Encoding surprise by retinal ganglion cells

Danica Despotović

SORBONNE UNIVERSITÉ

INSTITUT DE LA VISION PARIS

THÈSE DE DOCTORAT

DE SORBONNE UNIVERSITÉ

Spécialité : Neurosciences

École doctorale nº158: Cerveau, cognition, comportement

Sujet de la thèse :

Encoding surprise by retinal ganglion cells

réalisée

à l'Institut de la Vision

sous la direction de Serge Rosolen

présentée par

Danica Despotović

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

soutenue le 17 janvier 2022

devant le jury composé de :

Pr.	Stephanie Palmer	Rapporteur
Dr.	Georg Keller	Rapporteur
Dr.	Thierry Mora	Examinateur
Dr.	Florent Meyniel	Examinateur
Dr.	Serge Rosolen	Directeur de thèse

Encoding surprise by retinal ganglion cells

Abstract: The efficient coding hypothesis proposes that early sensory neurons transmit maximal information about sensory stimuli, given internal constraints such as metabolic cost of firing a spike or noise. This theory predicts that neurons should adapt to the stimulus statistics and invest most resources on encoding surprising stimuli. Previous results from Schwartz and colleagues showed that some ganglion cells in the axolotl retina respond to an omitted flash in a periodic sequence of flashes (termed the omitted stimulus response, OSR), suggesting that these cells code for unexpected stimuli rather than for physical luminance. However, so far there was no quantitative validation of this assumption. To test this hypothesis, we varied the level of surprise in the stimulus, and recorded the responses of retinal ganglion cells (RGCs) to stochastic sequences of full-field flashes. Our results suggest that, given a simple internal model of the stimulus statistics, neural responses are consistent with the idea of neurons encoding surprise. Moreover, the observed diversity in the RGC population can be explained by different confidence in the internal model of the stimulus statistics.

Keywords: Vision; Retina; Neural Coding; Omitted Stimulus Response (OSR); Predictive Coding; Retinal Ganglion Cells

Codage de la surprise par les cellules ganglionnaires rétiniennes

Résumé : L'hypothèse du codage efficace stipule que les neurones sensoriels maximisent l'information transmise sur les stimuli sensoriels, compte tenu de contraintes internes telles que le coût métabolique des potentiels d'action ou le bruit. Cette théorie prédit que les neurones s'adaptent aux statistiques des stimuli et investissent davantage de ressources dans l'encodage des stimuli surprenants. Des résultats antérieurs de Schwartz et de ses collègues ont montré que certaines cellules ganglionnaires de la rétine de l'axolotl répondent à l'omission d'un flash dans une séquence périodique de flashs (appelée réponse au stimulus omis), ce qui suggère que ces cellules codent pour des stimuli inattendus plutôt que pour la luminance. Cependant, il n'existe jusqu'à présent aucune validation quantitative de cette hypothèse. Pour tester cette hypothèse, nous avons fait varier le niveau de surprise du stimulus et enregistré les réponses des cellules ganglionnaires de la rétine (CGR) à des séquences stochastiques de flashs plein champ. Nos résultats suggèrent que, compte tenu d'un modèle interne simple des statistiques du stimulus, les réponses neuronales sont compatibles avec l'idée que les neurones codent la surprise. De plus, la diversité observée dans la population de CGR peut être expliquée par une confiance différente dans le modèle interne des statistiques du stimulus.

Mots clés : Vision ; Rétine ; Codage Neuronal ; Réponse à un Stimulus Omis ; Codage Prédictif ; Cellules Ganglionnaires de la Rétine

Acknowledgments

I would like to express my gratitude to all those who contributed along the way. First of all, my supervisor Mat Chalk: for the guidance, patience, and above all, everything you taught me. Given it was your first PhD supervision and my first PhD experience, it was a great ride. I thank the French taxpayer for funding our work through Matthew's ANR JCJC project.

Srđan Ostojić and Thierry Mora, for the constructive feedback as members of my thesis committee, as well as the members of the jury: Florent Meyniel, Thierry once more, and especially the readers, Stephanie Palmer and Georg Keller, for their comments.

Olivier, for inspiring discussions and all kinds of advice. Francesco, for being the comedian comrade in various stages of our PhDs. Many colleagues and friends during my stay at IDV: Baptiste, Elaine, Tristan, Ulisse, Giulia, Semih, Deby, Thomas, Gabriel, Matias, Greg, Simone, Samuele, Sarah, and Tom.

Corentin Joffrois, for his reliable and pedantic help with experiments, and Yannick Andéol for providing the axolotls. Serge Rosolen, for helping with all the administrative affairs promptly and with experience.

Dragana Bajić, Nebojša Božanić and Tatjana Lončar-Turukalo, my Bachelor and Master thesis advisers, for showing me the first steps towards biomedical signal processing and computational neuroscience. Petnica Science Center, for the first steps into science in general, and to which I am indebted for most of my closest friends to this day.

All the fun and lively people in Cité universitaire, especially in the Robert Garric residence. Simona and Nikola, for the joyful meals and a welcoming, cozy place to couchsurf in the last few months of frenzy.

My family, for lovingly bringing me up to the point where I can continue on my own. Stefan Vukmanović, for all the trivia, support, and laughs during trying times. Nataša Puzović, for friendship beyond words: "Frodo wouldn't have got far without Sam."

“The only thing that should surprise us is that there are still some things that can surprise us."

— Francois de La Rochefoucauld

Contents

List of Figures

Chapter 1

Introduction

The first stage of our interaction with the environment is to sense it. Early sensory systems are constantly bombarded with diverse stimuli from the external world. It is therefore essential for sensory systems to perform compression of their inputs in real-time, before transmitting this information to downstream brain areas. An open question is how do sensory neurons extract and compress sensory information, that is useful and relevant for the organism.

The influential efficient coding theory proposes that sensory neurons use their limited resources to encode maximal possible information about the natural environment [Attneave, 1954, Barlow et al., 1961]. To do this optimally, neural circuits had to adapt during their evolution to the statistical structure of their environment, so they could focus on conveying the information deemed unpredictable given what they have seen before [Simoncelli and Olshausen, 2001]. However, there is also evidence of certain neural systems, such as retina, adapting to presented stimuli on much shorter time-scales [Wark et al., 2007]. In this case, the time required to 'learn' the feature of the stimulus which is varying can range from hundreds of miliseconds to several minutes [Smirnakis et al., 1997, Shapley and Victor, 1978, Kim and Rieke, 2001].

According to efficient coding, sensory system's should reduce redundancy by preferentially encoding surprising elements of natural scenes in low-noise conditions. In the visual system, a 3-layer neural network called retina is the first processing step for all the visual information available to the rest of the brain. Thus, we can study the relationship between the presented stimulus statistics and

how it is encoded. It is known that retinal neurons can adapt their responses to simple visual features, such as average light level [Shapley and Victor, 1978], contrast [Kim and Rieke, 2001], and spatial correlations [Hosoya et al., 2005]. Still, all of the listed phenomena offer only a qualitative relationship between retinal activity and surprise encoding. Additionally, the extent to which retinal neurons adapt their responses to complex changes in the temporal statistics of presented stimuli is largely unknown. To investigate both questions, we focus on the omitted stimulus response (OSR), a phenomenon where the retina strongly responds to the abrupt stopping of a periodic sequence of flashes [Schwartz et al., 2007a].

In this thesis, we ask whether the unpredictable elements of complex temporal stimuli can be quantitatively related to surprise encoding in the retina. While previous studies have mostly looked at mechanistic models of the OSR [Werner et al., 2008, Gao et al., 2009], here we ask what could be the functional goal of this retinal computation. We investigate whether the retinal neurons match the external stimulus statistics in case of inputs with complex temporal structure. Our research attempts to gain better understanding of the assumptions of the neural network about stimulus surprise, as well as whether these different assumptions could explain the diversity of neural responses in the retinal population.

1.1 Thesis outline

Chapter 2 covers the basics of retinal biology and diversity of computations retina performs, with particular emphasis on the omitted stimulus response phenomena. We provide a brief overview of computational models used to explain and predict retinal activity.

Chapter 3 focuses on the theory of efficient coding and its extensions. The efficient coding hypothesis proposes that the early sensory systems have evolved

to transmit maximal possible information given intrinsic biological constraints, such as metabolic cost of firing a spike, or internal noise.

In Chapter 4 we test the hypothesis that the retinal ganglion cells determine what is surprising in the environment by comparing the stimulus to their internal expectations. We find that the responses can be explained by a simple normative model which combines neuron's expectations about the stimulus with leaky integration of recent events Additionally, we find that the expectation is similar for all the cells in the population, while the confidence in the expectation is what explains the diversity of responses.

Finally, in Chapter 5 we give an overview of our results and discuss possible future research directions. Better understanding of retinal computations could be beneficial for learning about other brain areas as well, since many of the underlying principles of neural coding are common across the neural system.

Chapter 2

Retinal processing

The aim of this chapter is to give an overview of the organisation and function of the retina. We will mostly restrict ourselves to the visual system of vertebrates, since the primary animal model studied in this thesis was the axolotl. This chapter should also provide the reader with information about the types of computations retina performs, as well as quantitative models used to advance our understanding of these neural systems.

2.1 First step of visual processing

Once the light reaches the eye and passes through the pupil, the cornea and lens focus the light so that the image is formed on the retina, a light-sensitive tissue in the back of the eye [Tessier-Lavigne, 2000]. The neural processing of a visual scene starts here, in a three-layered neural network (Figure 2.1). The conversion of light into an electrical signal is performed by photoreceptors, specifically rods and cones, which respond to light via a graded change in their membrane potential. The visual information is then transmitted through a layer of interneurons, where the graded potentials from photoreceptors are fed to bipolar cells, while being modulated by horizontal cells which connect laterally to rods and cones. The bipolar cells' outputs are affected by another class of interneurons, amacrine cells, which perform lateral inhibition. Finally, retinal ganglion cells (RGCs) receive the inputs from bipolar cells, and communicate the visual information to the rest of the central nervous system through the optic nerve,

comprised of RGC axons. Unlike retinal interneurons, which communicate on smaller scales, the RGCs generate action potentials i.e. spike in order to transmit the signal over long distances.

Apart from rods and cons, there is a third type of photosensitive retinal cell responsive directly to light, namely the intrinsically-photosensitive retinal ganglion cells (ipRGCs) [Morgan and Kamp, 1980, Foster et al., 1991, Hattar et al., 2002]. We will focus on the 'classical' retinal ganglion cells for the rest of the section since in this thesis we record and model their activity. Recently, a completely new class of retinal cells was hypothesized to exist, called the Campana cells, however there is still little known about them [Young et al., 2021].

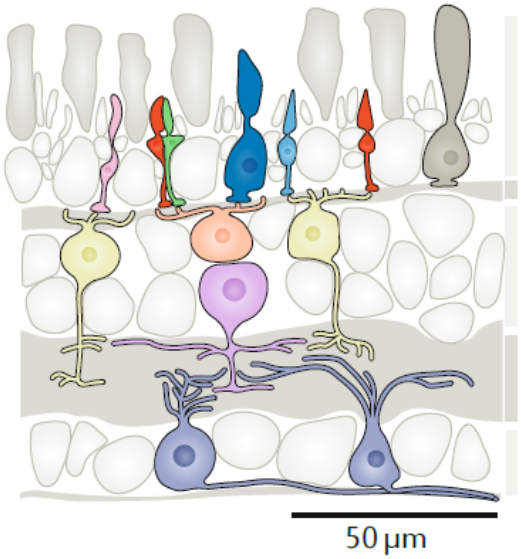

Fig. 2.1 *Structure of the larval tiger salamander retina. Light is transformed to electrical signal by one of six types of photoreceptors in the layer shown on top (pink, red, green, dark and light blue, gray). Next layer consists of interneurons: horizontal cells (peach), bipolar cells (yellow) and amacrine cells (purple). Lastly, retinal ganglion cells (pigeon blue) integrate outputs of other cells and transmit it to the rest of the brain. Adapted from [Baden et al., 2020].*

One of the first recorded retinal responses was the one related to luminosity changes, when first functional distinction between different types of ganglion

cells was discovered [Hartline, 1938]. Using an oscillograph to record a single intraocular fiber of bull-frog retina, Hartline was able to classify retinal ganglion cells according to their responses to light pulses. Some cells regularly responded to the increase in the light level (termed ON cells), while other did so for decrease (OFF cells); a combination of these classes, called ON-OFF cell, was activated by both increases and decreases of light intensity.

