise covariances on a testing set (blue points). By construction, the conditionally independent model predicts vanishing noise covariance (red points).

### F. Robustness of the inference of the couplings to the choice of stimulus

A major challenge with fitting complex encoding models to neural responses is that they rarely generalize well. Here we ask if the interaction network can be inferred from a qualitatively different stimulus and then applied to our two-bars stimulation. To test for this, we infer the couplings as in Sec. III C, but on the response to repeated whitenoise (random checkerboard) stimuli. In fig. 5A we compare the couplings inferred from the whitenoise stimulus, to the couplings inferred from the two-bar stimulus. We then use the coupling matrix $\mathbf{J}$ learned on the whitenoise stimulus in a complete population model for predicting the response to two-bar stimuli (fitted following Secs. III D and III E). In fig. 5B we show that this model predicts noise covariances when applied to the two-bar testing set.

This demonstrates that our inference method allows us to generalize from one stimulus type to the other, and that the inferred couplings between neurons are invariant to the stimulus.

## IV. DISCUSSION

We have introduced a model for the spiking activity of population of sensory neurons responding to a stimulus, in which extrinsic and intrinsic correlations are clearly separated. Our method is general and could be applied to other structures than the retina. It could also be extended to models where the influence of the stimulus on single cell activity is described by different non-linear models than illustrated here, e.g. a deep network with more layers [35].

Our inference strategy allows us to infer couplings between neurons that only reflect noise correlations between neurons, without the interference of stimulus effects due to artifacts of the learning procedure. Such effects can arise when the inference procedure tries to use the activity of other cells as a surrogate of the stimulus itself to improve the predictability of a given cell, compensating for non-linearities in the stimulus dependence that are unaccounted by the model. The inferred couplings thus show a weak dependence on the stimulus ensemble driving the neuronal response, in contrast to previous attempts where the stimulus ensemble had a major influence on the couplings [5]. Note that checkerboard and two-bars stimulations drive very different responses of the retina. It is thus a remarkable result that noise correlations in the response to a complex film of moving objects can be predicted by couplings inferred from responses to white-noise stimuli. This result can thus be seen as a first step toward the construction of a full model that accounts for large and heterogeneous stimulus ensembles.

Having models that account for noise correlations is a crucial first step to study their impact on coding complex stimuli. This impact can be quantified by comparing the population model to the conditionally independent model ($J_{ij} = 0$), or to alternative population models with different coupling matrices. By comparing the computations performed by these models, future works will be able to understand how these fast noise correlations affect coding in the retina. The same strategy can be used in other

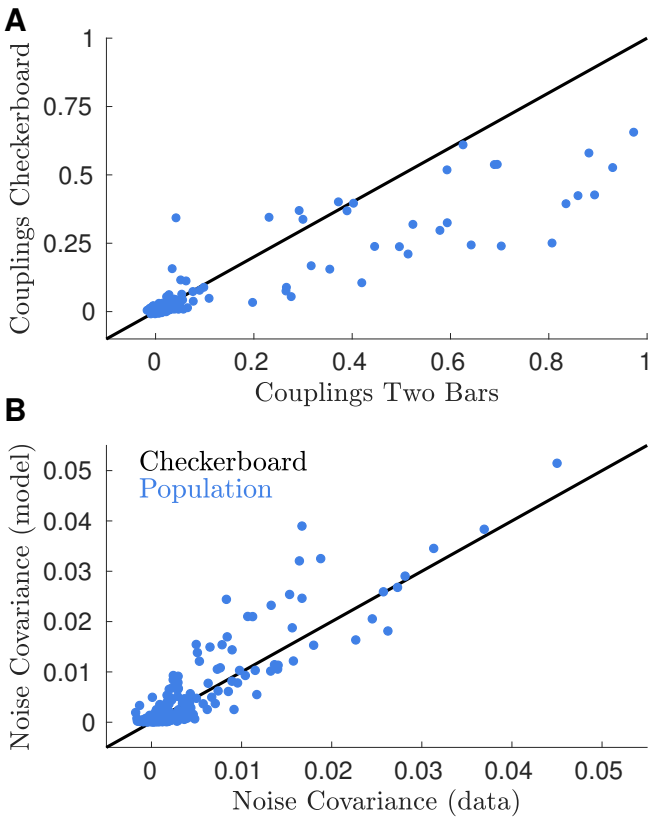


FIG. 5: **Inferred couplings generalize from one stimulus ensemble to another.** A) Scatterplot of the inferred couplings from the response to the two bars stimulation (x-axis) against those from the response to checkerboard. B) Population model prediction of empirical noise covariances in response to the two bars when the couplings are inferred from repeated checkerboard stimulation.

sensory structures or any other noisy input-output systems.

