Encoding surprise by retinal ganglion cells

Danica Despotović[1], Corentin Joffrois[1], Olivier Marre[1], Matthew Chalk[1,*]

1 Sorbonne Université, INSERM, CNRS, Institut de la Vision, 75012 Paris, France

* matthew.chalk@inserm.fr

Abstract

The efficient coding hypothesis posits that early sensory neurons transmit maximal information about sensory stimuli, given internal constraints. A central prediction of this theory is that neurons should preferentially encode stimuli that are most surprising. Previous studies suggest this may be the case in early visual areas, where many neurons respond strongly to rare or surprising stimuli. For example, previous research showed that when presented with a rhythmic sequence of full-field flashes, many retinal ganglion cells (RGCs) respond strongly at the instance the flash sequence stops, and when another flash would be expected. This phenomenon is called the 'omitted stimulus response'. However, it is not known whether the responses of these cells varies in a graded way depending on the level of stimulus surprise. To investigate this, we presented retinal neurons with extended sequences of stochastic flashes. With this stimulus, the surprise associated with a particular flash/silence, could be quantified analytically, and varied in a graded manner depending on the previous sequences of flashes and silences. Interestingly, we found that RGC responses could be well explained by a simple normative model, which described how they optimally combined their prior expectations and recent stimulus history, so as to encode surprise. Further, much of the diversity in RGC responses could be explained by the model, due to the different prior expectations that different neurons had about the stimulus statistics. These results suggest that even as early as the retina many cells encode surprise, relative to their own, internally generated expectations.

Introduction

Visual scenes are highly correlated, both in space and time. It has been hypothesized that neurons in early sensory areas have evolved to exploit this structure, by only encoding 'surprising' sensory signals, that cannot be predicted based on their spatio-temporal context. This efficient coding theory

can account for many qualitative aspects of neural responses in early sensory areas, such as the stimulus selectivity of neurons in the retina [Karklin and Simoncelli, 2011, Doi et al., 2012, Soto et al., 2020], as well as primary visual [Rao and Ballard, 1999, Olshausen and Field, 1996, Van Hateren and van der Schaaf, 1998] and auditory [Lewicki, 2002, Smith and Lewicki, 2006] cortices.

A central prediction of the efficient coding theory is that neurons should best encode stimuli that are surprising, given the recent stimulus history. There appears to be some evidence for this in early visual and auditory areas, where neurons have been found that respond most strongly to rare or surprising stimuli [Ulanovsky et al., 2003, Gill et al., 2008]. In the retina, previous studies found that when a sequence of full-field light flashes are presented, many neurons respond most strongly at the moment the sequence of flashes is stopped, and where another flash would be expected. This phenomenon was labelled the 'omitted stimulus response' (OSR), since it can be considered to be a response to the stimulus that was unexpectedly omitted [Schwartz et al., 2007].

However, if neurons in the retina really do encode surprise, then their responses should vary in a graded way as one varies the level of stimulus surprise. Unfortunately previous studies [Schwartz et al., 2007, Schwartz and Berry 2nd, 2008, Werner et al., 2008] did not test for this, since there were typically only two alternatives: either the stimulus was surprising (e.g the sequence of flashes ends) or it was unsurprising (e.g. the sequence of flashes continues). As a result, it is hard to conclude from these studies whether neurons in the retina encode surprise.

To address this question, we presented retinal ganglion cells (RGCs) with extended sequences of stochastically occurring full-field flashes. With this stimulus, the degree of 'surprise' for each flash (or period of silence between flashes) could be quantified mathematically, and was observed to vary in a graded manner depending on the previous sequence of flashes and silences. We could thus test how RGC responses varied with the level of surprise. Interestingly, we found that the responses of RGCs to these stimulus sequences could be well explained by a simple normative model, which described how neurons optimally combined their prior expectations about the stimulus with the recent stimulus history to encode surprise. Further, we found that much of the diversity in the responses of different recorded RGCs could be explained by this model, due to the different levels of 'confidence' that different neurons had in their prior expectations. Our study provides support for the predictive coding model of retinal coding, while shedding light on the different prior expectations that different RGCs have about environment. More generally, it shows that, already at the stage of the retina, many ganglion cells do not encode the physical stimulus itself, but how unexpected this stimulus is, with different prior expectations for different cells.

Results

RGC responses to flash sequences

We used a multi-electrode array to record retinal ganglion cells (RGCs) of an axolotl. We presented a visual stimulus, consisting of random sequences of full-field dark flashes, interleaved with periods of silence (Fig. 1A; see Methods: Stimulus statistics for details). Recorded neural activity was sorted into single unit responses using SpyKing Circus [Yger et al., 2016].

We were interested in neurons that exhibited an 'omitted stimulus response' (OSR), where they responded to the absence of a flash, following several flashes presented in a row [Schwartz et al., 2007]. We thus selected 48 out of 114 single unit responses for further analysis, that showed (i) high quality recording (quantified by low number (<1%) of refractory period violations, where refractory period is 2 ms), and (ii) the presence of an OSR (quantified as a peak around 120 ms after the omitted flash).

Fig. 1B shows the example responses of one of these cells to a varying number of flashes presented in a row. As can be seen, this cell responded strongly to the first flash in a sequence, and shortly after the sequence had ended (i.e. the OSR). The size of the OSR increased monotonically with the number of flashes presented in a row.

For our analysis, we converted the stimulus to a binary variable, which was set to 1 or 0 depending on whether there was a dark flash (stim. = 1) or a period of silence (stim. = 0) within a given 120ms window. Neural responses were taken to be the number of spikes that occurred within each 120ms window.

To see how the OSR varied with the number of consecutive flashes, we computed the average response of each neuron, given a 'stimulus history' consisting of a varying number of consecutive flashes followed by silence (Figure 1C). The OSR increased monotonically with the number of flashes for all cells. However, we observed differences in the rate of increase as well as the maximum firing rate for different cells (Fig. 1C).

Finally, to see how neural responses depended on all possible stimulus sequences (and not the number of consecutive flashes), we constructed 'tree-plots' (Fig. 1D), showing each neuron's average response to all possible sequences of flashes and silences of a given length. The top branch of this tree plot corresponds to the OSR, shown in Fig. 1D. However, many cells that showed a qualitatively similar OSR (i.e. that increased with the number of consecutive flashes) exhibited very different tree-plots, identifying clear differences in how they responded to different patterns of flashes and silences (e.g. Fig. 1D).

Modeling 'surprise encoding' by RGCs

We asked whether RGC responses were consistent with them encoding surprise. To test this, we constructed a simple model of how RGCs could combine their internal stimulus expectations with their recent stimulus history to compute surprise (Fig. 2A). Following [Shannon, 1948, MacKay et al., 2003], we defined surprise at time t, s_t, as the negative log probability of a stimulus, x_t, given the recent stimulus history, $x_{<t}$, and the neuron's internal model of the stimulus statistics (parameterised by θ):

$$s_t = -\log p\left(x_t | x_{<t}, \theta\right) \tag{1}$$

The mean firing rate was then obtained by applying a simple non-linear mapping:

$$r_t = f\left(a s_t + b\right) \tag{2}$$

where a and b are free parameters and $f(\cdot)$ was assumed to be a softplus ($\log(1+e^x)$) non-linear function to prevent firing rates being negative.