Moreover, Hartline found that responses can be found in a certain fiber only if a restricted area of the retina is stimulated by light dots, effectively discovering receptive field [Hartline, 1938, Hartline, 1940]. The receptive field is usually constituted of two typically concentric circles, center and surround, which might be of same or opposite polarity (ON/OFF) [Hartline, 1940, Kuffler, 1953]. The structure of center and surround is fluid, and cannot always be considered to be a regular shape [Levick, 1967, Liu et al., 2009]. A cell with a RF of opposite polarity of center and surround exhibits the center-surround antagonism. For example, if an RF with ON center and OFF surround is flashed with a bright spot, the excitatory center and inhibitory surround will cancel out. This can be understood intuitively as a way of preventing energy expenditure on parts of visual scenes with uniform luminosity, such as cloudless blue sky or a white wall. On contrary, if the center of RF is stimulated with a bright spot, while the surround is presented with a dark band, the response from the center and surround will be combined into a stronger one. The element of visual scene that corresponds to this situation is a high-contrast, such as an edge.

While retinal cells can be divided into aforementioned 5 broad classes, each class has a high number of anatomically and morphologically distinct cell types: for example, in mammalian retina the lower estimate is around 60 cell types [Masland, 2012]. There is an abundance of cell types even if we focus solely on retina's output cells, ganglion cells. For instance, by stimulating with white noise and chirp stimulus, it was revealed that the mouse retina has more than

30 functionally distinct types of ganglion cells, which respond in different way to the same stimulation pattern [Baden et al., 2016]. Around 17 types were found in the primate retina so far [Grünert and Martin, 2020], although 5 types - ON and OFF midget, ON and OFF parasol, and small bistratified cells – together make up for 75% of all cells [Dacey, 2004b]. The conclusions of studies of the axolotl retina, which is the model animal we primarily discuss here, has been somewhat ambiguous regarding the number of RGC types, but the latest studies estimate presence of 5-7 types [Segev et al., 2006, Marre et al., 2012, Rozenblit and Gollisch, 2020].

A distinct RGC type is frequently found to uniformly cover the visual field in a regular lattice structure, displaying mosaic organisation (again, with a certain degree of uncertainty in case of salamander, since only some of the cell types were found to tile the space without an overlap) [Segev et al., 2006, Marre et al., 2012, Kastner and Baccus, 2011, Kühn and Gollisch, 2016]. This enables the retina to uniformly sample the visual space, creating a 'sensory map', where each RGC cell type then extracts a certain low-level feature of the visual scene [da Silveira and Roska, 2011].

2.2 Computations in the retina

Until quite recently it was thought that the retina's role is mainly one of a 'camera sensor', adapting to the light intensity and performing spatio-temporal filtering using center-surround antagonism [Meister and Berry, 1999]. This view would assume the visual scene is transmitted to the downstream areas as a matrix of pixels that are sharpened in both space and time. However, such pixel-by-pixel representation seems unlikely given two facts: (i) the number of photoreceptors is 2 orders of magnitude higher than the number of ganglion cells (both in mouse and human, as example), (ii) the diversity of retinal cell types. These imply the

information has to be re-packaged to pass this bottleneck in a meaningful way. Additionally, as Gollisch and Meister point out, there is a paradox in assuming simple operations such as adapting to changing light levels and image sharpening would require such a complex network comprised of such a variety of neuron types [Gollisch and Meister, 2010]. One of the possible explanations for the diversity of cell types is that different retinal computations require parallel pathways to transmit different features of the visual scene [Wässle, 2004, Dacey, 2004a]. In other words, there is a need for diversity of cell types to fulfill various functions the retina performs.

A good example of feature extraction happening as early as the retina are the direction-selective retinal ganglion cells (DS RGCs). When a moving stimuli, such as a grating or bar, passes across its receptive field, these cells fire spikes with clear preference for one direction [Barlow et al., 1964, Demb, 2007]. This illustrates how the information from the visual scene can be compressed already at the first stage of neural processing, with subset of cells - DS RGCs - conveying nothing else apart from the signal about object direction. It also allows for downstream areas to directly read-out said direction by integrating activity of several direction-selective RGCs. This kind of computation is an illustration of explicit coding of a certain property of the visual environment. Furthermore, it is not the only such example: previous studies were able to decode various features of the external stimuli in the activity of ganglion cells, such as contrast [Shapley and Victor, 1978, Smirnakis et al., 1997, Goldin et al., 2021], local and global motion [Oyster, 1968, Ölveczky et al., 2003, Kühn and Gollisch, 2016], texture motion [Enroth-Cugell and Robson, 1966, Kaplan and Shapley, 1986, Petrusca et al., 2007], and approach sensitivity [Münch et al., 2009]. In fact, it is not unusual for multiple features to be encoded by the same cells, such as object motion and direction [Kühn and Gollisch, 2016], or object position and speed [Deny et al., 2017] (for review, see [Gollisch and Meister, 2010]).

2.2.1 Retinal anticipation

The signal transmission through the photoreceptor cascade introduces delays of around 30-100 ms, which might be critical for the flight or fight response [Gollisch and Meister, 2010]. A subset of retinal computations is related to retina's apparent ability to counter these intrinsic delays by anticipating future stimulus states. If an object is moving over the retina, we could expect the prediction of the object position to lag behind object's actual position. However, in an experiment with a smoothly moving bar, it was revealed that the peak of RGCs population activity in fact corresponds to the current position of the bar, or even its position slightly in the future [Berry et al., 1999] (Figure 2.2). The retina compensates for the processing delays by extrapolating the upcoming bar position given the regularity of its movement. This finding was surprising given that at the time motion anticipation was proposed to be generated by some higher-level brain area [Nijhawan, 1994, De Valois and De Valois, 1991].

An analogous effect, found in psychophysics, is the flash-lag effect: participants were shown a bar moving at fixed speed and another bar flashed in continuation of the moving one. Despite the two bars being aligned, participants would report the moving bar being ahead, suggesting another example of motion extrapolation at hand [Nijhawan, 2002]. The results of Berry et al. suggest that spatial anticipation is not unique for the visual cortex, but can also be computed at the first stage of visual processing as well. Furthermore, the computation of flash-lag effect was also more recently associated with known feed-forward retinal mechanisms [Subramaniyan et al., 2018, Nijhawan, 2002, Rust and Palmer, 2021].

Continuing the work on anticipation of bar motion, Schwartz et al. asked the following: if the retina is extrapolating the motion of the bar, what would be the response in the case where the movement is interrupted? In the

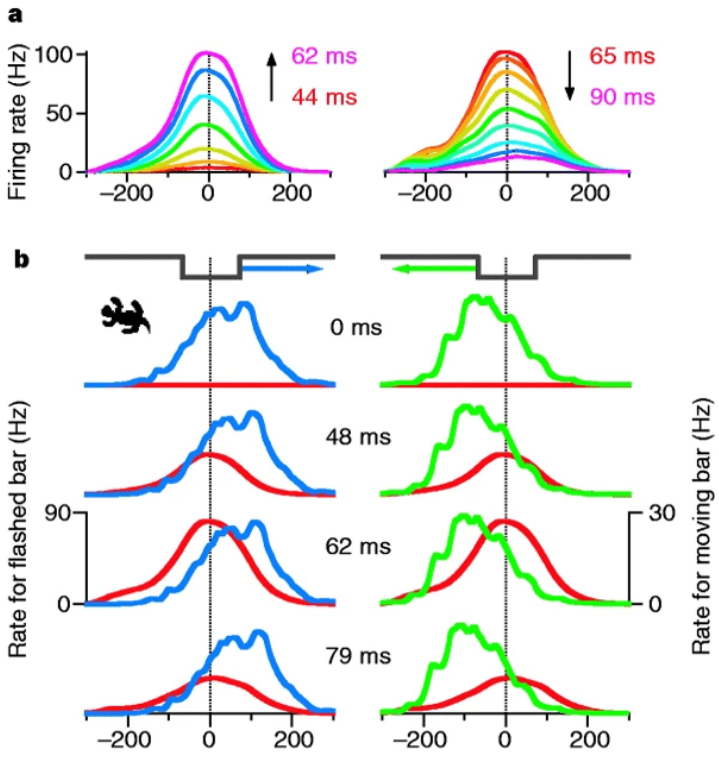

Fig. 2.2 *RGC responses to flashed and moving bar. A. Spatial profile of the firing rate in response to a flashed bar (15 cells from salamander retina). The peak is centered at 0, corresponding to true bar position. B. Responses to flashed bar from (A) (red, population response) compared to responses to bar moving to right (blue, population response) and left (green, population response). The 'visual image' and 'neural image' are aligned despite the processing delay. Adapted from [Berry et al., 1999].*

experiment where a moving bar makes a sudden turn and changes direction, the RGC population will at first continue to signal the position as if the bar didn't turn, but it will quickly, after several tens of milliseconds, update on the new bar direction and continue with correct predictions of its position [Schwartz et al., 2007b]. In the case of 180 degrees reversal, there is a brief synchronized burst of spikes, possibly signalling an error in anticipated motion to downstream areas. Similarly, a sudden onset of movement elicits a stronger response than smooth motion, possibly because it contradicts the expectation of having a stationary

object [Chen et al., 2013].

2.2.2 Omitted stimulus response

We have seen how the retina responds to stimuli with a predictable spatial component. To test whether similar findings stand for temporal patterns, Schwartz and colleagues stimulated the retina with a sequence of periodic flashes [Schwartz et al., 2007a]. They found that once that sequence is abruptly stopped, the RGCs respond strongly (Figure 2.3, e.g. third row). This is yet another nonlinear phenomena: the omitted stimulus response (OSR), also known as the omitted stimulus potential (OSP) [Bullock et al., 1990a]. Similar to motion reversal effect, here the temporal regularity - the periodic nature of the flashes - is violated, causing the RGCs to seemingly signal the deviation from the prediction. Moreover, the timing of the OSR is not constant but instead carries information: it depends on the inter-flash period in the range between 6 and 20 Hz [Schwartz et al., 2007a]. The retina appears to 'learn' the exact interval between two flashes and the latency of the OSR is consistently shifting with it (Figure 2.4). The robustness of OSR was probed by jittering the periods between flashes, changing flashes to different shapes, etc, however the response persisted despite the noise [Schwartz and Berry 2nd, 2008]. Possibly the most surprising discovery is the variety of behaviour in the recorded responses, as can be seen from 10 different combinations of responses to beginning and ending of the flash sequence (Figure 2.3).

The diversity of cell types might seem like a good candidate for explaining the variety of responses, since different types are assumed to encode [Wässle, 2004]. However, so far there was no clear correspondence found between the response type and either receptive field size, ON-OFF index, or spike-triggered average [Schwartz and Berry 2nd, 2008, Deshmukh, 2015]. An important caveat is that most of the OSR studies were done in salamander, where cell typing is still not well-established and boundaries between types are less clear, so further research

into response diversity might benefit from focusing on the mouse as the model animal.

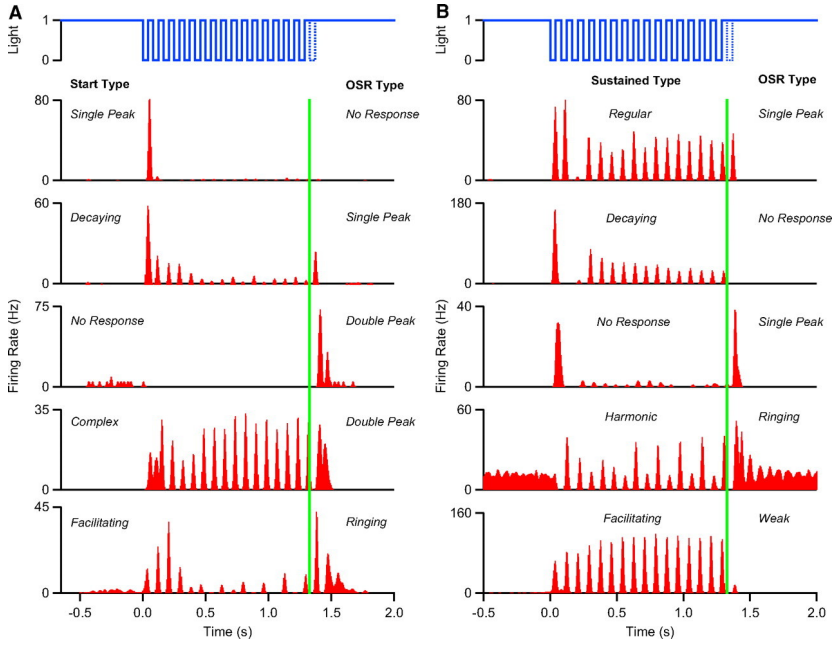

Fig. 2.3 *Diversity of OSR responses shown. RGC activity is presented using PSTH. Solid blue line represents stimulus, dotted blue line the omitted flash, while the green line is the timing of the omitted flash. Note that not all the cells exhibit an OSR. A. Variety of responses depending on the first few flashes ranges from no activity to strong complex response. B. Rest of the flash sequence also elicits different, sustained responses: while some cells follow the stimulus intensity and respond to each flash, others show increase in firing rate only for the beginning and ending of the sequence. Reproduced from [Schwartz and Berry 2nd, 2008]*

An OSR-like phenomena was spotted across numerous animal species. So far it was observed not only in the retina of salamander and mouse [Schwartz et al., 2007a, Schwartz and Berry 2nd, 2008, Werner et al., 2008], but also in humans using electroretinogram [Bullock et al., 1994, McAnany and Alexander, 2009, Fradkin, 2020], turtles [Prechtl and Bullock, 1994], zebrafish larvae [Sumbre et al., 2008], both cartilaginous and bony fish [Bullock et al., 1990b, Bullock et al., 1990a, Bullock et al., 1993], bullfrog [Chen et al., 2014], and even invertebrates such as crayfish [Ramón et al., 2001]. Similar findings by all these studies suggest

there is a computational goal of such a response that extends beyond one species: any animal needs to be able to make predictions about the environment while saving the energy when something predictable is happening. This is also one of the reasons for emphasising importance of studying retina's circuitry across different species (for review, see [Baden et al., 2020]).