## V. ACKNOWLEDGEMENTS

We like to thank C. Gardella for useful discussion. This work was supported by ANR TRAJECTORY, the French State program Investissements d'Avenir managed by the Agence Nationale de la Recherche [LIFESENSES: ANR-10-LABX-65], a EC grant from the Human Brain Project (FP7-604102)), NIH grant U01NS090501 and AVIESAN-UNADEV grant to OM.

### Appendix A: Generalized linear model analysis

We fitted a generalized linear model (GLM) to the responses of $N_r = 25$ off-cells in response to the moving bar stimuli. We discretized the response and stimulus using temporal bins of length $1.667ms$ (i.e. 600Hz; 10 times smaller than the temporal bins used for the model described in the main paper). The response of the $i^{th}$ neuron at time $t$ is described by an integer $n_i(t)$ denoting the number of spikes fired by the $i^{th}$ neuron in the $t^{th}$ time bin. The spatial location of the presented bar stimuli were discretized into $N_x = 100$ equi-sized bins. The stimulus at time $t$ was denoted by a binary vector $\boldsymbol{x}_t$, of which the $i^{th}$ component $(x_i(t))$ was set to 1 if one of the bars was centred on the corresponding spatial location at time $t$, and 0 otherwise.

In a GLM, the spiking response of each neuron is assumed to be Poisson distributed, with the mean number of spikes of the $i^{th}$ neuron at time $t$ given by $\exp(r_i(t))$, where:

$$r_i(t) = \sum_{j=1}^{N_x} \sum_{k=0}^{N_w-1} w_{ijk} x_j(t-k) + \sum_{j=1}^{N_r} \sum_{k=1}^{N_v} v_{ijk} n_j(t-k) + b_i. \tag{A1}$$

In this equation, $b_i$ denotes a constant bias term, $w_{ijk}$ is an element of the temporal stimulus filter (of size $[N_r, N_x, N_w]$)), and $v_{ijk}$ is an element of the recurrent filter (of size $[N_r, N_r, N_v]$)). We used a stimulus filter of length $N_w = 200$ (i.e. ∼300ms), and recurrent filters of length $N_v = 15$ (i.e. ∼25ms).

Model parameters were fitted by maximizing an objective function consisting of the log-likelihood (which, for a Poisson distribution, is given up to constant by $L = \sum_{i,t} n_i(t) r_i(t) - e^{r_i(t)}$), minus an L2 norm regularization term that promoted smooth filters and thus prevented overfitting. The regularization parameters were chosen to maximize the log-likelihood on test data, held out during model fitting. Further, to reduce the number of parameters, and thus further reduce over–fitting we assumed that the stimulus filter could be described as the sum of $N_{rank} = 3$ spatio-temporally separable filters, each given by a spatial filter multiplied by a temporal filter (i.e. $w_{ijk} = \sum_{l=1}^{N_{rank}} u_{ij}^l a_{ik}^l$, where $u$ and $a$ here denote the spatial and temporal filters, respectively). Relaxing this assumption (by increasing $N_{rank}$) did not improve the quality of the model fit.

We evaluated the performance of the GLM model in fitting the covariances between the responses of pairs of neurons (Fig. 6A). Interestingly, despite giving a reasonable fit of the total response covariances (Pearson correlation = 0.87), the model gave a poor fit of the noise covariances (compare Fig. 6B with Fig 4C). We wondered whether this could be because the recurrent filters, $v$, did not just capture interactions between neurons, but also compensated for the inability of the feed-forward filters, $w$, to fully capture the effects of the stimulus on neural firing rates. To see if this was the case, we compared the coupled 'population model' described above, and with firing rates given by equation A1, with an uncoupled 'conditionally-independent' model, where the recurrent filters, $v$, were set to zero. We found that the coupled GLM model resulted in improved predictions of the recorded PSTH for nearly all recorded OFF-cells, compared to the uncoupled model (Fig. 6C-D). This suggests