The computed 'surprise' for each cell thus depends on their expectations or 'internal model' of the stimulus statistics (parameterized by θ). We first assumed the simplest possible internal model: a 'Markov model', in which the probability of observing a flash, $x_t = 1$, only depends on whether there was a flash or not in the previous time bin ($x_t = 0/1$). This binary Markov model has two free parameters: the probability of a flash occurring if there was/wasn't a flash in the previous time-step ($\theta_0 = p\left(x_t = 1 | x_{t-1} = 0\right)$, and $\theta_1 = p\left(x_t = 1 | x_{t-1} = 1\right)$). The parameters of the response function (a and b) and internal model (θ) were fitted for each neuron using maximum likelihood, assuming that the responses were generated by a Poisson distribution with mean r_t (see Methods: Neural model).

Fig. 2B shows the average firing rate of a single neuron (black) for a given stimulus sequence (above) (see Methods: Data analysis). This model accounted for the most prominent feature of the neuron's responses: that it responded strongly to the first flash in a sequence, and the first silence in a sequence (i.e. the OSR). However, the model was unable to replicate the dependence of the OSR on the number of flashes presented in a row, observed for this (Fig. 2C) and many other cells (Fig. 2D). This was because, by design, with a Markov model the computed surprise, only depends on the stimulus in the previous time-bin, and thus the predicted response is also independent of the stimulus history, beyond one time-bin (Fig. 2E).

Adaptive surprise model

To account for the observed variations in the OSR with number of consecutive flashes, we next considered a more complex 'dynamic belief' internal model. Here, we assume that the transition probabilities ($\theta_i \equiv p\left(x_t = 1 | x_{t-1} = i\right)$), are not known *a priori* by each neuron, but must be inferred. We assume neurons combine their prior expectations ($p\left(\theta_i\right)$) with the recent stimulus history ($p\left(x_t, x_{t-1}, \ldots | \theta\right)$) using Bayes' law: $p\left(\theta | x_t, x_{t-1}\right) \propto p\left(x_t, x_{t-1}, \ldots | \theta_i\right) p\left(\theta_i\right)$. We assumed a beta-distribution for the prior over θ_i, with parameters α_i and β_i. This results in a simple expression for the inferred probability of observing $x_t = 1$, given $x_{t-1} = i$:

$$p\left(x_t = 1 | x_{t-1} = i, x_{t-2}, \ldots\right) = \frac{n_{i\to 1} + \alpha_i}{n_{i\to 0} + n_{i\to 1} + \beta_i + \alpha_i}, \tag{3}$$

where $n_{i\to j}$ is the number of occurrences of the transition $i \to j$ in the sequence $\{x_1, x_2, \ldots, x_t\}$, and α_i and β_i are parameters of the prior. We assume that the parameters of the prior (α_i, β_i) are different for each neuron. Note that, in the limit where the prior is very strong (i.e. $n_{i\to j} \ll \alpha_i$ and $n_{i\to j} \ll \beta_i$), this model becomes identical to the 'fixed-belief' model described in the previous section, where the transition probabilities for each neuron are stimulus-independent.

If neurons had 'infinite' memory then, given a sufficiently long stimulus sequence, their prior expectations would have no effect. Instead, we assume a more biologically plausible model where neuron's have a finite memory, and $n_{j\to i}$ are estimated using a leaky integration of past observations (see Methods: Adaptive surprise model). This requires one additional parameter (the time-scale of integration i.e. the leak parameter), which we kept fixed for all neurons. In Supplementary section Dynamic surprise model we show how qualitatively similar results can be obtained by assuming neurons perform Bayesian inference, given a model where the transition probabilities have a small probability of changing on each time-step. However, optimal Bayesian inference in this setting required complex numerical integration, and it is thus hard to see how it could be implemented feasibly by individual neurons. As a result, we focus on the simpler 'leaky integration' model for the rest of the paper.

We fitted the 4 parameters of the prior (plus the bias and gain of the LN model) for each neuron, using maximum likelihood, assuming Poisson noise [Pillow et al., 2005]. Fig. 3A shows the predicted firing rate for one neuron (blue) to a short stimulus sequence (above). The 'adaptive surprise' model was able to capture aspects of the neuron's response that could not be accounted for by the previous 'fixed surprise model'. For example, it could capture how the size of the OSR increased with the number of flashes presented in a row (Fig. 3B-C). Further, it captured individual differences in the OSR decay for different neurons (compare cells 1-3 in Fig. 3B). Overall, the correlation between the estimated firing rates and the model prediction was significantly higher for the adaptive, compared to the fixed, surprise

model (Fig. 3D).

To further investigate how the adaptive surprise model could account for the diverse responses of different cells, we plotted tree-plots showing the average firing rates predicted by the model for stimulus sequences of different lengths (Fig. 4A). The adaptive belief model captured much of the structure in the neural responses to stimulus sequences of varying length, as well as the diversity across different cells. This was supported by plotting the correlation coefficient between the model predictions for each node of the tree and the data, which decayed slowly with the tree depth (Fig. 4B), compared to the fixed belief model which reduced dramatically for tree depth greater than 2.

To further test our adaptive belief model, we compared it to a more complex fixed belief model, with a comparable number of free parameters. To do this, we implemented a 'Markov-2 model' in which the probability of observing a flash is depends on the observed stimulus in the previous two time-bins. This model's prior has 4 parameters, ($\theta_{ij} = p(x_{t+1} = 1|x_t = i, x_{t-1} = j)$), which is the same as the adaptive surprise model (aside from the leak parameter, which we kept the same for all cells). The behaviour of this model is shown in Fig. 5. While the Markov-2 model outperformed the fixed surprise model model described earlier, it could not account for increases in the OSR that occurred for sequences of more than 2 consecutive flashes (Fig. 5B-C), or any structure in the tree plots at a depth greater than 2 (Fig. 5D, emphasized with a dashed ellipse). Finally, the correlation coefficient between predicted and observed firing rates was significantly worse for the Markov-2 model than the adaptive surprise model (Figure. 5E) despite them having the same number of free parameters ($p = 2 \cdot 10^{-8}$, Wilcoxon signed-rank test).

Differences in the internal expectations for individual cells

We were interested to see how the inferred expectations (the 'prior') varied for each cell. Recall that we assumed a beta-prior over the transition probabilities $\theta_i = p\left(x_t = 1|x_{t-1} = i\right)$, with parameters α_i and β_i. The mean of this prior is determined by the ratio of these two parameters, α_i/β_i, while its width (i.e. the level of prior uncertainty) is determined by their sum, $\alpha_i + \beta_i$. Figure 6A shows how the parameters of the prior varied for different cells. Interestingly, we found that while for different cells there was a large variation in the sum, $\alpha_i + \beta_i$, the ratio, α_i/β_i, was relatively constant. Thus, while the width of the prior, which determines how much weight is accorded to prior expectations versus new observations, varied greatly across cells, the prior mean was roughly constant.

Focusing on the prior parameters, α_1 and β_1, which determines neural responses to 'flash→flash' and 'flash→no-flash' transitions (i.e. the OSR), we observed two clusters of cells (Fig. 6A, right panel), with different levels of prior uncertainty (determined by the sum, $\alpha_1 + \beta_1$; Fig. 6B). We asked what effect

this would have on these cells responses. We reasoned that cells with a strong prior (i.e. large $\alpha_i + \beta_i$) would not adapt their posterior belief much depending on recent observations, and hence their responses would be well predicted by the fixed surprise model. In contrast, cells with a weak prior (i.e. small $\alpha_i + \beta_i$) would be strongly influenced by recent observations, and thus their responses would be poorly predicted by the fixed-surprise model. This turned out to be the case. Fig. 6C shows the correlation coefficient between the prediction of the fixed surprise model and recorded responses, versus the adaptive surprise model. There was a trend for cells with a strong prior (high $\alpha_i + \beta_i$; colour coded in yellow) to be equally well-fit by both models, while cells with a weak prior (low $\alpha_i + \beta_i$; colour coded in blue) were better fit by the adaptive surprise model. We asked whether this same effect could be observed without reference to the model fits. To do this, we compared the average neural responses to stimulus sequences of length 2, to the average response to longer sequences, of length 10. As expected, we found that the average responses of cells with a strong prior (i.e. $\log(\alpha_1 + \beta_1) > 7$) only depended on the most recently presented stimuli (Fig. 6D, yellow). This tended not to be the case for stimuli with a weak prior (Fig. 6D, blue) ($p = 0.066$, Wilcoxon rank-sum test).