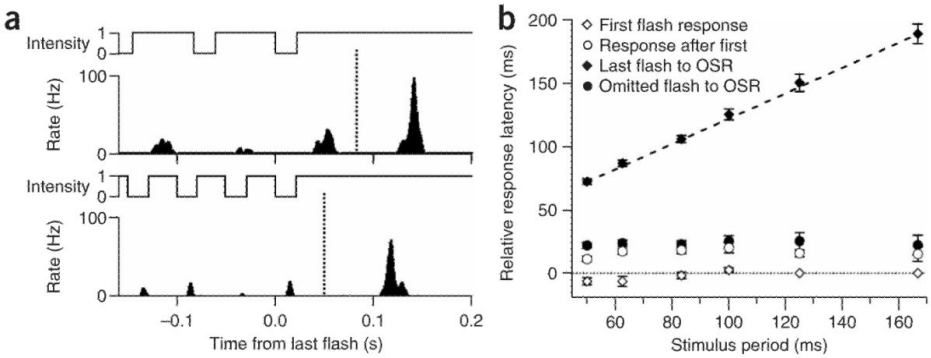

Fig. 2.4 *Entrainment of OSR to the flash period. A. RGC firing rate after showing 40 ms dark flashes with different frequency (12 Hz (top), 20 Hz (bottom)). Dotted line represents the omitted flash. B. Relation between flash period (1/frequency) and response latency. The timing of OSR is linearly related to the period between flashes. Dotted line is the unity line. Reproduced from [Schwartz and Berry 2nd, 2008].*

A common question regarding the nature of OSR is whether maybe the retina treats the array of flashes as a prolonged ON step stimulation, and whether then OSR can be considered as a released suppression. There are two arguments against this hypothesis: (i) the timing of OSR is dependent on the period between flashes, indicating flashes are treated as separate events and neural circuitry learns that period (see Figure 2.4); (ii) an OSR persists even when the average luminosity during flash sequences and between trials is kept constant [Schwartz et al., 2007a].

We will briefly cover the possible mechanistic explanations of the OSR. The first thing to note is that depending on polarity of the flashes (dark or bright), there are most likely two distinct pathways activated [Weidmann, 2009]. The

first attempt to find a neural circuitry that is responsible for OSR found that the phenomena is still present when inhibition from amacrine cells was blocked [Schwartz and Berry 2nd, 2008]. The same study found that the ON bipolar cells were required for the dark-flash OSR, forming the basis for a model in which the OSR is produced by combining the oscillatory response of the ON bipolar cells with the inputs from OFF bipolar cells [Schwartz and Berry 2nd, 2008, Gao et al., 2009, Deshmukh, 2015].

An alternative proposal claimed OSR is in fact not a response to the omitted flash, but instead a byproduct of temporal integration of the flash sequence [Werner et al., 2008]. In this model, it was possible to reproduce OSR with a combination of time-shifted RGC ON and OFF channels. However, the main criticism of this result is that the model predicts that the relationship between flash period and OSR latency, while still linear, does not lie on the unity line; however, the experiments in [Werner et al., 2008] show that for different cells, the slope was on average 0.85, with variations from 0.4 to 1.8. Recently, there were also attempts to discover the mechanism using convolutional neural networks trained on the retinal data [McIntosh et al., 2016]. By deriving the minimal necessary subset of channels i.e. cells, Tanaka et al. find that in their model, a mixture of 2 ON and 1 OFF channel is sufficient to generate an OSR-like behaviour, which successfully reproduces the slope distribution [Tanaka et al., 2019]. To the best of author's knowledge, this hypothesis remains to be experimentally tested.

2.3 Modelling retinal responses

Now that we have seen some of the computations occurring in the retina's output layer, we want to know how to quantitatively describe neural activity. A basic way to model cell's responses to stimuli would be to represent it as a cascade model (Figure 2.5). First, the stimulus $\vec{x}$ is filtered by the cell's receptive field,

which is represented as a linear spatio-temporal filter $\vec{k}$. Since the firing rate cannot be negative and the neuron cannot fire with infinite frequency, i.e. there is a saturation point due to the biophysical properties of neurons, the output of the linear filter has to be rectified [Dayan and Abbott, 2001]. This can be done by filtering with a nonlinear function f, also known as the link function, giving as a result the instantaneous firing rate λ:

$$\lambda = f(\vec{k} \cdot \vec{x})$$

To generate stochastic spike trains, it is approximated that neuron's firing corresponds to a Poisson process with mean λ. This kind of cascade model is termed Linear-Nonlinear Poisson model (LNP), or Poisson Generalized Linear Model (GLM) [Simoncelli et al., 2004]. Although LNP model has limited relation to biophysical processes, it provides a compact way to relate input stimulus x with the recorded responses r and obtain a prediction of average firing rate, such as peri-stimulus time histogram (PSTH) [Pillow et al., 2005, Pillow, 2007]. For estimating instantaneous firing rate, several parameters have to be fitted

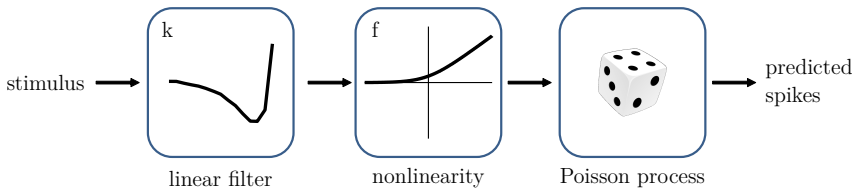

Fig. 2.5 *Schematic of a linear-nonlinear (cascade) Poisson model. The stimulus is filtered with linear kernel k, passed through a rectifying nonlinearity f (here rectified linear unit, ReLU), followed by Poisson spike generation.*

to describe the LN model: the coefficients of the linear filter k, and, depending on the functional form of nonlinearity, the parameters governing it (usually f is an exponential or other rectyfier function). In case of white noise stimuli, it

is possible to estimate the linear filter from the data by estimating the average stimulus that elicits a spike, called the spike-triggered average (STA) [Chichilnisky, 2001] (Figure 2.6). Under certain conditions, STA is an unbiased estimate for the linear filter, making the LN model easy to estimate and computationally tractable [Chichilnisky, 2001, Paninski, 2004]. A more general approach to fit the parameters of the linear filter is to maximize the likelihood, $p(r|x)$ (for derivation, see [Pillow, 2007]). This optimisation is convex, making the inference of LN parameters relatively straightforward. Thanks to the accurate performance in predicting average firing rate and relatively easy inference, LN model found its application in many areas, successfully explaining the responses of retinal ganglion cells [Pillow et al., 2008], ipRGCs [Milosavljevic et al., 2018], lateral geniculate nucleus [Babadi et al., 2010], visual [Kotekal and MacLean, 2020], motor [Truccolo et al., 2005] and parietal [Park et al., 2014] cortex.

However, the explanatory power of LN model is by design limited when it comes to more complex stimuli with spatio-temporal correlations, such as natural images or videos [Heitman et al., 2016, McIntosh et al., 2016]. Although the simplicity of the LN model is one of its advantages, it also means there is a lack of different sources of nonlinear interactions, which are highly present even in the retina, and more so later in the visual cortex [Latimer et al., 2014]. Some of the shortcomings can be mitigated by extending the LN model to include post-spike filters or couplings with other neurons [Pillow et al., 2008], multiple linear filters i.e. filter banks [Gollisch and Meister, 2008] or stacked, two-layered LN also known as LN-LN [Shapley and Enroth-Cugell, 1984, Deny, 2016], or independently infer parameters of stimulus filter and coupling filter [Mahuas et al., 2020]. For example, for the previously described phenomena of motion extrapolation, it was possible to explain the responses using an LN model with an added contrast gain control [Berry et al., 1999]. Yet, LN models still fail to predict responses in certain cases, such as (i) when neurons are presented with natural stimuli, (ii) when neurons

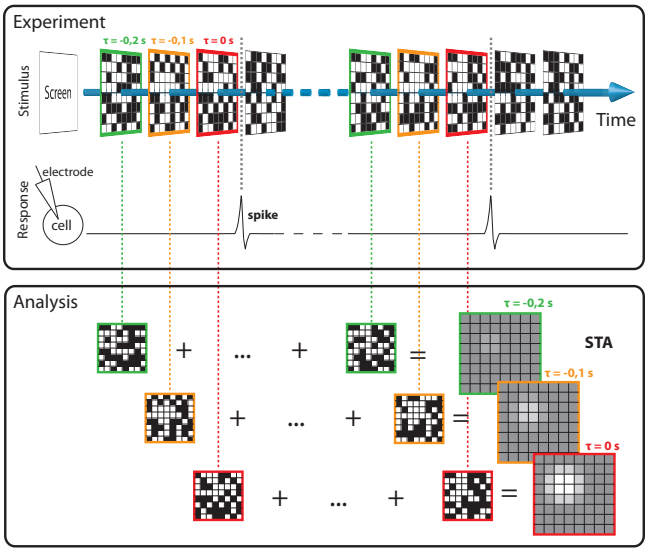

Fig. 2.6 *Computing spike-triggered average. Random white noise, i.e. checkerboard stimulus, is presented while recording neuron's spiking activity. The average number of spikes elicited by a certain sequence of frames of stimulus is calculated (here, 3 time-bins depicted in different colours). In this example, the neuron responds strongly to the increase of light in upper left corner of the scene. The analysis part shows the frame summation, the averaging is not displayed. Reproduced from [Deny, 2016].*

have sub-Poisson spiking variability, or (iii) when the polarity of a cell changes [Heitman et al., 2016, McIntosh et al., 2016, Maheswaranathan et al., 2019].

As a more complex and flexible model which might be able to resolve these issues, convolutional neural networks (CNNs) were proposed as another possible model of visual systems [Yamins et al., 2014, Cadena et al., 2019], including the retina [McIntosh et al., 2016, Maheswaranathan et al., 2019, Tanaka et al., 2019, Goldin et al., 2021]; for review, see [Lindsay, 2021]. Inspired by natural vision, the basic computation in CNNs is the convolution of the input image with a set of learned filters, which are then combined to predict neural responses [LeCun et al., 1998]. Note that a two-layered CNN is in principal equivalent to an LN-LN model. Unlike LN models, learning the parameters of CNN is a slow and computationally expensive task, partially due to the fact that

number of parameters required for CNN can be several orders of magnitude higher than for the LN models (e.g. in [McIntosh et al., 2016], 150 thousand vs 4 thousand).

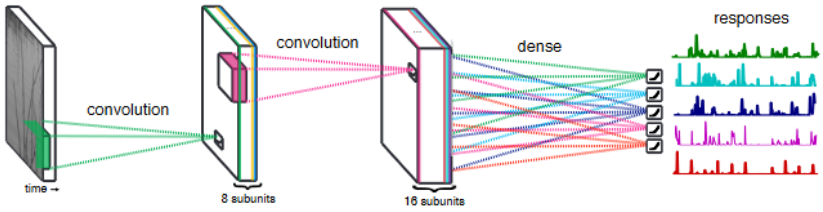

Fig. 2.7 *Example of a CNN architecture. The input image is convolved with two layers of spatiotemporal filters which are learned from the data. Lastly, the outputs are combined linearly and filtered through rectifying non-linearity to output the predicted firing rate. Reproduced from [McIntosh et al., 2016].*

Recently, McIntosh and colleagues found that a three-layered CNN model, trained on retinal responses to natural movies, outperforms LN models with and without spike history filters not only on the responses elicited by natural scenes, but white noise stimulus as well [McIntosh et al., 2016, Maheswaranathan et al., 2019]. Moreover, the CNN is able to reproduce multiple retinal phenomena related to motion encoding and predictive coding, including the omitted stimulus response. However, it requires additional work to compare these complex models to biological systems that they emulate, and extract information about computations a CNN model performs [Tanaka et al., 2019].