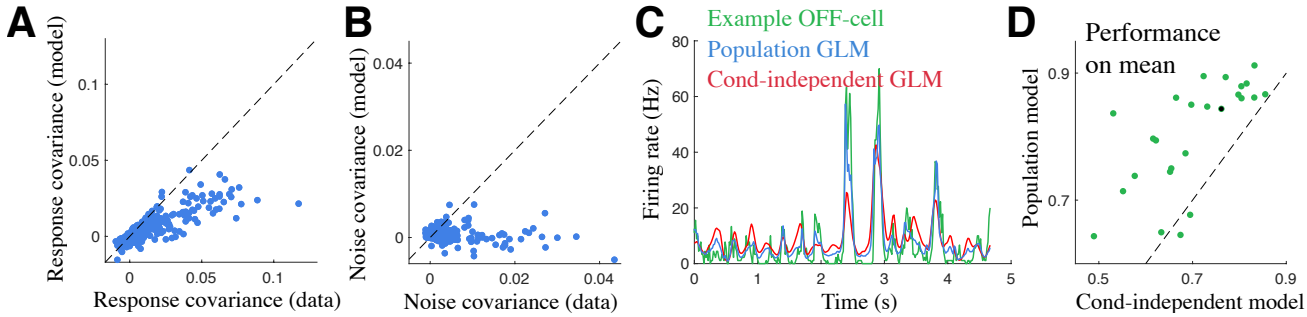


FIG. 6: (A) Total response covariances predicted by the coupled GLM model (y-axis) versus the response covariances measured from the data (x-axis). Each point corresponds to a single neuron pair. (B) Noise covariances predicted by the coupled GLM model (y-axis) versus the noise covariances measured from the data (x-axis). (C) Firing rate of a single example cell (green; also plotted in Fig 2A), alongside the predictions of the coupled ('population') (blue) and uncoupled ('conditionally-independent') (red) GLM models. (D) Correlation coefficient between the firing rate predicted by the coupled (y-axis) and uncoupled (x-axis) GLM models and data, for each cell. The cell shown in panel C is indicated in black.

that, rather than just fitting the interactions between different neurons, the coupled GLM model used the recurrent filters, $v$, to improve the prediction of how each neuron responded to the stimulus. However, it remains to be seen whether further differences between the GLM and the Ising model (e.g. time-dependent recurrent filters, Poisson distributed firing rates) could also contribute to their different performances in predicting the noise covariances.

### Appendix B: Construction of the mean-field theory and Thouless-Anderson-Palmer correction

We are interested in computing the TAP correction to the fields $h_i(t)$ due to the addition of the coupling term $\mathbf{J}$, see Eq.(1). Because we are not interested in the TAP expression for couplings nor in that for covariances, we can construct the mean-field theory for a single time-bin. Otherwise, because the couplings are constant in time we should have considered the whole model. To apply the Plefka expansion we introduce:

$$F[\alpha, \boldsymbol{h}, J] \equiv -\ln \sum_{\mathbf{n}} \exp\left\{-H_1(\mathbf{n}) - \alpha H_2(\mathbf{n})\right\} \tag{B1}$$

where we neglect the $t$ dependence of $H_1$ because here we focus on a single time-bin. The Legendre transform of (B1) reads:

$$G[\alpha, \boldsymbol{\lambda}, J] = \left( \sum_i h_i \lambda_i + F[\alpha, \boldsymbol{h}, J] \right|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} \tag{B2}$$

where $\tilde{\boldsymbol{h}} = \tilde{\boldsymbol{h}}[\alpha, \boldsymbol{\lambda}]$ is defined implicitly from:

$$\left. \frac{\partial \left( \sum_i h_i \lambda_i + F[\alpha, \boldsymbol{h}, J] \right)}{\partial h_i} \right|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} = \lambda_i - \left. \langle n_i \rangle^{(\alpha)} \right|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} = 0 \tag{B3}$$

where $\langle \dots \rangle^{(\alpha)}$ is the average with respect to the distribution related to the free energy (B1). Our goal is to expand $G(\alpha, \boldsymbol{\lambda}, j)$ in power of $\alpha$ up to the second order. At first we evaluate the derivatives:

$$G'[\alpha, \boldsymbol{\lambda}, J] = \langle H_2(\mathbf{n}) \rangle^{(\alpha)} \tag{B4}$$