In Fig. 6A we observed that the prior mean, determined by α_i/β_i, remained near-unity across different cells. We thus, asked whether it would be possible to fit neural responses using a reduced adaptive surprise model with only two parameters (i.e. the sum, $\alpha_i + \beta_i$), and α_i/β_i held fixed at unity. Fig. 7A shows that, while this reduced adaptive model performed worse than the full adaptive surprise model, this reduction in performance was small ($< 10\%$ reduction in correlation coefficient), despite having having only 2 free parameters per cell (compared to 4 parameters, for the full model). Notably, the reduced model was able to capture similar qualitative features of neural responses, such as how the OSR increased with the number of consecutive flashes (Fig. 7B, Supp. Fig. 5). Its performance was also significantly better than the fixed surprise model, which had the same number of free parameters per cell ($p = 4 \cdot 10^{-5}$, Wilcoxon signed-rank test).

Discussion

We observed how neural responses in the retina showed non-trivial dependencies on the precise order of flashes and silences in random stimulus sequences (Fig. 1C-D). Interestingly, RGC responses were well predicted by a simple model, which assumed that they depended on how 'surprising' stimuli were, relative to an internally generated expectation (Fig 3-4). Moreover, our model showed how the different 'expectations' of different neurons could account for the diverse way they responded to presented stimuli (Fig 6-7).

Our approach contrasts with previous ideal observer models, which assume that neurons are

perfectly adapted to the 'true' presented stimulus statistics [Geisler, 1989, Geisler, 2003, Smeds et al., 2019, Chichilnisky and Rieke, 2005]. Instead, we found that neural responses could be well explained by assuming that each neuron has learned its own internal model of the stimulus statistics (with the parameters of the prior fitted separately for each cell). Interestingly, we found that different neurons had very similar prior expectations about which stimuli were most likely to occur (determined by the mean of the prior). What varied was the degree of confidence they had about their own prior expectations (determined by the width of the prior). Furthermore, recorded cells could be divided into two categories: those with weak confidence in their prior, and those with strong confidence in their prior expectations. In the future, it would be interesting to elucidate the reason for this split, and whether, for example, it corresponded to different types of ganglion cell identified in previous work [Baden et al., 2016].

Our modelling framework was adapted from a previous model of Meyniel et al., that sought to explain psychophysical data showing how subjects' behaviour (such as their reaction time and accuracy) depended on the statistics of sequentially presented sensory stimuli [Meyniel et al., 2016]. Meyniel and colleagues showed how their data could by explained if subjects used a Bayesian inference model, as described here, to predict new stimuli based on what came before. Here, we extended this model to include a variable 'prior' distribution, whose parameters could be fit to describe the diverse responses of different ganglion cells. Nonetheless, the fact that a similar type of model can be used to describe both neural responses in the retina and subjects behaviour in different tasks is intriguing, raising the question of whether similar computations may be present ubiquitously in the brain when subjects are presented stimuli with complex temporal statistics.

Previous experimental [Schwartz and Berry 2nd, 2008, Werner et al., 2008, Deshmukh, 2015] and computational [Maheswaranathan et al., 2019, Tanaka et al., 2019, Chen et al., 2017] studies sought to understand the neural mechanisms underlying the OSR. However, there remains some controversy over which of the proposed theories could explain all of the experimentally observed features of the OSR, such as e.g. the fact that the delay before the OSR varies linearly with the time between flashes. Our work provides further constraints to distinguish between different theories, by showing how the OSR varies depending on the precise sequence of flashes and silences (Fig 1).

The stimuli in our experiment, which consisted of sequences of full-field flashes, were chosen to be sufficiently rich so as to permit many different levels of 'surprise', while simple enough to permit a straight-forward analysis of neural responses. Nonetheless, in the future it would be interesting to investigate neural responses to more naturalistic stimuli, which for example, varied spatially as well temporally [Keller et al., 2012]. This would allow us to investigate, for example, the degree to which neurons' internal model is adapted to the statistics of natural scenes, as predicted by the efficient coding hypothesis [Machens et al., 2005].

Methods

Experimental setup

The recordings were performed in the axolotl retina, using a multi-electrode array with 252 electrodes with 60 μm spacing (procedure described in detail in [Marre et al., 2012]). The experiment was performed in accordance with institutional animal care standards of Sorbonne Université. The raw signal, recorded at 20 kHz sampling rate, was high-pass filtered at 100 Hz and then sorted offline using SpyKing Circus software [Yger et al., 2016]. The stimulus consisted of full-field dark flashes. The reason for using dark flashes was the dominance of OFF type cells in axolotl. The dark flashes had a duration of 40 ms, with 80 ms period between the flashes, ($\sim$12 Hz frequency) (as in [Schwartz et al., 2007]).

Stimulus statistics

We generated sequences of flashes and silences (i.e. where no flash occured in a 120ms window) using a stochastic model. The number of flashes and periods of silent states presented in a row was drawn from a negative binomial distribution, with parameters r and p. In the case of flashes, we varied the first parameter, p, at 20 minute intervals between three different values (0.98, 0.8 and 0.01, consecutively). The second parameter, r was adjusted so as to maintain a constant mean, of 7 flashes presented in a row. The length of the silence sequence was drawn from a geometrical distribution with a fixed mean p ($p = 9$). Changing p alters the degree to which the distribution is clustered around the mean. However, we observed no difference in the neural responses recorded with different values of p. As a result we concatenated data from neural responses to all three stimulus distributions for the rest of our analysis.

Data analysis

To generate the spike raster plots shown in Fig. 1, we aligned the spiking responses of neurons to a sequence of n flashes presented in a row. The peri-stimulus-time-histogram (PSTH) plotted in the bottom row of Fig. 1 was computed by averaging the spike count recording over all the stimulus repeats, and then averaging over a 5 ms time bin.

For the remainder of the analysis, we discretised the neural responses and stimulus into time bins of length 120 ms (the time between consecutive flashes). The stimulus presented in each time-bin was treated as a binary variable: '1' if there was a (dark) flash, '0' otherwise. The average firing rate in each bin was computed by average the spike count over all repetitions of a stimulus sequence of length n. Except where stated explicitly in the text, we set $n = 8$ (so there were 256 distinct stimulus sequences used to compute the average firing rate).

Neural model

For our model, we assumed that at each time-bin, t, neurons fire spikes drawn from a Poisson distribution with mean, λ_t, given by:

$$\lambda_t = f(as_t + b) \tag{4}$$

where s_t is the encoded surprise at time t, f is a non-linearity, and a and b are parameters describing the gain and bias, respectively. The non-linearity, $f(x) = \log(1 + e^x)$ (soft-ReLU), was kept fixed for all the cells.

The surprise at time t is defined as:

$$s_t = -\log p(x_t|x_{t-1}, x_{t-2}, \ldots, \theta) \tag{5}$$

where $p\left(x_t|x_t, x_{t-2}, \ldots, \theta\right)$ is the probability of observing no flash or a flash at time t ($x_t = 0$ or 1 respectively) given the stimulus x at previous times, and the internal model of the cell, parameterized by θ.