Chapter 3

Efficient coding

The natural world is full of redundancies. The information received from the visual scenes is no exception, as Attneave points out in his seminal work [Attneave, 1954]. Attneave presents a thought experiment to illustrate the redundancy of visual stimulus inspired by Shannon's guessing game with English language [Shannon, 1951]. In Shannon's experiment, a subject is tasked to fill in the missing letters from an unknown paragraph. Out of 129 letters, the subject guessed 89 correctly (69%), implying most of the letters are not needed for understanding the text, but can be predicted *frm th cntxt*.

As Attneave suggested, similar investigation into visual redundancy can be done: if a subject would be asked to describe a simple image of N pixels containing three shapes of different colour (Figure 3.1), they could make better-than-chance guesses to predict neighbouring pixel's colour by using significantly less than N tries. Later work by Kersten sought to quantify precisely the level of redundancy, using psychophysics experiments with grey-coloured images which have certain share of deleted pixels [Kersten, 1987]. He found that the spatial redundancy ranges from 46% for a complex image, to 74% for a picture of a face. This result demonstrates the presence of various forms of redundancy, such as continuity of texture and homogeneity of colour, enabling the subject to make assumptions about the neighbouring pixel. The errors subjects make would be at points where there is an unpredictable change in shape or colour, or as Attneave points it out, 'information is concentrated along contours'.

A more rigorous, quantitative formulation of Attneave's observations is

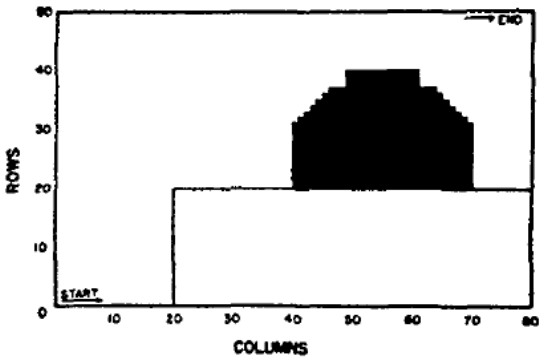

Fig. 3.1 *Illustration of a redundant visual scene for the described thought experiment described in [Attneave, 1954]. Three objects of different colour (an ink bottle, a table and background wall) are shown: a subject guessing the neighbouring pixel is likely to make error around contours of these objects.*

given by Barlow [Barlow et al., 1961]. Following Shannon's ground-laying work in mathematically defining concepts like information and redundancy [Shannon, 1948], Barlow applied fundamental ideas from communications theory in the context of neuroscience to define efficient coding. According to this hypothesis, the neurons in sensory areas have adapted to transmit maximal possible information to the rest of the brain, despite the internal constraints such as metabolic cost and noise. Barlow reduces a neural system to an information processing system which has a limited capacity, and therefore has to choose an appropriate neural code to transform the stimulus in order to transmit only the high-information content. The prediction that follows is then that the guiding principle of neural circuits should be to remove statistical redundancies from the sensory signals, i.e. the 'redundancy reduction' hypothesis (see [Barlow, 2001] for review).

One of the early tests of this hypothesis was in the compound eye of the fly [Srinivasan et al., 1982]. Srinivasan and colleagues asked whether neurons take advantage of the structure of the natural images by implementing a center-surround receptive field (RF). Their proposal was that the excitatory center and inhibitory

surround combine as a linear weighted sum of input luminosity. Contrary to what might be expected, this RF structure implies RGCs are not encoding constant luminance, but contrast. Based on their model, they predicted that at low light intensities, the inhibition should be weaker in order to faithfully represent the luminosity in the center. Moreover, they found experimental support for this assumption by comparing their predicted RFs to recorded RFs of interneurons in the fly retina at different light levels (i.e. different levels of noise). Several later studies focused on the statistics of natural scenes, and found the codes visual system has evolved to utilise are indeed the ones which are well-tuned for decorrelating visual information ([Field, 1987, Switkes et al., 1978, Rieke et al., 1995], reviewed in [Atick, 1992, Simoncelli and Olshausen, 2001]).

3.1 Efficient coding in the retina

As we have seen in Chapter 2, in most of the mammals there is a thousand-fold reduction in number of cells between the photoreceptor layer and the retinal ganglion cells output. This bottleneck makes the retina a good candidate to test the efficient coding theory, since in such conditions it is suggested that the retina would have to compress the incoming information. Moreover, it is possible to record a representative sample the whole retinal output, which makes validating the theoretical predictions with experimental data feasable.

Atick and Redlich were able to predict the variation of receptive field shape depending on the noise conditions by tarting from efficient coding hypothesis as a design principle. In the low-noise setting, the center-surround structure of retinal ganglion cells receptive fields is used to integrate inputs from within their RF while suppressing the stimuli in their immediate surround (Figure 3.2A). This finding is in accordance with Barlow's original hypothesis, since he assumed a noiseless channel, hence being efficient in this case means the optimal strategy is

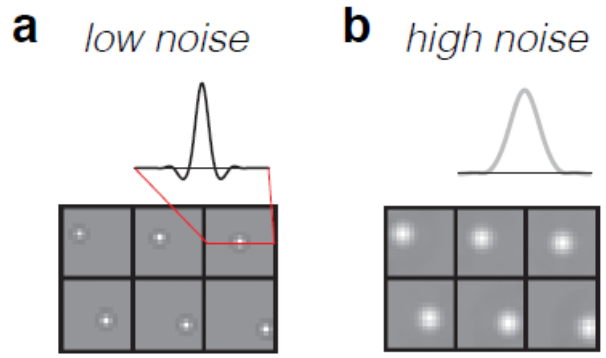

Fig. 3.2 *Receptive field shape prediction depending on the noise level.* ***A.*** *In case of low noise, the RF shows excitatory center and inhibitory surround, and therefore whitens the image.* ***B.*** *In high-noise setting, the surround suppression is weak, enabling neurons to use this RF shape to average out the noise. Adapted from [Doi and Lewicki, 2007].*

to reduce the redundancy and decorrelate the input stimulus.

In contrast, the efficient coding theory makes an opposite prediction when sensory inputs are corrupted by a high level of noise. Here, the optimal code is actually the one in which neurons respond redundantly to their inputs, so as to average out the noise, leading to an increase in signal-to-noise (SNR) ratio. As a result, the efficient coding theory predicts that the neural code should change qualitatively with varying input noise, acting as a whitening filter at low noise, and a smoothing filter at high noise. Interestingly, Atick and Redlich showed that this can explain the observed changes in the RF shape of RGCs with decreasing visual contrast, which become broader and have a weaker suppressive surround at lower contrast levels (Figure 3.2B). However, this should not come as a surprise. The optimality of the code depends on the input: a code which is efficient for a certain input statistics is not necessarily optimal for another [Simoncelli and Olshausen, 2001].

Atick and Redlich made a number of simplifying assumptions about the nature of the neural code, where RGCs are assumed to behave as linear deterministic filters of their inputs. Since then, a number of authors have investigated what

happens in the more general case, where neural responses are noisy and non-linear. It was shown that, with more realistic neural models, efficient coding can account for many qualitative aspects of retinal organisation, such as the ratio between ON and OFF cell types [Karklin and Simoncelli, 2011, Ratliff et al., 2010], the overlap between RFs [Doi and Lewicki, 2007], changes in RFs with varying retinal eccentricity [Doi and Lewicki, 2014, Ocko et al., 2018]. Likewise, starting from efficient coding principles it is possible to explain how having both ON and OFF cell pathways leads to a lower metabolic cost on average [Gjorgjieva et al., 2014]. In recent work, Doi et al. directly compared predictions of efficient coding with simultaneous recordings from cone photoreceptors and RGCs [Doi et al., 2012]. They found that ganglion cells exhibited high ($\sim 80\%$) efficiency in transmitting spatial information, relative to their model. Recently, Ocko et al. found that it is possible to start from first principles (statistics of natural movies and realistic energy constraints), and reconstructed the spatial and temporal sensitivity, cell spacing, ratio of cells types, as well as how distribution of cell types changes with eccentricity in primate retina [Ocko et al., 2018].

Despite these successful predictions of qualitative features of retinal organisation, efficient coding theory has several limitations. The universal nature of its formulation makes it both very flexible for various interpretations and elusive for applying to data: it is a difficult task to pinpoint what information is encoded, the statistics of the natural scenes, or what constraints should be taken into account without making the model intractable. As a result, it has thus far proved hard to move beyond qualitative features to make direct, quantitative comparisons to the neural activity. In the example of visual system, there is a higher count of cortical neurons ($\sim 10^9$) than ganglion cells in the retina ($\sim 10^6$), therefore there would be no need for compression of the signals coming downstream from the retina, which seems to be opposite of what would be predicted by efficient coding [Barlow, 2001]. However, the complexity and diversity of visual tasks is also

increasing as the signal move through the hierarchy of sensory processing, as well as the timescales over which visual cortex has to integrate the incoming inputs. It is important to note that in this criticism there is an implicit assumption that the number of neurons is the only relevant factor as if each neuron can convey an equal information load, which is still an open question [Simoncelli and Olshausen, 2001, Barlow, 2001].

Lastly, since efficient coding is envisioned as a general guiding principle for neural organisation, it posits all sensory information is treated equally. This goes against the empirical evidence showing that neural systems prioritise behaviourally relevant stimuli, & not just statistically likely ones [Machens et al., 2005]. For example, for interactions in human world it is significant to recognize a face of a fellow human, unlike distinguishing one treetop from another. Supporting this idea, Machens et al. demonstrate how grasshoppers auditory receptor neurons preferentially encode calling songs for mating over other natural stimuli: the ecological importance of the former would not be captured by the efficient coding hypothesis.

3.2 Coding for predictions

To overcome agnosticism about the what type of stimuli is relevant, an alternative hypothesis was proposed: that neural systems encode maximal information about those stimuli that are predictive about the future, while discarding other, non-predictive, information [Bialek et al., 2006, Salisbury and Palmer, 2015]. This 'predictive coding' theory is motivated by the fact that only 'predictive' information allows the organism to respond adaptively to changes in the environment and improve its chances of survival. In case of sensory system, this can be motivated by the need to compensate for processing delays and act based on these anticipated future states.

The relationship between the information about the stimulus and the responses can be formalised using information bottleneck (IB) method [Tishby et al., 2000]. In the context of coding for predictions, the IB theory prescribes how to encode a stimulus so as to preserve maximal information about the future, given a certain amount of information encoded about the past. Namely, it is an optimisation problem which a neural system attempts to solve: consider input stimulus X, which is coded using response variable R, in order to transmit information about some relevant variable Y. The question is how to transmit maximal possible information about the future, I_{pred}, given that the information about the past, I_{past}, has to be compressed. Therefore, the goal is to minimise the following loss function:

$$L = I(X;R) - \beta I(R;Y)$$

where I stands for mutual information, and β is the parameter determining trade-off between accurate representation of the past (compression) and information about the future (prediction). In case of coding for predictions, the assumption is that the target variable Y is the state of input stimulus Δ steps ahead, i.e. $Y = X(t + \Delta)$.

Recently, Palmer et al. tested this idea of extraction of predictive information in the context of retinal encoding [Palmer et al., 2015]. They measured the spiking activity of a population of RGCs in response to a moving bar. In support of the theory, they found that RGCs encoded close to maximal information about the future trajectory of the bar, given the amount of information they encoded about its past trajectory (Figure 3.3). Notably, Palmer and colleagues also show that downstream neurons can almost optimally read-out predictive information in such form, even if they receive no other inputs [Sederberg et al., 2018]. An open question remains how does this approach translate to more naturalistic stimuli.

Also, how to more explicitly quantify this 'closeness' to optimal encoding is not yet clear, with one recent proposition given in [Młynarski et al., 2021].

Work so far assumed neurons encode predictive information redundantly, via their 'instantaneous' responses i.e. using the information from previous time-point only (this is also implicitly assumed by Palmer et al.) [Wiskott and Sejnowski, 2002, Creutzig and Sprekeler, 2008, Berkes and Wiskott, 2005]. In this view, the theory predicts that neurons should preferentially encode slowly varying (i.e. temporally correlated) stimulus features, since they persist into and are predictive of the future stimulus (i.e. 'smoothing'). Note that this is an extension of the initial interpretations of efficient coding, which (at low-noise) predicts that neurons should temporally decorrelate the stimulus, and only respond when it transiently changes on fast timescales (i.e. 'whitening').

Subsequent work provides a way to explain both cases in a single framework [Chalk et al., 2018]. Chalk et al. show different coding objectives can lead to very different coding strategies which are efficient in their respective conditions.