$$G''[\alpha, \boldsymbol{\lambda}, J] = -\langle H_2(\mathbf{n})^2 \rangle_c^{(\alpha)} - \sum_g \frac{\partial \tilde{h}_g}{\partial \alpha} \langle H_2(\mathbf{n})(n_g - \lambda_g) \rangle^{(\alpha)} \tag{B5}$$

$$\frac{\partial \tilde{h}_g}{\partial \alpha} = -\frac{\langle H_2(\mathbf{n})(n_g - \lambda_g) \rangle^{(\alpha)}}{\langle n_g^2 \rangle_c^{(\alpha)}} \tag{B6}$$

where to obtain the last equality we applied the implicit function theorem to Eq. (B3). The Plefka approximation consists in estimating $G[\boldsymbol{\lambda}, J] = G[\alpha = 1, \boldsymbol{\lambda}, J]$ from the expansion around $\alpha = 0$ evaluated at $\alpha = 1$:

$$G[\boldsymbol{\lambda}, J] \approx G[0, \boldsymbol{\lambda}, J] + G'[0, \boldsymbol{\lambda}, J] + \frac{1}{2} G''[0, \boldsymbol{\lambda}, J] \ . \tag{B7}$$

We need $G[\alpha, \boldsymbol{\lambda}, J]$ and its derivatives at $\alpha = 0$. To this aim we note that $\tilde{h}_i[\alpha = 0, \boldsymbol{\lambda}] = \tilde{h}_i[\lambda_i]$, as for $\alpha = 0$ the system units become independent and consequently $\tilde{h}_i$ depends only on $\lambda_i$. For $\alpha = 0$, in fact, the distribution over $\{n_i\}_{i=1}^N$ factorizes over a set of single variable distributions. this allows us to compute model expectations at $\alpha = 0$. For future convenience, we define the moments of such distributions:

$$P^{(s)} \equiv \langle n^s \rangle^{(\alpha=0)} \tag{B8}$$

so that the terms in the expansion become:

$$G[0,\boldsymbol{\lambda},J] = \sum_i \tilde{h}_i\lambda_i + F[\alpha=0,\tilde{h}] \quad \text{(B9)}$$

$$G'[0,\boldsymbol{\lambda},J] = -\sum_{i<j} J_{ij}\lambda_i\lambda_j - \sum_i J_{ii}P_i^{(2)} \quad \text{(B10)}$$

$$G''[0,\boldsymbol{\lambda},J] = -\sum_{i<j} J_{ij}^2\big(P_i^{(2)}-\lambda_i^2\big)\big(P_j^{(2)}-\lambda_j^2\big) - \sum_i J_{ii}^2 W_i \quad \text{(B11)}$$

where $W$ has been defined in Eq. (7).

The mean-field equation for the fields $\boldsymbol{h}$ can be easily obtained by a reverse Legendre transform of Eq. (B7):

$$h_i = \frac{\partial G[\boldsymbol{\lambda},J]}{\partial\lambda_i} \quad \text{(B12)}$$

$$= \tilde{h}_i[\lambda_i] + \frac{\partial G'[0,\boldsymbol{\lambda},J]}{\partial\lambda_i} + \frac{1}{2}\frac{\partial G''[0,\boldsymbol{\lambda},J]}{\partial\lambda_i} \quad \text{(B13)}$$

which provides the expression (5) for the TAP correction.