Internal model

The computed surprise depends on each cell's internal model of the stimulus statistics. We first considered a binary Markov model, where the probability of observing a flash at time t is assumed to depend only on whether a flash was observed in the previous time bin. This model has two parameters: $\theta_0 = p\left(x_t = 1|x_{t-1} = 0\right)$, and $\theta_1 = p\left(x_t = 1|x_{t-1} = 1\right)$. For the Markov 2 model, we simply extend the observed history to 2 previous states, yielding a total of 4 parameters: $\theta_0 = p\left(x_t = 1|x_{t-1} = 0, x_{t-2} = 0\right)$, $\theta_1 = p\left(x_t = 1|x_{t-1} = 1, x_{t-2} = 0\right)$, $\theta_2 = p\left(x_t = 1|x_{t-1} = 0, x_{t-2} = 1\right)$, and $\theta_3 = p\left(x_t = 1|x_{t-1} = 1, x_{t-2} = 1\right)$.

Inferring the transition probabilities

Next, we considered an 'adaptive belief model' where the transition probabilities, $\theta_i = p\left(x_t|x_{t-1} = i\right)$, are not known in advance, but must be inferred by combining each cell's prior belief with newly observations, using Bayes' law, as follows:

$$p\left(\theta|x_t, x_{t-1}, \ldots\right) \propto p\left(x_t|x_{t-1}, \theta\right) p\left(\theta|x_{t-1}, x_{t-2}, \ldots\right) \tag{6}$$

where $p\left(x_t|x_{t-1}, \theta\right)$ is the likelihood of observing x_t given x_{t-1}, and is described by a Bernoulli distribution:

$$p\left(x_t|x_{t-1} = i, \theta_i\right) = \theta_i^{x_t}\left(1 - \theta_i\right)^{1-x_t}. \tag{7}$$

Let us assume that at time $t-1$, the posterior distribution over θ_i, $p\left(\theta|x_{t-1}, x_{t-2}, \ldots\right)$, is described by the beta distribution with parameters α_i^{t-1} and β_i^{t-1}:

$$p\left(\theta_i|x_{t-1}, x_{t-2}, \ldots\right) \propto \theta^{\alpha_{t-1}^i - 1}\left(1-\theta\right)^{\beta_{t-1}^i - 1}. \tag{8}$$

Now, at time t, multiplying this distribution by the likelihood according to Bayes law (Eqn 6) will result in a new beta-distribution, with parameters:

$$\alpha_t^i \leftarrow \alpha_t^i + x_t x_{t-1}^i \left(1-x_{t-1}\right)^{1-i} \tag{9}$$

$$\beta_t^i \leftarrow \beta_t^i + (1-x_t) x_{t-1}^i \left(1-x_{t-1}\right)^{1-i} \tag{10}$$

The probability of observing $x_t = 1$ given previous observations is then given by:

$$p\left(x_t = 1|x_{t-1} = i, x_{t-2}, x_{t-3}, \ldots\right) = \int_\theta p(x_t = 1|x_{t-1} = i, \theta_i) p\left(\theta_i|x_{t-1}, x_{t-2}, \ldots\right) \tag{11}$$

$$= \frac{\alpha_{t-1}^i}{\alpha_{t-1}^i + \beta_{t-1}^i} \tag{12}$$

$$= \frac{n_{i\to 1}}{n_{i\to 1} + n_{i\to 0}}, \tag{13}$$

where $n_{i\to 0}$ and $n_{i\to 1}$ describe the number of occurrences of the transitions $i \to 0$ and $i \to 1$, respectively.

Adaptive surprise model

The statistics of the external world are not static, but change in time. To take this into account, we could assume there a non-zero probability of transition matrix changing between two observations (a 'dynamic belief model'). Performing exact Bayesian inference in this case requires expensive numerical integration, which may be difficult to perform by individual neurons in the retina. However, in [Meyniel et al., 2016] they found that such a dynamic belief model could be approximated by a 'forgetful' model, where recent observations are weighted more strongly than the ones in the past. In contrast to the optimal Bayesian model, their leaky integration model results in simple linear parameter updates, and could thus be easy to implement neurally.

In practice, we can implement the 'forgetful' model of Meyniel et al., by modifying the update rules described earlier for α^i and β^i as follows:

$$\alpha_t^i \leftarrow (1-\eta)\,\alpha_t^i + \eta\alpha_0^i + x_t x_{t-1}^i \left(1-x_{t-1}\right)^{1-i} \tag{14}$$

$$\beta_t^i \leftarrow (1-\eta)\,\beta_t^i + \eta\beta_0^i + (1-x_t) x_{t-1}^i \left(1-x_{t-1}\right)^{1-i} \tag{15}$$

where $1 > \eta > 0$ is a leak term that results in forgetting observations far in the past, while α_0^i and β_0^i determine the steady state values of α_t^i and β_t^i in the absence of new observations. With this update rule, the probability of observing a $x_t = 1$ given $x_{t-1} = i$ is given by:

$$p\left(x_t = 1 | x_{t-1} = i, x_{t-2}, x_{t-3}, \ldots\right) = \frac{\tilde{n}_{i\to 1} + \alpha_0^i}{\tilde{n}_{i\to 1} + \tilde{n}_{i\to 0} + \beta_0^i}, \tag{16}$$

where $\tilde{n}_{i\to j}$ is the 'effective' number of observations of a transition $i \to j$, after taking into account the leak, when $\eta > 0$:

$$\tilde{n}_{i\to j} = \sum_{k=0}^{\infty} (1-\eta)^k x_{t-k}^j x_{t-1-k}^i \left(1 - x_{t-k}\right)^{(1-j)} \left(1 - x_{t-1-k}\right)^{(1-i)}. \tag{17}$$

In practice we assumed that the leak, η was the same for all cells. We used a value of $\eta = 0.2$. However, similar results were obtained when we increased or decreased the leak by a small amount.

In contrast, the parameters of the prior, α_0^i and β_0^i, were allowed to vary for different cells. This us allowed to investigate how different cells' 'prior expectations' for different transitions affected their responses. (Note that for notational simplicity we dropped the subscript '0' in the main text.)

Model fitting

We fitted the internal model parameters (see previous section), and the gain and bias of the response curves (Eqn 4) using Maximum Likelihood (ML) algorithm [Doya et al., 2007]. For this, we assumed a Poisson noise model, resulting in a log-likelihood:

$$\mathcal{L} = \sum_t n_t \log f_t - f_t \tag{18}$$

where f_t and n_t are the spike count predicted by the model and observed spike count at time t, respectively. All models were fitted using algorithms with multiple starting points (MultiStart in MATLAB, 50 starting points, random initial parameters).

The data analysis and model fitting were done in MATLAB R2021a. Code and data will be available upon paper acceptance.

Acknowledgments

This work was funded by ANR JCJC grant Optimal predictive coding. The authors would like to thank Yannick Andéol for providing the axolotls. We would also like to thank Ulisse Ferrari, Francesco Trapani, Matías Goldin and Samuele Virgili for useful discussions.

References

[Baden et al., 2016] Baden, T., Berens, P., Franke, K., Rosón, M. R., Bethge, M., and Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350.

[Chen et al., 2017] Chen, K. S., Chen, C.-C., and Chan, C. (2017). Characterization of predictive behavior of a retina by mutual information. *Frontiers in computational neuroscience*, 11:66.

[Chichilnisky and Rieke, 2005] Chichilnisky, E. and Rieke, F. (2005). Detection sensitivity and temporal resolution of visual signals near absolute threshold in the salamander retina. *Journal of Neuroscience*, 25(2):318–330.

[Deshmukh, 2015] Deshmukh, N. R. (2015). *Complex computation in the retina*. PhD thesis, Princeton University.