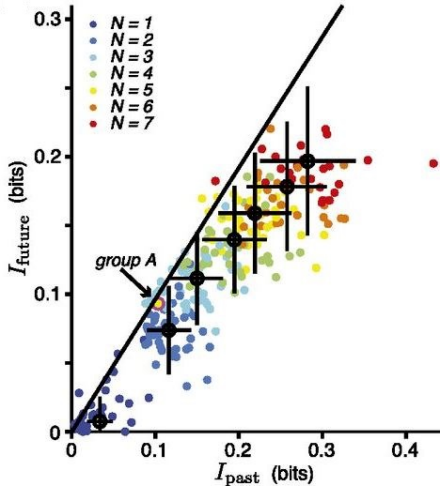

Fig. 3.3 *Information encoded by RGCs about the future vs. information about the past. Each point is a group of cells, with colours indicating group size N. Theoretical upper bound (solid line) is defined by the statistics of the bar motion. Adapted from [Palmer et al., 2015].*

In this theory, neural responses (in a time window of length τ are hypothesised to transmit maximal information about the stimulus at some time-lag, Δ, constrained on the total information encoded about previous inputs, C (Figure 3.4A). Thus, the predicted neural code depends on three optimisation parameters (τ, Δ, and C), which together describe the functional goals and constraints faced by the system (Figure 3.4B).

Previous coding theories emerge as special cases of this more general theory, given specific choices of optimisation parameters. For example, efficient coding is obtained by assuming a temporally extended code ($\tau \gg 0$) and negative decoding lag ($\Delta < 0$; blue region in Figure 3.4B). On the other hand, 'instantaneous predictive coding' (as described in Palmer et al.) is obtained by using a short coding window ($\tau \sim 0$), and positive decoding lag ($\Delta > 0$; red region in Figure 3.4B).

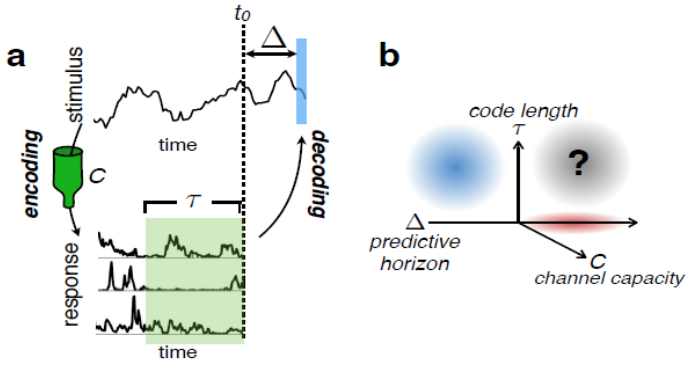

Fig. 3.4 *Unifying theoretical framework for coding strategies.* ***A.*** *A time-varying stimulus (top) is encoded by neurons (bottom) with limited capacity C (bottleneck) during a coding window τ with the goal of predicting stimulus of Δ steps ahead.* ***B.*** *Landscape of coding strategies. Optimal code is determined depending on optimisation parameters: efficient coding of past inputs at present time (blue circle), Markovian decoding of future inputs (red circle), while the black circle is largely unexplored. Adapted from [Chalk et al., 2018].*

3.3 Encoding surprise

As we have seen so far, a way to exploit the redundancy in visual inputs in low noise conditions would be to communicate only the unpredictable information whenever possible. In this case, the emphasise of neural responses would be on the stimuli considered 'unexpected', 'deviant', 'oddball', or less probable by the internal model; in contrast, an 'expected', 'standard' or more likely stimuli will elicit a weaker response in comparison. This view allows to frame efficient coding as 'surprise encoding' or 'surprise salience'.

The typical experiment design includes presenting a rare, oddball stimulus interleaved with a standard stimulus of a different physical feature (such as colour, tone, intensity, etc) and significantly higher probability [Ulanovsky et al., 2003]. Gill et al. provide a model for the oddball paradigm by proposing that neurons encode surprise instead of intensity or change in intensity [Gill et al., 2008]. They found that while classical spatiotemporal linear filters fail to replicate the responses of neurons in zebra finch auditory system, 'surprise spatiotemporal receptive fields' succeed to explain the recorded data. These surprise-based method meant that instead of convolving the filter with the sound intensity, they convolved it with the stimulus surprise (here surprise is defined as a negative logarithm of probability given the recent stimulus history and birdsong-related expectations).

As we have seen in Chapter 2, there is also a number of experimental observations in the retina that seem to fit the description of retina making predictions. One of the most straightforward examples is how RGCs anticipate the bar position [Berry et al., 1999], as well as the motion reversal phenomena [Schwartz et al., 2007b]. Moreover, the retina has the ability to dynamically adapt to the changing correlation structure of the environment. Hosoya and colleagues explored adaptation to presented stimulus in even more detail and found how the RGCs receptive fields adapt to efficiently encode the stimulus statistics [Hosoya

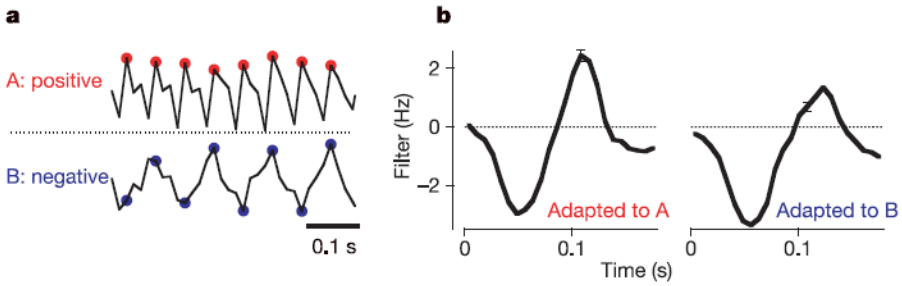

Fig. 3.5 ***A.*** *Adaption to positive (A) or negative (B) correlations in the stimulus. The intensity at each time-step could be predicted by taking the value of the stimulus 60ms into the past with either positive (A) or negative (B) sign.* ***B.*** *Shape of receptive field depends on which stimulus is it adapted to. The strongly biphasic RF (left) suppresses the response to stimuli with positive correlation. It becomes less pronounced once the cell is exposed to the other environment (right). Adapted from [Hosoya et al., 2005].*

et al., 2005].

Their work discovered presence of adaptations to both spatial (image correlations, grating orientation) and temporal features of the images performed by the retinal ganglion cells. An example of retina's ability to adapt to different temporal statistics is showcased by two stimuli with fixed light level and contrast, but different temporal correlation structure (Figure 3.5A). Depending on which stimulus was the neuron exposed to, the shape of the receptive field varies. In fact, when adapted to positive correlation (A in Figure 3.5A), the RF is clearly biphasic (Figure 3.5B, left), effectively suppressing the predictable response. This changes towards less pronounced biphasic RF (Figure 3.5B, right) when the same neuron is presented with a stimuli of negative correlation (B in Figure 3.5A).

Another illustration of such phenomena is the omitted stimulus response (OSR) [Schwartz et al., 2007a]. The retinal ganglion cells adapt dynamically to the period between two flashes, and show a wide range of different responses that could not be directly related to their type [Schwartz and Berry 2nd, 2008]. While the function of this kind of response has been assumed to have a predictive nature

[Schwartz and Berry 2nd, 2008], there is only one study looking into it from the information-theoretic point of view [Chen et al., 2017]. Chen et al. randomly sample the inter-flash interval from a Hidden Markov Model (HMM), and find that in a range of values which corresponds to inter-flash periods typically used in OSR, the mutual information $I(S;R)$ actually carries information about the future flash intervals. However, there was so far no comparison to a quantitative model which would explicitly compute the surprise as a function of stimulus statistics. Given the diversity of spatiotemporal features the retina encodes - contrast, direction, looming motion, to name just a few - a relevant target could also be the surprise.

Chapter 4

Surprise encoding in the retina

4.1 Introduction

A central prediction of the predictive coding theory, elaborated in Chapter 3, is that at low noise neurons should preferentially encode stimuli that are surprising, based on what came before. Several studies suggest that this may be true in the retina, as described in Chapter 2. For example, RGCs respond vigorously to unexpected changes in visual motion [Schwartz et al., 2007b]. Moreover, a series of experiments by Schwartz et al. [Schwartz et al., 2007a, Schwartz and Berry 2nd, 2008] demonstrated how RGCs show a diverse range of adaptive behaviours to repeated patterns of illumination. Notably, they observed that many RGCs responded strongly to violations in the presented temporal pattern: a phenomenon they called the omitted stimulus response ('OSR').

However, despite these results, we still lack direct quantitative evidence relating the responses of RGCs to the degree of 'surprise' for a given stimulus. For example, previous studies looking at the OSR only used a very limited range of different temporal sequences (e.g. n flashes presented in a row, repeated multiple times), and thus it is unclear how RGC responses would vary when certain sequences are more 'surprising' than others. To investigate this, we presented RGCs with extended sequences of stochastically occurring full-field flashes. The stochastic nature of our stimulus meant there was a large range of different levels of surprise, depending on how many flashes had been presented beforehand.

We found that the responses of RGCs to these stimulus sequences could

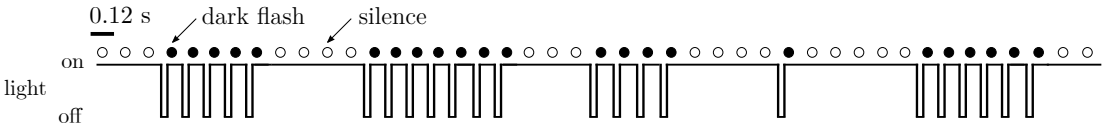

Fig. 4.1 *Stimulus excerpt, showing periodic sequences of dark flashes. Each flash lasts 40 ms with 80 ms inter-flash period. A 120 ms bin containing a dark flash is marked with a dark dot, while a bin without it, called 'silence', is marked by a white dot.*

be well explained by a simple model describing how neurons combine their prior 'expectations' with the recent stimulus history, to encode 'surprise'. The diversity of neuron's responses was captured by describing each cell with its own internal model. Interestingly, while different neurons had similar expectations about which stimuli were most likely, their degree of confidence in their prior expectations varied considerably across cells. Furthermore, these differences were sufficient to explain much of the diversity of neural responses that we observed across the population.

4.2 Results

RGCs responses to flash sequences

We recorded retinal ganglion cells (RGCs) of an axolotl with a multi-electrode array. We presented a visual stimulus, consisting of random sequences of full-field dark flashes, interleaved with periods of silence (Fig. 4.1; see Methods: Stimulus statistics for details). Recorded 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 neurons respond to the absence of a flash, following several flashes presented in a row [Schwartz et al., 2007a]. We thus selected 48 out of 114 single unit responses for further analysis, that showed (i) high quality recording

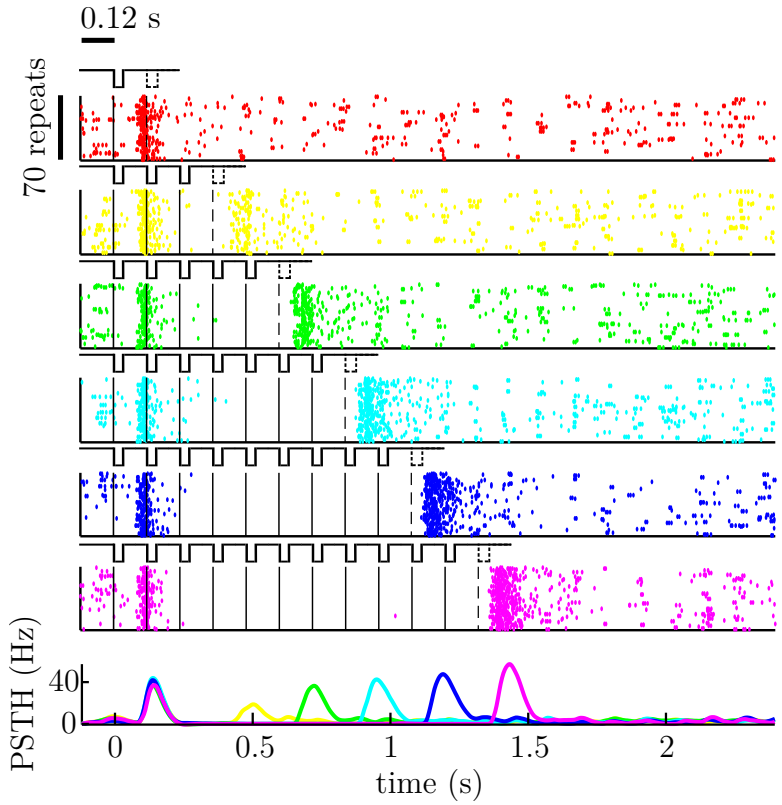

Fig. 4.2 *Raster plot of spikes for an example cell. Solid line represents a flash, while dashed line stands for the timing of omitted flash. Each row represents a response to a different number of consecutive flashes (from 1 to 11 with step 2), with 70 repeats shown in each raster row (order of the repeats is not chronological, but shuffled randomly). The last row combines the peri-stimulus time histogram (PSTH) for different number of flashes (PSTH colors corresponds to the number of flashes). We observe an increase in firing rate after the missing flash, which we consider to be the omitted stimulus response (OSR). Furthermore, the condensed representation of responses show there is an increase in OSR strength with the number of flashes.*

(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 in 120 ms following the omitted flash).