[1] Schneidman, E, Bialek, W, & Berry, M. (2003) *Journal of Neuroscience* **23**, 11539–11553.
[2] Schneidman, E, Still, S, Berry, M. J, Bialek, W, et al. (2003) *Physical review letters* **91**, 238701.
[3] Schneidman, E, Berry, M, Segev, R, & Bialek, W. (2006) Nature **440**, 1007.
[4] Shlens, J, Field, G. D, Gauthier, J. L, Grivich, M. I, Petrusca, D, Sher, A, Litke, A. M, & Chichilnisky, E. (2006) *Journal of Neuroscience* **26**, 8254–8266.
[5] Cocco, S, Leibler, S, & Monasson, R. (2009) Proc. Natl. Acad. Sci. USA **106**, 14058.
[6] Ganmor, E, Segev, R, & Schneidman, E. (2011) PNAS **108**, 9679–9684.
[7] Tkacik, G, Granot-Atedgi, E, Segev, R, & Schneidman, E. (2013) *Phys. Rev. Lett.* **110**, 058104.
[8] Tkacik, G, Marre, O, Mora, T, Amodei, D, Berry, M, & Bialek, W. (2013) *Journal of Statistical Mechanics: Theory and Experiment* **2013**, P03011.
[9] Tkacik, G, Marre, O, Amodei, D, Schneidman, E, Bialek, W, & M.J, B. (2014) PloS Comput. Biol. **10(1)**, e1003408.
[10] Marre, O, El Boustani, S, Frégnac, Y, & Destexhe, A. (2009) *Phys. Rev. Lett.* **102**, 138101.
[11] Hamilton, L. S, Sohl-Dickstein, J, Huth, A. G, Carels, V. M, Deisseroth, K, & Bao, S. (2013) Neuron **80**, 1066–76.
[12] Tavoni, G, Ferrari, U, Battaglia, F, Cocco, S, & Monasson, R. (2017) *Network Neuroscience* pp. 1–27.
[13] Nghiem, T.-A, Telenzuk, B, Marre, O, Destexhe, A, & Ferrari, U. (2018) *bioRxiv* p. 243857.
[14] Meshulam, L, Gauthier, J. L, Brody, C. D, Tank, D. W, & Bialek, W. (2017) *Neuron* **96**, 1178–1191.
[15] Jaynes, E. T. (1982) Proc. IEEE **70**, 939.
[16] Bloomfield, S & Völgyi, B. (2009) *Nature Reviews Neuroscience* **10**, 495–506.
[17] Völgyi, B, Chheda, S, & Bloomfield, S. (2009) *J. Comp. Neurol.* **512**, 664–687.
[18] Brivanlou, I, Warland, D, & Meister, M. (1998) Neuron **20**, 527–539.
[19] Ala-Laurila, P, Greschner, M, Chichilnisky, E, & Rieke, F. (2011) *Nature neuroscience* **14**, 1309–1316.
[20] Smith, M. A & Kohn, A. (2008) *Journal of Neuroscience* **28**, 12591–12603.
[21] Cohen, M. R & Kohn, A. (2011) *Nature neuroscience* **14**, 811.
[22] Granot-Atedgi, E, Tkacik, G, Segev, R, & Schneidman, E. (2013) *PLOS Computational Biology* **9**, 1–14.
[23] Pillow, J, Shlens, J, Paninski, L, Sher, A, Litke, A, Chichilnisky, E. J, & Simoncelli, E. (2008) Nature **454**, 995–999.
[24] Deny, S, Ferrari, U, Mace, E, Yger, P, Caplette, R, Picaud, S, Tkačik, G, & Marre, O. (2017) *Nature communications* **8**, 1964.
[25] Ferrari, U, Deny, S, Marre, O, & Mora, T. (2018) *bioRxiv* p. 243543.
[26] Thouless, D, Anderson, P, & Palmer, R. (1977) Phil. Mag. **35**, 593.
[27] Kappen, H & Rodriguez, F. (1997) Neural Comput. **10**, 1137–1156.
[28] Tanaka, T. (1998) Phys. Rev. E **58**, 2302.
[29] Plefka, T. (1982) *Journal of Physics A: Mathematical and general* **15**, 1971.
[30] Onsager, L. (1936) *Journal of the American Chemical Society* **58**, 1486–1493.
[31] Yger, P, Spampinato, G. L. B, Esposito, E, Lefebvre, B, Deny, S, Gardella, C, Stimberg, M, Jetter, F, Zeck, G, Picaud, S, Duebel, J, & Marre, O. (2016) *bioRxiv.*
[32] Trenholm, S, McLaughlin, A. J, Schwab, D, Turner, M, Smith, R, Rieke, F, & Awatramani, G. (2014) *Nature neuroscience* **17**, 1759–1766.
[33] Ferrari, U. (2016) *Phys. Rev. E* **94**, 023301.
[34] McFarland, J, Cui, Y, & Butts, D. (2013) *PLoS Comput Biol* **9**, e1003143.
[35] McIntosh, L, Maheswaranathan, N, Nayebi, A, Ganguli, S, & Baccus, S. (2016) *Deep learning models of the retinal response to natural scenes.* pp. 1361–1369.