[Doi et al., 2012] Doi, E., Gauthier, J. L., Field, G. D., Shlens, J., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., Mathieson, K., Gunning, D. E., et al. (2012). Efficient coding of spatial information in the primate retina. *Journal of Neuroscience*, 32(46):16256–16264.

[Doya et al., 2007] Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

[Geisler, 1989] Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological review*, 96(2):267.

[Geisler, 2003] Geisler, W. S. (2003). Ideal observer analysis. *The visual neurosciences*, 10(7):12–12.

[Gill et al., 2008] Gill, P., Woolley, S. M., Fremouw, T., and Theunissen, F. E. (2008). What's that sound? auditory area clm encodes stimulus surprise, not intensity or intensity changes. *Journal of neurophysiology*, 99(6):2809–2820.

[Karklin and Simoncelli, 2011] Karklin, Y. and Simoncelli, E. (2011). Efficient coding of natural images with a population of noisy linear-nonlinear neurons. *Advances in neural information processing systems*, 24.

[Keller et al., 2012] Keller, G. B., Bonhoeffer, T., and Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815.

[Lewicki, 2002] Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363.

[Machens et al., 2005] Machens, C. K., Gollisch, T., Kolesnikova, O., and Herz, A. V. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3):447–456.

[MacKay et al., 2003] MacKay, D. J. et al. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Maheswaranathan et al., 2019] Maheswaranathan, N., McIntosh, L. T., Tanaka, H., Grant, S., Kastner, D. B., Melander, J. B., Nayebi, A., Brezovec, L., Wang, J., Ganguli, S., et al. (2019). The dynamic neural code of the retina for natural scenes. *BioRxiv*, page 340943.

[Marre et al., 2012] Marre, O., Amodei, D., Deshmukh, N., Sadeghi, K., Soo, F., Holy, T. E., and Berry, M. J. (2012). Mapping a complete neural population in the retina. *Journal of Neuroscience*, 32(43):14859–14873.

[Meyniel et al., 2016] Meyniel, F., Maheu, M., and Dehaene, S. (2016). Human inferences about sequences: A minimal transition probability model. *PLoS computational biology*, 12(12):e1005260.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

[Pillow et al., 2005] Pillow, J. W., Paninski, L., Uzzell, V. J., Simoncelli, E. P., and Chichilnisky, E. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *Journal of Neuroscience*, 25(47):11003–11013.

[Rao and Ballard, 1999] Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79.

[Schwartz and Berry 2nd, 2008] Schwartz, G. and Berry 2nd, M. J. (2008). Sophisticated temporal pattern recognition in retinal ganglion cells. *Journal of neurophysiology*, 99(4):1787–1798.

[Schwartz et al., 2007] Schwartz, G., Harris, R., Shrom, D., and Berry, M. J. (2007). Detection and prediction of periodic patterns by the retina. *Nature neuroscience*, 10(5):552–554.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

[Smeds et al., 2019] Smeds, L., Takeshita, D., Turunen, T., Tiihonen, J., Westö, J., Martyniuk, N., Seppänen, A., and Ala-Laurila, P. (2019). Paradoxical rules of spike train decoding revealed at the sensitivity limit of vision. *Neuron*, 104(3):576–587.

[Smith and Lewicki, 2006] Smith, E. C. and Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, 439(7079):978–982.

[Soto et al., 2020] Soto, F., Hsiang, J.-C., Rajagopal, R., Piggott, K., Harocopos, G. J., Couch, S. M., Custer, P., Morgan, J. L., and Kerschensteiner, D. (2020). Efficient coding by midget and parasol ganglion cells in the human retina. *Neuron*, 107(4):656–666.

[Tanaka et al., 2019] Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S., and Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in neural information processing systems*, 32.

[Ulanovsky et al., 2003] Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature neuroscience*, 6(4):391–398.

[Van Hateren and van der Schaaf, 1998] Van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366.

[Werner et al., 2008] Werner, B., Cook, P. B., and Passaglia, C. L. (2008). Complex temporal response patterns with a simple retinal circuit. *Journal of neurophysiology*, 100(2):1087–1097.

[Yger et al., 2016] Yger, P., Spampinato, G. L., Esposito, E., Lefebvre, B., Deny, S., Gardella, C., Stimberg, M., Jetter, F., Zeck, G., Picaud, S., et al. (2016). Fast and accurate spike sorting in vitro and in vivo for up to thousands of electrodes. *BioRxiv*, page 067843.

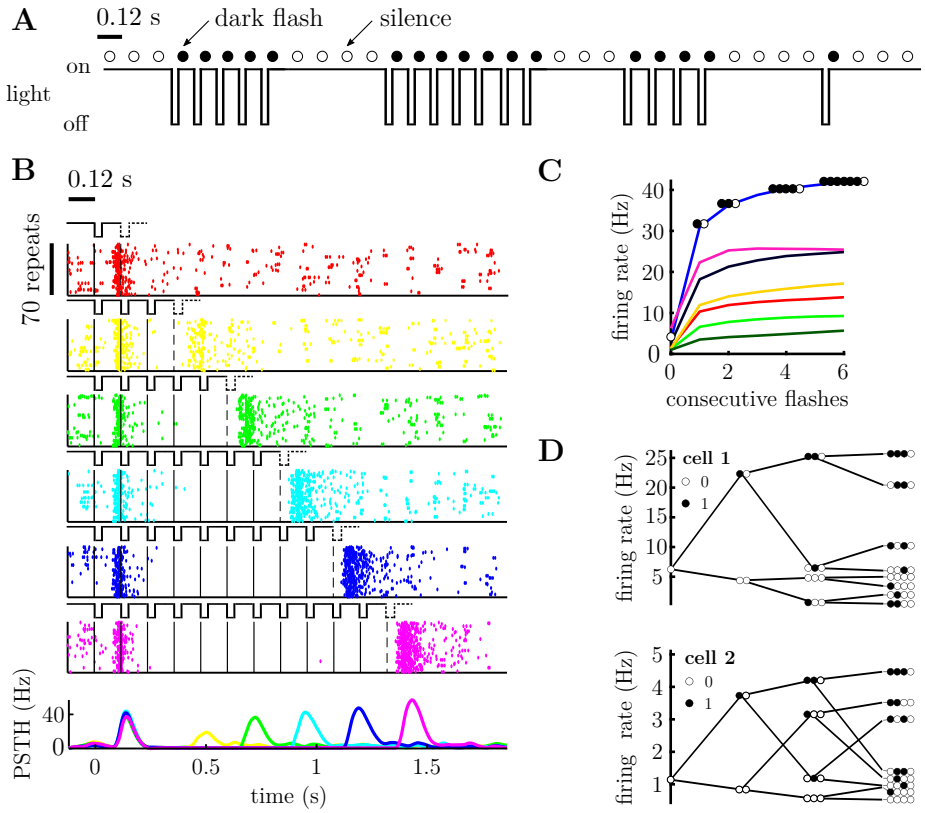

Figure 1: **Retinal ganglion cells' responses to a sequences of dark flashes and silences. A.** Stimulus excerpt, showing periodic sequences of dark flashes. Each flash lasts 40 ms with 80 ms between. The 120 ms bin containing a single dark flash is marked with a filled circle. A bin without a flash (called a 'silence') is marked by an open circle. **B.** Raster plot for one cell. A solid vertical line marks the occurrence of each flash; a dashed line indicates an 'omitted' flash, following a sequence of flashes. Each raster plot shows the cell's response to a different number of consecutive flashes (ranging from 1 to 11), with 70 repeats shown in each row of the raster. The bottom row shows the peri-stimulus time histogram (PSTH) for different numbers of consecutive flashes (colors denote the number of flashes). There is an increase in firing rate after the missing flash, called the omitted stimulus response (OSR). The OSR magnitude increases with the number of flashes. **C.** OSR for 7 cells, following a varying number of consecutive flashes (filled circles) followed by silence (open circles). **D.** Tree-plot, showing the mean response of two representative cells to different sequences of flashes (filled circles) and silences (empty circles). Cell 1 is the cell plotted in pink in panel C. Each column of the tree-plot shows the average response of the neuron to all stimulus sequences of a given length that end with silence (tree-plots corresponding to sequences ending with a flash are shown in Supp. Fig. 1). Moving right-ward the tree-plot branches out to include the effect of stimuli presented further in the past. The top branch of the tree-plot shows the cells' responses to a series of consecutive flashes followed by silence, as in panel C. Other branches show the cells' responses to all the different possible flash sequences of a given length.