Figure 4.2 shows an example of the responses of one cell to a varying number of flashes presented in a row. As can be seen, this cell fired responded strongly to the first flash in a sequence, and shortly after the sequence had ended,

but where another flash would expected (i.e. the OSR). Moreover, the size of the OSR increased monotonically, depending on 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 computed within each 120ms.

To discover 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 flashes in a row followed by silence (Figure 4.1). 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 difference cells (Figure 4.2).

To see how neural responses depended on all possible stimulus sequences (and not just a series of flashes, presented in a row), we constructed 'tree-plots' (Figure 4.4), consisting of a neuron's average response to all possible sequences of flashes and silences of a given sequence length. Note that the top branch of this tree plot corresponds to the OSR, shown in Figure 4.3A. In Figure 4.4 we can see two cells with quite pronounced differences in structure of the responses, apart from the strongest OSR, i.e. being elicited by silence ('0'; white circle) after several consecutive flashes ('1', black circle).

Modeling 'surprise encoding' by RGCs

We next asked whether RGCs encode surprise. To test this, we constructed a simple model of how RGCs could combine their internal stimulus expectations with the recent stimulus history to compute surprise (Fig. 4.5). Following [Baldi, 2002], we defined surprise at time t, s_t, as the negative log probability of observing a stimulus, x_t, given the recent stimulus history, $x_{<t}$, and

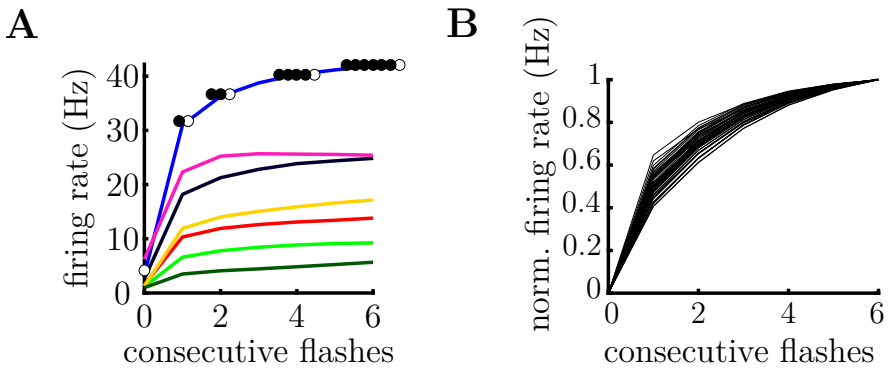

Fig. 4.3 ***A.*** *Diversity of omitted stimulus response (OSR) for seven cells (shown in different colours). Each point of a curve was obtained by averaging the response to one stimulus sequence (represented by circles on the blue line, black for flash, white for silence).* ***B.*** *Range of OSR curves for the full population of 48 cells. Normalized firing rate allows for comparison of cells with different basal firing rates. All cells have in common the rising trend, but vary in the slope at which they achieve their maximum.*

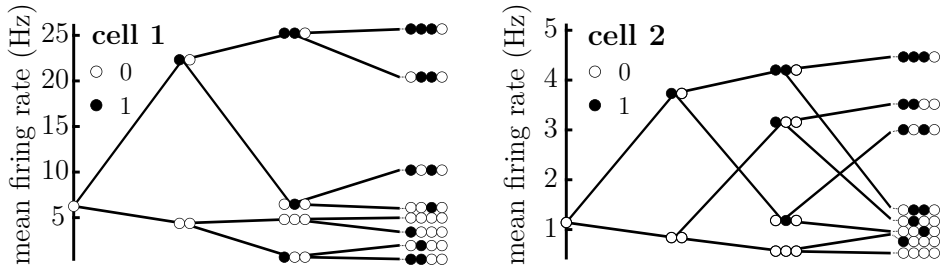

Fig. 4.4 *Tree representation reveals neuron's responses beyond OSR. A dot represents average firing rate in a 120 ms bin with either a dark flash (black circle) or no flash (white circle). The top branch corresponds to the OSR after a different number of consecutive flashes, same as in Fig. Figure 4.1. As the number of flashes before silence increases, the firing rate gets stronger. While both cells show the strongest response to the omitted flash, their other responses are quite different.*

the neuron's internal model of the stimulus statistics (parameterized by θ):

$$s_t = -\log p\left(x_t | x_{<t}, \theta\right) \tag{4.1}$$

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

$$r_t = f\left(gs_t + b\right) \tag{4.2}$$

where g and b are free parameters, and $f\left(\cdot\right)$ is a rectifying 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 or 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, g and b, and parameters of the 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 section: Model

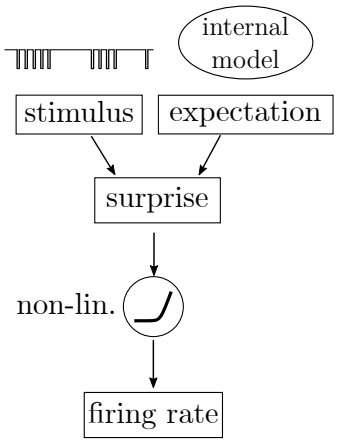

Fig. 4.5 *General schematic of a surprise model. The stimulus is compared to neuron's expectation, which is generated using the neuron's internal model. The surprise is then filtered with static non-linearity to obtain the firing rate. We assume neural responses are generated from a Poisson distribution with this mean firing rate.*

fitting).

Figure 4.6A shows the average firing rate of a single neuron (black) for a given stimulus sequence (above) (we estimated this using the average number of spikes elicited by all repetitions of a stimulus sequence eight time-bins long; see Methods: Data analysis). Despite its simplicity, our model was able to account 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 completely unable to replicate the dependence of the OSR on the number of flashes presented in a row, observed for this (Fig. 4.6B) and many other cells (Fig. 4.6C). 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. 4.6D).

We asked whether it was necessary to have a separate internal model for each cell, or instead the whole population can share one internal model of the stimulus statistics. While it is a plausible assumption, this would obviously not give the range of responses observed in Figure 4.3B. Still, we controlled for that option: we found that even a simple heterogeneous fits the responses better than such homogeneous model (Figure 4.6E).

Adaptive belief model

To account for the observed variations in the OSR with number of consecutive flashes, we next considered a more complex 'dynamic belief' model of surprise encoding. 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*, but must be inferred. Each neuron combines its prior expectations ($p\left(\theta\right)$) with the recent stimulus history ($p\left(x_t, x_{t-1}, \ldots | \theta\right)$) using Bayes' law: $p\left(\theta | x_t, x_{t-1}\right) \propto p\left(x_t, x_{t-1}, \ldots | \theta\right) p\left(\theta\right)$. With binary stimuli, this results in a simple expression for the inferred probability of

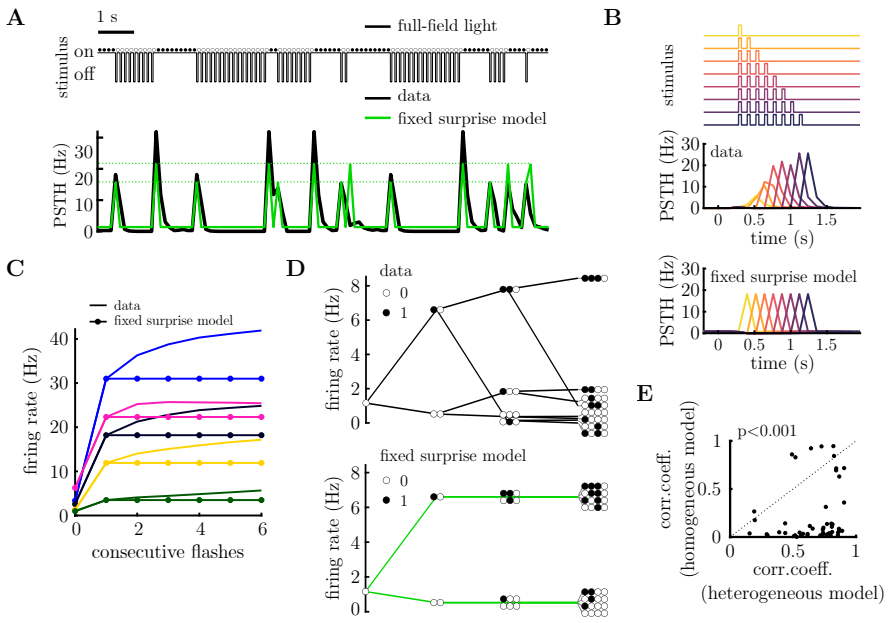

Fig. 4.6 ***Only basic features of the response are captured by a simple surprise model.*** ***A.*** *Stimulus excerpt (top, black) and corresponding response peri-stimulus time histogram (PSTH, bottom, black) compared to predicted firing rate for the fixed surprise model (bottom, green). The fixed model has limited flexibility, effectively displaying only 4 possible values of firing rate (2 shown in dashed green line). Pearson correlation coefficient: r=0.82.* ***B.*** *Flash sequences of different length (top); PSTH of the responses (middle) and model prediction (bottom) when presented with a variable number of consecutive flashes. Each colour corresponds to a different length of the flash sequence. The fixed surprise model predicts same strength of OSR independent of the number of flashes, which is not what can be seen in the data.* ***C.*** *Average responses to an increasing number of flashes is not reproduced by the fixed surprise model, which can take into account only the previous state (flash or no flash). The flash history beyond that does not play a role in estimation of expectation. Different colours stand for individual cells.* ***D.*** *Tree-plot of data (left) and fixed model prediction (right). The depth of the tree corresponds to the number of observed time-bins; for every depth, combinations of silence ('0'; white circle) and flashes ('1', black circle) is shown. Other half of the tree, for codewords ending in '1', is omitted. Fixed surprise model can capture the mean response for codewords of length 1 and 2, but not beyond.* ***E.*** *Comparison of heterogeneous (each cell has its own expectation) to homogeneous model (one for the whole population). For majority of the cells the fit is significantly better with individual expectations.* $p = 6.64 \cdot 10^{-9}$*, Wilcoxon rank sum test.*

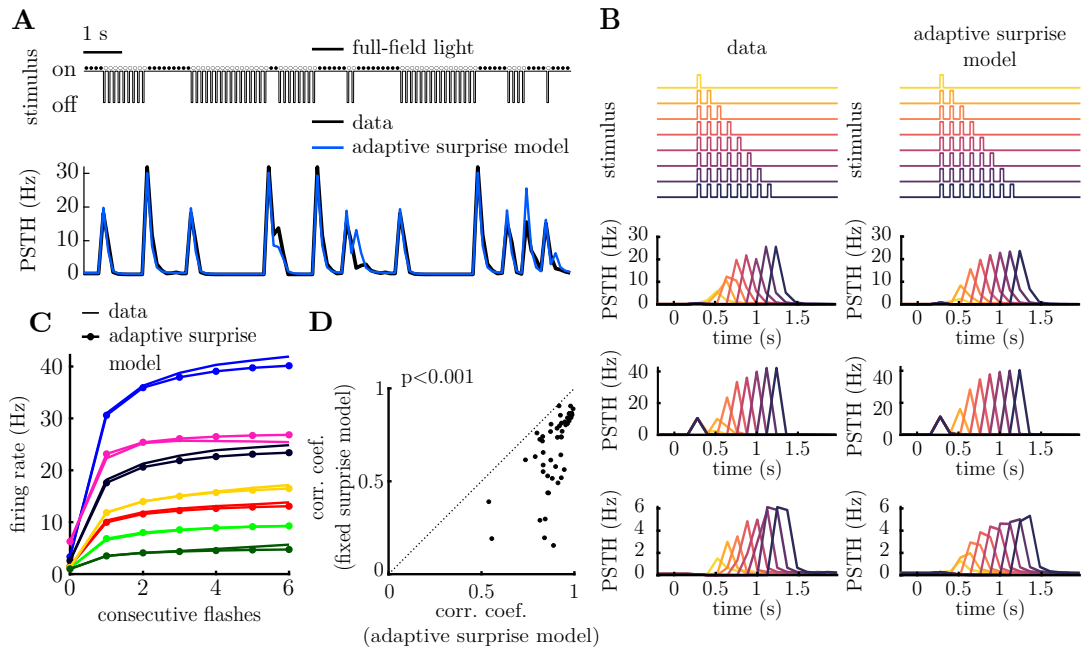

Fig. 4.7 ***Adaptive surprise model can capture both qualitative and quantitative features of the responses.*** ***A.*** *Stimulus excerpt (top, black) with response firing rate (bottom, black) compared to predicted firing rate for the fixed surprise model (bottom, blue). Pearson correlation coefficient r=0.95.* ***B.*** *Responses to flash sequences from 1 to 8 flashes (top). Condensed OSR display showing responses and model predictions for 3 cells (rows 2-4, one cell per row). Adaptive surprise model captures not only the overall trend of OSR rise, but also different decay patterns.* ***C.*** *Rise in OSR with the number of consecutive flashes for seven cells (each one represented with a different colour), data (solid line) vs. adaptive surprise model (solid line with circles). The range on x-axis starts from the mean response to silence, ending in mean response to silence after 6 consecutive flashes.* ***D.*** *Goodness of fit comparison, Pearson correlation coefficient for two surprise models. Adaptive surprise model outperforms the fixed surprise one. Each dot represents a cell in the population. Dashed line is the unity line.* $p = 5 \cdot 10^{-11}$*, Wilcoxon rank sum test.*

seeing x_t, given $x_{t-1} = i$:

$$p\left(x_t = 0 | x_{t-1} = i, x_{t-2}, \ldots\right) = \frac{n_{i\to 0} + \beta_{i\to 0}}{n_{i\to 0} + n_{i\to 1} + \beta_{i\to 0} + \alpha_{i\to 1}} \tag{4.3}$$

$$p\left(x_t = 1 | x_{t-1} = i, x_{t-2}, \ldots\right) = \frac{n_{i\to 1} + \alpha_{i\to 1}}{n_{i\to 0} + n_{i\to 1} + \beta_{i\to 0} + \alpha_{i\to 1}}, \tag{4.4}$$

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 $\alpha_{i\to 1}$ and $\beta_{i\to 0}$ are parameters of the prior, that can be thought of as the 'effective' number of observations of the transition $i \to j$. We assume that the parameters of the prior are different for each neuron. Note that, in the limit where the prior is very strong (i.e. $n_{i\to j} \gg \alpha_{i\to 1}$ and $n_{i\to j} \gg \beta_{i\to 0}$), 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 neuron's had 'infinite' memory then, given a sufficiently long stimulus sequence, the prior would have no effect (as we would always have $n_{i\to j} \gg \alpha_{i\to 1}$ and $n_{i\to j} \gg \beta_{i\to 0}$). Instead, we assume a more realistic model where neuron's have a finite memory, and $n_{i\to j}$ are estimated using a leaky integration of past observations (see Methods: Internal model of stimulus). This required one additional parameter (the time-scale of integration), which we kept fixed for all neurons.

We fitted the 4 parameters of the prior (plus 2 parameters of the LN model) for each neuron, using maximum likelihood, with a Poisson noise model [Pillow et al., 2005]. Figure 4.7A shows the predicted firing rate for one neuron (blue) to a short stimulus sequence (above). This 'adaptive belief' model was able to capture aspects of the neuron's response that could not be accounted for by the previous 'fixed belief model'. For example, it could capture how the size of the OSR increased with the number of flashes presented in a row (Fig. 4.7B-C). Further, it captured individual differences in the OSR decay for different neurons (compare cells 1-3 in Fig. 4.7B). Overall, the correlation between the estimated

firing rates and the model prediction was significantly higher for the adaptive, compared to the fixed, belief model (Fig. 4.7D).

To investigate further how the adaptive belief model could account for the diverse responses of different cells, we plotted 'tree-plots' describing the average firing rates predicted by the model for stimulus sequences of different lengths (Fig. 4.8). 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. 4.9A), compared to the fixed belief model which reduced dramatically for tree depth greater than 2. On the other hand, we wondered if the structural similarity is also better described by the adaptive surprise model. We computed the 'edit distance': minimum number of permutations needed to arrive at the recorded neuron's tree structure (Fig. 4.9B).

A control: internal Markov model with longer history

To assess the validity of our adaptive belief model, we decided to compare it to a more complex fixed belief model, with a similar number of free parameters. To do this, we implemented a 'Markov-2 model' in which the probability of observing a flash is assumed to depend on the observed stimulus in the previous two time-bins. This model has 4 parameters, $(\theta_{ij} = p(x_{t+1} = 1|x_t = i, x_{t-1} = j))$, which is the same as the adaptive belief model (putting aside the leak parameter, which was kept the same for all cells). The behaviour of this model is shown in Figure 4.10. While the Markov-2 model outperformed the Markov-1 model described earlier, it could not account for increases in the OSR that occurred for sequences of more than 2 consecutive flashes (Fig. 4.10B-C), or structure in the tree plots at a depth greater than 2 (Fig. 4.10D). Finally, the correlation coefficient between predicted and observed firing rates was significantly worse for

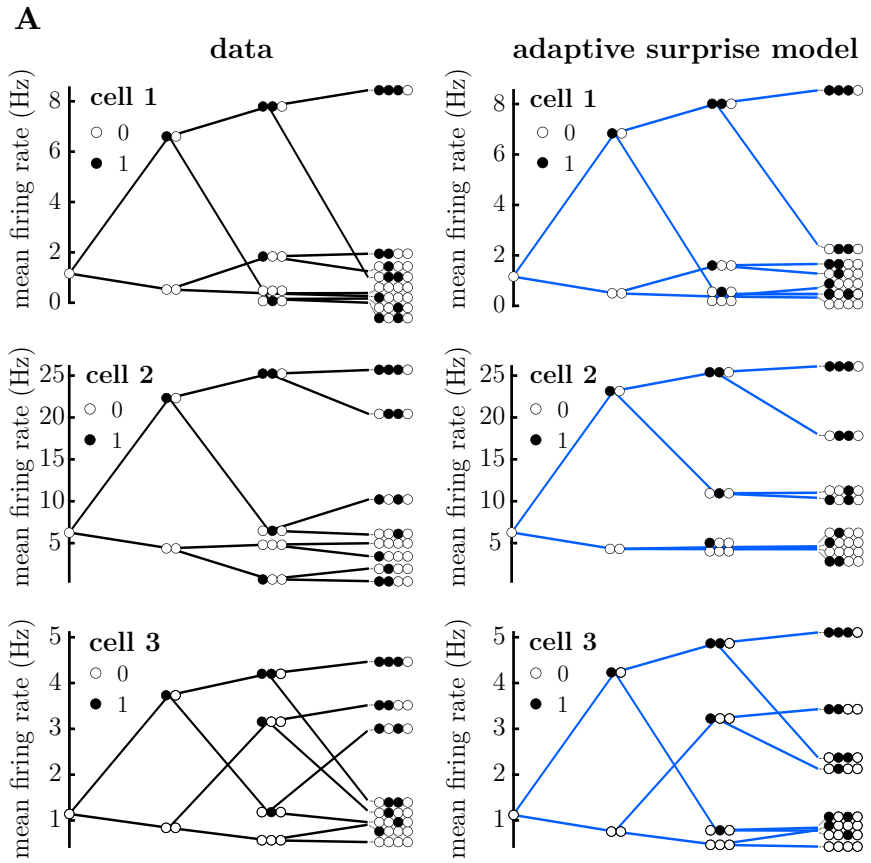

Fig. 4.8 *Tree diagram of responses (left) vs. model predictions (right). A dot represents average firing rate in a 120 ms bin with either a dark flash (black circle) or no flash (white circle). The top branch corresponds to the OSR after a different number of consecutive flashes, same as in Fig. Figure 4.1. The adaptive model manages to reproduce the tree structure, capturing different possible history patterns (here only patterns ending in silence are shown).*

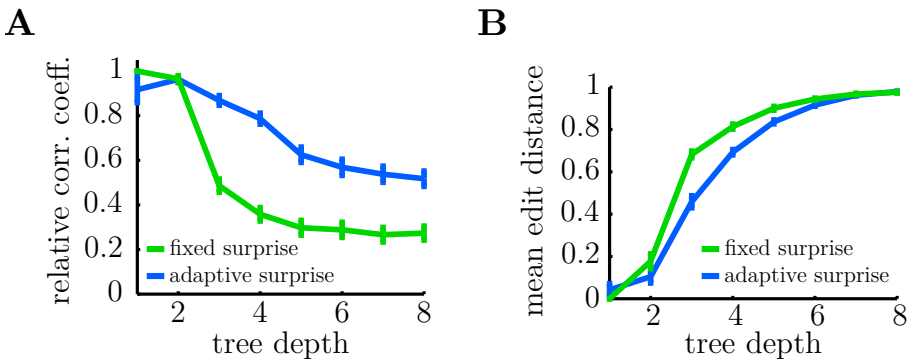

Fig. 4.9 ***Response diversity is better captured with the adaptive surprise model. A.*** *Correlation coefficient between the response-based tree and model-based tree is computed at each tree depth. On overall, the tree structure is better captured by the adaptive model, even though for depth 1, aka average response to flash or silence, the fixed model outperforms it.* ***B.*** *Normalized edit distance vs. the tree depth: the number of operations needed to transform the predicted tree order at given depth to actual, neuron's tree structure, divided by the total number of states in that depth ($2^d epth$). The fixed surprise model performs worse in this regard, until both models arrive at the point of almost completely shuffled structure (for that case, edit distance = 1).*

the Markov-2 model than the adaptive surprise model (Fig. 4.10E) despite them having the same number of free parameters ($p = 5 \cdot 10^{-3}$, Wilcoxon rank sum test).

It would also be possible to further increase the order of the Markov model and consider longer look-back when estimating the probability of current stimulus state. However, this would lead to an exponential increase in the number of parameters used to describe a total of 2^N code-words (here $N = 8$). Moreover, it is not feasible in practice, since the model complexity also increases as 2^M, where M is the order of Markov model.

Differences in the prior between different cells

We were interested to see how the inferred expectations (the 'prior') varied across different cells. Recall that the parameters of the prior ($\alpha_{i\to 1}$ and

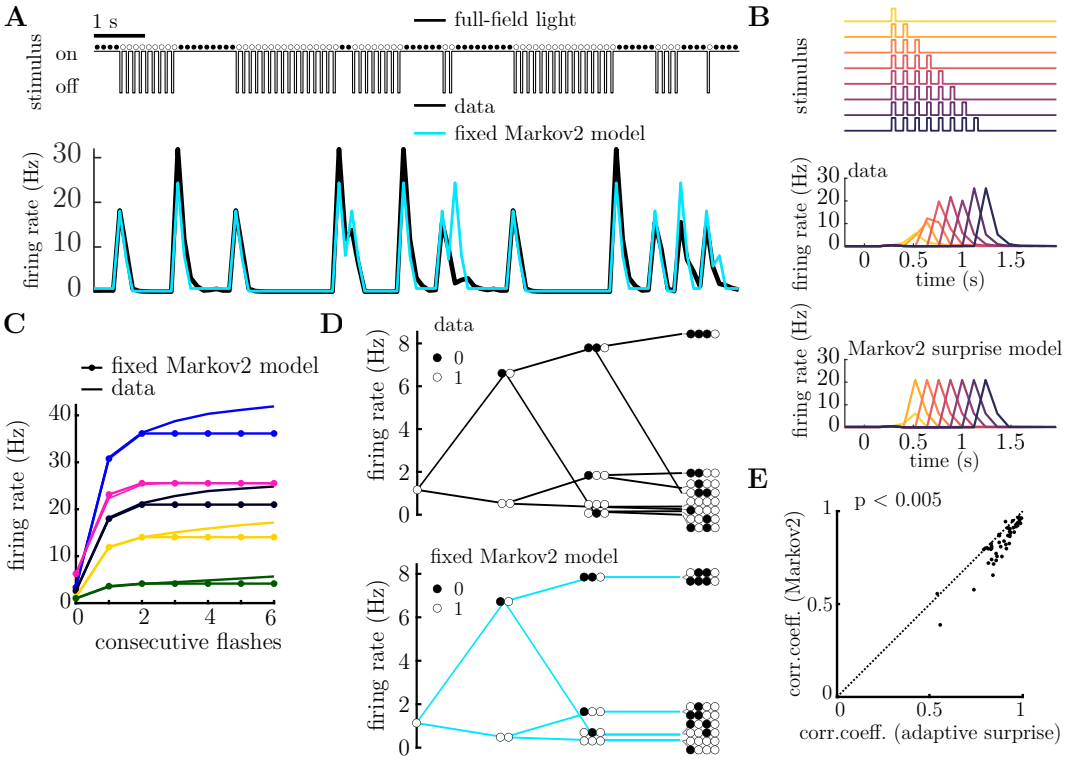

Fig. 4.10 ***A fixed model with longer past - Markov2 - cannot match the adaptive surprise model predictions.*** ***A.*** *Stimulus excerpt (top, black) with average firing rate of the responses (bottom, black) compared to predicted firing rate for the Markov2 surprise model (bottom, cyan). Pearson correlation coefficient: r=0.91.* ***B.*** *Mean firing rate of the responses and model prediction to variable number of flashes. Each colour corresponds to a different number of consecutive flashes.* ***C.*** *OSR increase with the number of flashes is not completely reproduced by the fixed surprise model, which can take into account only the previous two states. The flash history beyond that does not play a role in estimation of expectation. Different colours stand for individual cells.* ***D.*** *Markov2 surprise model can capture the mean response for codewords up to length 3, but not further. Other half of the tree, for codewords ending in 1, is omitted.* ***E.*** *Adaptive surprise model achieves better correlation coefficient than Markov2. Each dot represents a cell in the population. Dashed line is the unity line.* $p = 5 \cdot 10^{-3}$*, Wilcoxon rank sum test.*