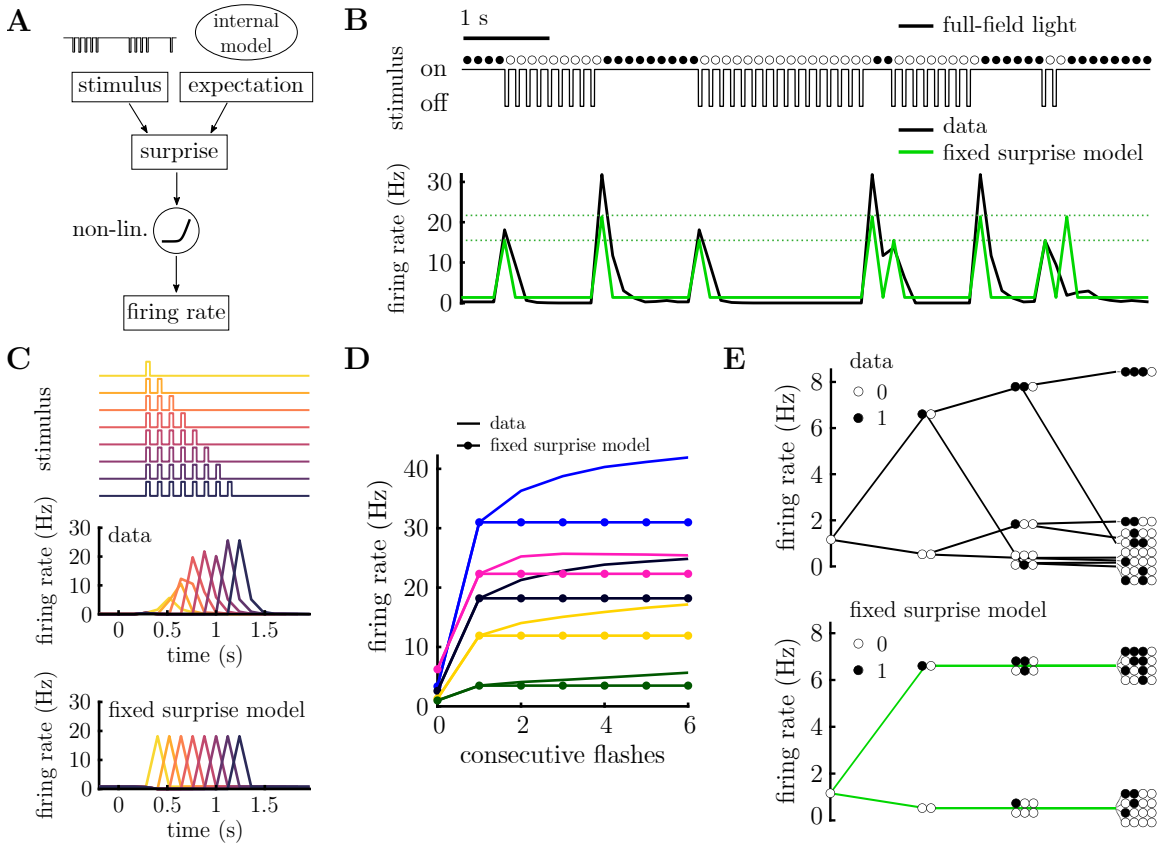

Figure 2: **Fixed surprise model. A.** Schematic of modeling framework. The stimulus is compared to the neuron's expectation, which depends on their internal model, to compute surprise. The encoded surprise is then transformed via a static non-linearity to obtain the neuron's firing rate. **B.** Stimulus excerpt (above) and recorded PSTH (below, black), and prediction of the fixed surprise model (below, green). The fixed surprise model has limited flexibility, only permitting four possible firing rates (indicated with dashed lines). **C.** Response to flash sequences of varying length (top). PSTH for a single neuron (middle) and model prediction (bottom) to the stimulus sequences shown above. Each colour corresponds to a different length of flash sequence. The fixed surprise model predicts the OSR magnitude to be independent of the number of flashes. **D.** OSR for 5 cells (solid lines) after a varying number of consecutive flashes. The fixed surprise model (lines with filled circles) cannot account for the increase in the OSR with increasing number of flashes. **E.** Tree-plot for a single cell (above) and fixed model prediction (below). The fixed surprise model can capture the mean response for stimulus sequences of up to length 2, but not beyond. Tree-plots corresponding to sequences ending with a flash are shown in Supp. Fig. 2.

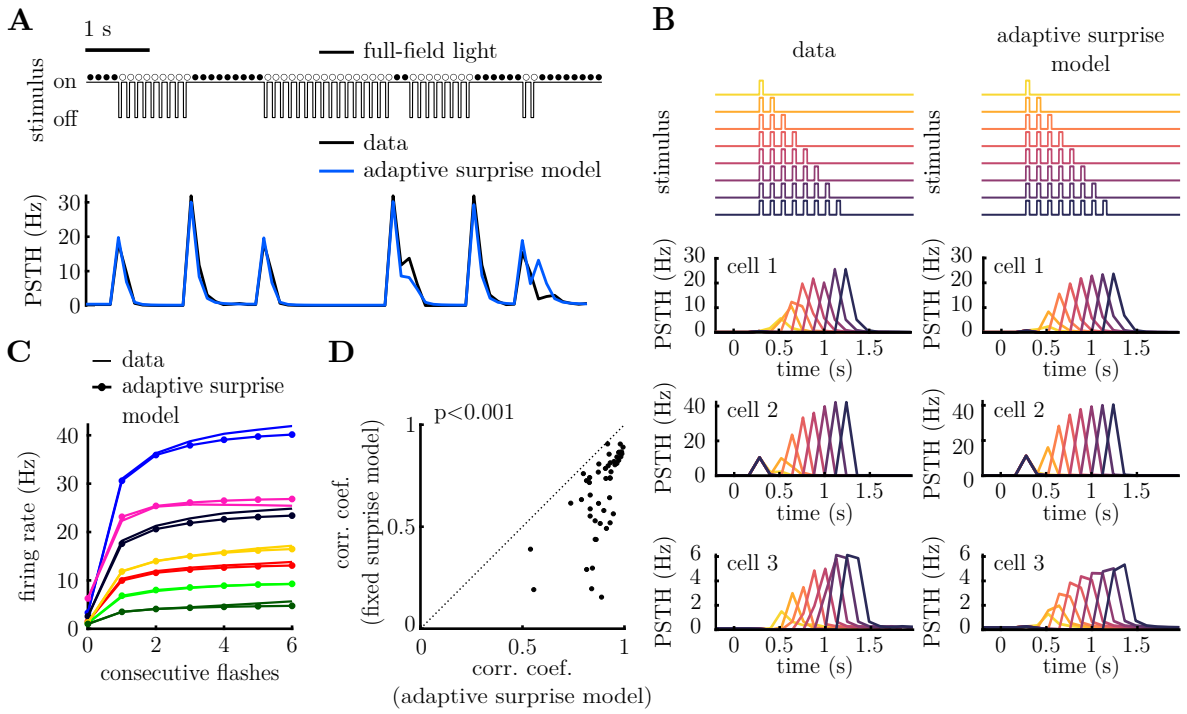

Figure 3: **Adaptive surprise model. A.** Stimulus excerpt (above) and recorded PSTH (below, black), alongside prediction of the adaptive surprise model (below, blue). **B.** Neural responses to varying number of consecutive flashes (above). Recorded PSTH of three neurons is shown to the left, while model predictions are shown to the right. Each colour corresponds to a different number of consecutive flashes. The adaptive surprise model captures variations in both the magnitude and width of the OSR. **C.** Increase in the OSR with the number of consecutive flashes for seven cells (each cell plotted with a different colour). The data (solid line) is plotted alongside the predictions of the adaptive surprise model (solid lines with circles). **D.** Pearson correlation coefficients between each cell's PSTH and the model predictions, for the fixed surprise model (y-axis) versus the adaptive surprise model (x-axis). The adaptive surprise model significantly outperforms the fixed surprise model ($p = 1 \cdot 10^{-9}$, Wilcoxon signed-rank test).

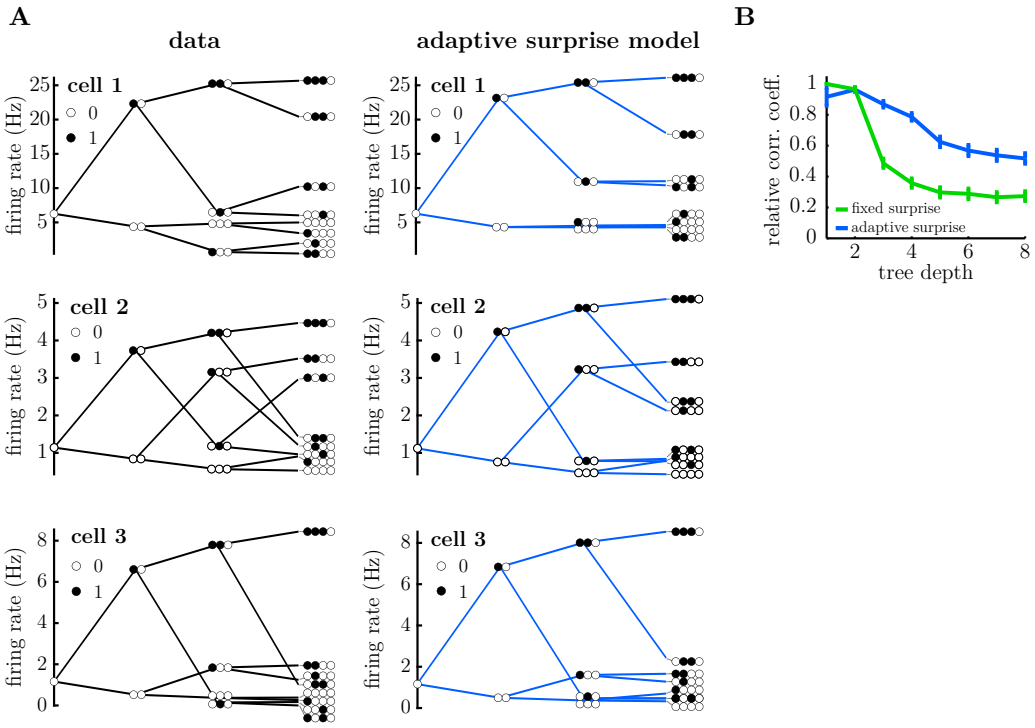

Figure 4: **Neural responses to different stimulus sequences, for the adaptive surprise model. A.** Tree-plot, showing the mean response of three representative cells to all possible sequences of flashes (filled circles) and silences (empty circles) of a given length. As we move rightward, the tree branches to show responses to take into account stimuli presented further in the past. The data is shown on the left and the model predictions on the right. The adaptive model is a able to reproduce qualitative aspects of each tree-plot, beyond the top branch (which shows how the OSR magnitude varies with the number of consecutive flashes). Tree-plots corresponding to sequences ending with a flash are shown in Supp. Fig. 3. **B.** Correlation coefficient between the tree-plot obtained with the adaptive surprise model and the data, computed separately for each tree-depth (i.e. stimulus sequence length). The adaptive model is significantly better at capturing the shape of the tree for stimulus sequences longer than 2, compared to the fixed surprise model.

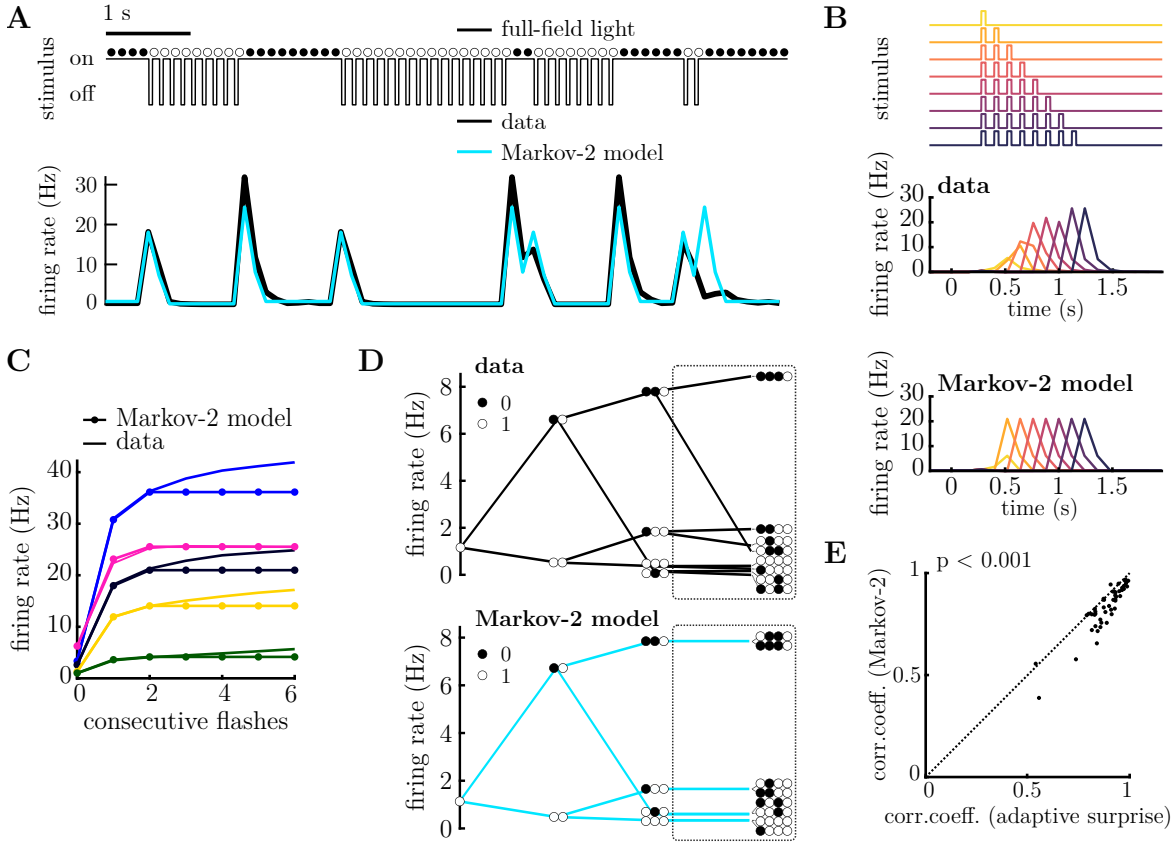

Figure 5: **Fixed surprise model with longer past (Markov-2 model). A.** Stimulus excerpt (above) and recorded PSTH (below, black), and firing rate predicted by the Markov-2 model (below, blue). **B.** Response to varying number of consecutive flashes (top). PSTH for a single neuron (middle) and model prediction (below) to the stimulus sequences shown above. Each colour corresponds to a different length of flash sequence. The Markov-2 surprise model predicts the OSR magnitude to be dependent on the previous two state only. **C.** Average responses of 5 cells (solid lines) to flash sequences of varying lengths. The Markov-2 surprise model (lines with filled circles) cannot account for the increase in the OSR beyond 2 consecutive flashes. **D.** Tree-plot for a single cell (above) and fixed model prediction (bottom). The Markov-2 surprise model cannot capture the response for stimulus sequences greater than length 2 (highlighted with dashed circle). Tree-plots corresponding to sequences ending with a flash are shown in Supp. Fig. 4. **E.** The correlation coefficient between the adaptive surprise model and recorded PSTH for each cell is significantly better than for the Markov-2 model ($p = 2 \cdot 10^{-8}$, Wilcoxon signed-rank test).

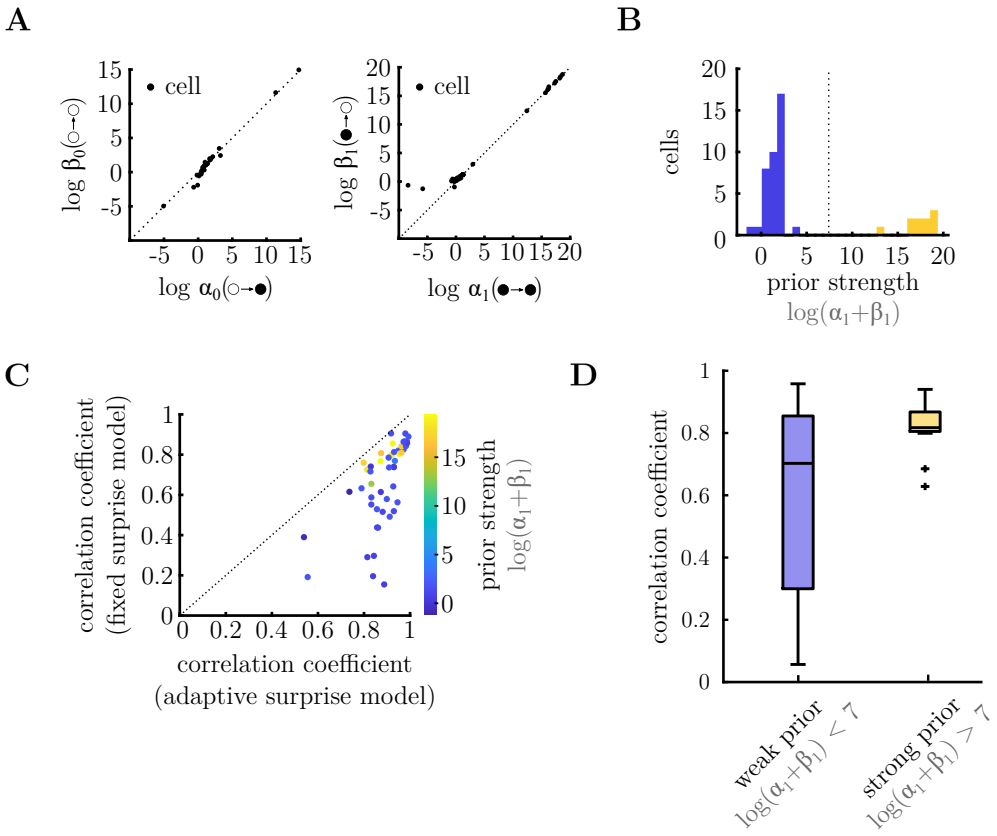

Figure 6: **Parameters of internal model. A.** Parameters of the inferred prior (α_i and β_i) for each cell. These parameters determine each cell's prior expectation for the different transitions from silence (left) or flash (right). In both cases, the ratio between these parameters, α_i/β_i (which determine the mean of the prior) is close to unity for all the cells, while their sum, $\alpha_i + \beta_i$ (which determines the strength of the prior) varies across different cells. **B.** Histogram of $\log(\alpha_1 + \beta_1)$ for different cells. The population could be split into two groups: cells with low confidence in the prior (i.e. small $\alpha_i + \beta_i$; blue) and cells with high confidence in the prior (i.e. large $\alpha_i + \beta_i$; yellow). **C.** Correlation coefficient between the responses predicted by the adaptive surprise model, versus the fixed surprise model. Each circle is colour coded according to the parameters of the inferred prior for that cell $\log(\alpha_1 + \beta_1)$. Cells that had a strong prior (yellow) tended to be better fit by the fixed surprise model, relative to the adaptive surprise model. **D.** For each cell we computed the correlation coefficient between the average neural responses to stimulus sequences of length 2, versus average responses that take into account stimulus sequences of length 10. The responses of cells with a strong prior (right, yellow), but not a weak prior (left, blue), could be reasonably well predicted just by looking at the most recent stimulus transition.

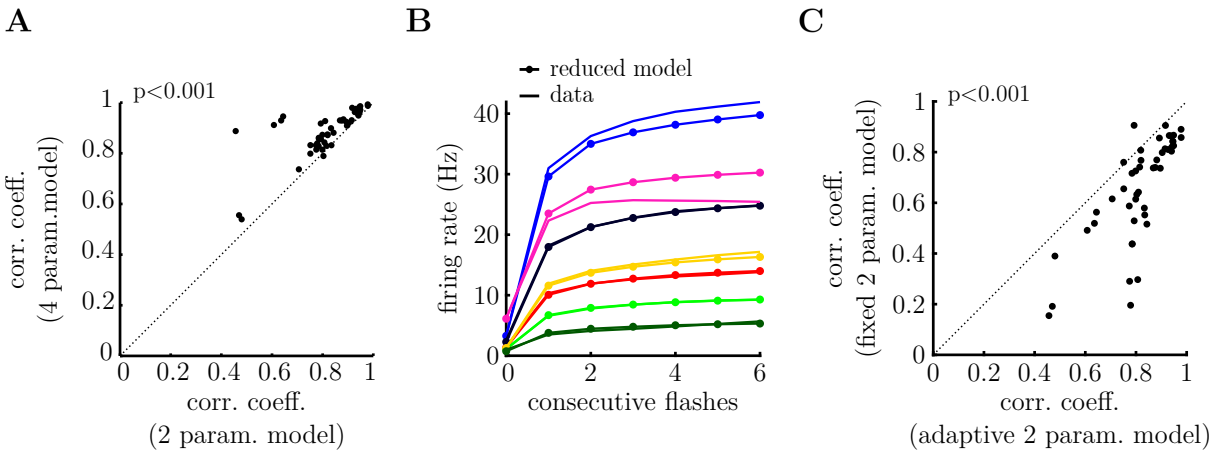

Figure 7: **Reduced adaptive surprise model, with a fixed prior mean (i.e. $\alpha_i/\beta_i = 1$). A.** The reduced model performs almost as well as the adaptive surprise model despite having half the number of free parameters. ($p = 3 \cdot 10^{-9}$, Wilcoxon signed-rank test). **B.** Mean response of 7 cells following a variable number of flashes presented in a row. The increase in the OSR with the number of flashes is well-fitted by the reduced model. **C.** The reduced adaptive surprise model performs significantly better at fitting the recorded neural responses, despite both models having the same number of free parameters per cell. ($p = 4 \cdot 10^{-5}$, Wilcoxon signed-rank test.)