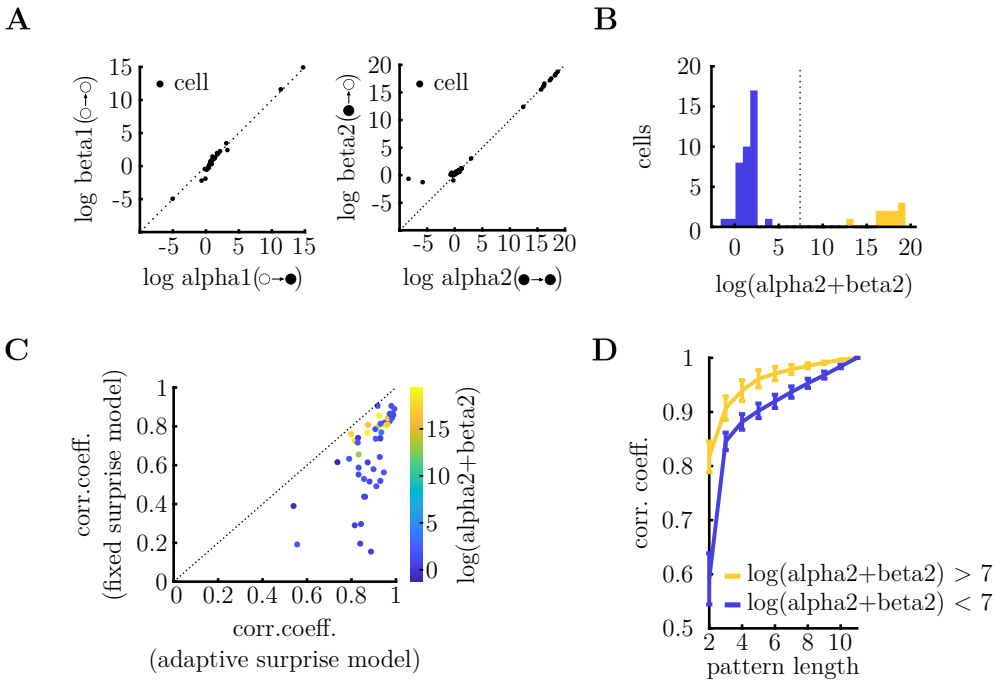

Fig. 4.11 ***Parameters of internal model.*** ***A.*** *Scatter plot of parameters of prior ($\alpha_{i\to 1}$ and $\beta_{i\to 0}$) for each cell, describing prior belief for transitions originating in 'no flash' state (left) or 'flash' (right). For both pairs of parameters, the ratio between eff. number of transitions for changing state and staying in the same state is around 1 for all the cells, while the confidence in prior varies for different cells (see text).* ***B.*** *Histogram of $\alpha_{1\to 1}$ ('alpha2') and $\beta_{1\to 0}$ ('beta2') values i.e. confidence in prior observations. The population can be split into two groups of cells: the ones with low confidence (blue) and high confidence (yellow).* ***C.*** *Correlation coefficient of the adaptive surprise model vs. the fixed surprise, with cells coloured according to $\log(\alpha_{1\to 1} + \beta_{1\to 0})$. Cells that were fitted well by the fixed surprise model, i.e. the ones near the unity line (dashed), were the ones with high value of confidence (yellow).* ***D.*** *Correlation between the average response to pattern of different length (x-axis) with the average response to pattern of length 10. We split the population of recorded cells into two groups, based on $\log(\alpha_{1\to 1} + \beta_{1\to 0})$, which is either low (yellow) or high (blue), and show the group average correlation coefficient. Error bars represent standard error.*

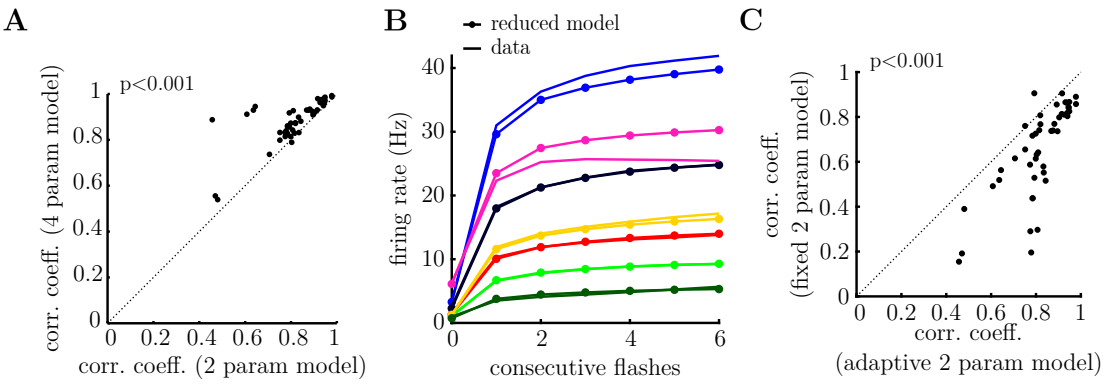

Fig. 4.12 ***Reduced model with fixed prior mean. A.*** *The reduced model of 2 parameters per cell is fitting the data similarly to full adaptive surprise model of 4 parameters per cell (see text for fitting details).* $p = 3 \cdot 10^{-9}$. ***B.*** *Average response at stim=0 for 0 to 6 flashes, 7 different cells are shown (different colours). The OSR trend is well-fitted by a simplified adaptive surprise model.* ***C.*** *Comparison of fixed surprise model to the reduced adaptive surprise model (both have 2 parameters per cell). The reduced model still outperforms the fixed surprise for almost all cells.* $p = 4 \cdot 10^{-5}$, *Wilcoxon rank sum test.*

$\beta_{i\to 0}$ for $i = 0/1$), can be considered as the number of 'effective prior observations' of different transitions. We plotted how these parameters of the prior varied for different cells (Fig. 4.11A). Interestingly, we found that while for different cells there was a large variation in the total number of effective observations ($\alpha_{0\to 1}+\beta_{0\to 0}$ & $\alpha_{1\to 1}+\beta_{1\to 0}$), their ratio ($\frac{\alpha_{0\to 1}}{\beta_{0\to 0}}$ & $\frac{\alpha_{1\to 1}}{\beta_{1\to 0}}$) remained relatively constant across cells. Thus, while the confidence in prior, which determines how much weight is accorded to prior expectations versus new observations, varied greatly across cells, the mean of the prior (which depends on just the ratio between different parameters) was relatively constant. Moreover, focusing on the prior parameters related to transitions to 'no flash' state ('0'), on the level of population there seems to be two clearly separable groups of $\alpha_{1\to 1} + \beta_{1\to 0}$ values (Fig. 4.11B).

We wanted to see what does it mean for a cell to have a weak or strong prior and how it would be reflected in the responses. What reasoned that cells with a high total effective observations would not adapt their posterior belief as much

depending on local observations, and hence their response would be well predicted by the fixed surprise model. This turned out to be the case: if we colour-code the cell's goodness of fit based on its confidence in prior for both adaptive and fixed surprise models, the cells with high $\log(\alpha_{1\to1} + \beta_{1\to0})$ (parameters of the prior relevant for the OSR) can all be found near the unity line (Figure 4.11C).

The main difference between cells with low and high confidence in prior could be how strongly their responses reflect the local history. For a cell with a strong prior, the stimuli from the recent past will influence its responses less; therefore, the similarity between average firing rate computed with shorter and longer look-back will be higher than in the case of a cell which is more uncertain about the prior. To illustrate the degree of this similarity, we compute the correlation coefficient of the average firing rate for stimulus sequence of length 10 with the stimulus sequence of shorter history (x-axis in Figure 4.11D).

Knowing the distribution of prior parameters is bimodal (Figure 4.11B), the population can be split into two groups based on the confidence in prior. The cells with high confidence in prior (yellow) hold the similarity between responses higher than the cells with low confidence in prior, indicating there is less to lose when shortening the code-words. Thus, cells with high confidence in prior could be thought of as less-adaptive to surprise, unlike the other extreme where cells integrate over a longer history of recent stimulus transitions. While the fixed model can account for the cells with high confidence in prior, the local updates present in adaptive model are necessary to explain the responses of cells with low confidence in prior.

Given the ratio between prior parameters remained approximately constant (Figure 4.11A), we can simplify the adaptive surprise model by assuming the prior parameters lie on the unity line, and their position on the line is described with one parameter, the total effective number of observations (two per cell in total, one for each pair of transitions). We fitted this 'reduced adaptive' surprise

model without a noticeable decrease in performance Figure 4.12A, B. Moreover, when compared to the fixed surprise model, the reduced adaptive surprise model retains clear advantage in fitting the responses Figure 4.12C despite having the same number of parameters. It emphasises the importance of taking into account recent transitions as well as the prior belief in estimation of surprise.

4.3 Discussion

Here we present the evidence that retinal ganglion cells encode surprise. We showed that the RGCs activity is well-fitted by a simple normative model that assumes neural responses are proportional to surprise. It has been suggested that there are many different functional goals that retinal ganglion cells fulfill [Gollisch and Meister, 2010]. These results might present an addition to the already impressive list of features retina can compute and convey to the downstream areas: mean luminance [Barlow and Levick, 1969, Shapley and Enroth-Cugell, 1984], local and global contrast [Demb, 2008, Kastner and Baccus, 2011], direction [Vaney et al., 2012], position [Deny, 2016], speed [Deny, 2016], object motion [Ölveczky et al., 2003, Borst and Euler, 2011], approaching motion [Münch et al., 2009].

Our modelling work suggests that neurons' expectations might not be equivalent to the true stimulus statistics. We instead find that the diversity of RGC responses can be captured better by computing surprise with respect to an individual expectation of each cell. This contrasts with previous work, which assumed retinal neurons are 'ideal observers' of the environment, which gives the theoretical upper bound for a performance on a given task [Geisler, 1989, Geisler, 2003, Chichilnisky and Rieke, 2005, Smeds et al., 2019]. Here, the task would be to predict the length of flash sequence given the previously observed sequences. Therefore, such 'ideal observer' neuron could adapt perfectly to

underlying statistics [Hosoya et al., 2005, Chichilnisky and Rieke, 2005]. However, we found that this is not the case for the complex temporal stimuli we presented, which motivated us to look further into other possible explanations (see Appendix A for details).

The modelling framework we use here is adapted from psychophysics work explaining several experimental findings on sequential effects, where subject's responses depend on local regularities in the sequence [Meyniel et al., 2016]. Meyniel and colleagues formulate a Bayesian model which combines recent stimulus with a uniform prior using leaky integration, enabling it to dynamically estimate the probability of presented stimuli. The surprise is then computed with respect to this constantly updated expectation. What Meyniel and colleagues found was that all six of the experimental studies could be explained by the same class of models, which learned the transition probabilities of the sequence and thus could reproduce the responses of subjects depending on the probability of the presented stimulus. The fact that the retina's activity can be predicted with a similar model might suggest similar basic principles in computations in brain areas of completely different complexity and tasks (see Chapter 5 for more details).

Although the sequences of full-field flashes are an artificial stimulus, they present a good starting point to probe how retina responds by manipulating the level of surprise in the stimuli. An advantage of having a carefully varied feature of the stimulus is the clear assumption of what is encoded, giving the possibility to design a model that should then be validated on natural scenes [Rust and Movshon, 2005]. Given the relationship between surprise and RGC responses that we show, it would be interesting to next explore a more naturalistic stimulus. One direction would be to expand the set of possible stimulus states which was in our case binary. Another possible path could be to use a stimulus which is not spatially uniform. Related studies are done in visual cortex when studying the mismatch response: a certain part of the visual field is perturbed i.e. carries

surprising information [Keller et al., 2012]. However, the difficulty of defining surprise for a more complex spatiotemporal scene can be a limiting factor.

The parameters of neurons' expectations allow us to investigate in more detail where the range of responses comes from. Surprisingly, we find that RGCs have in common the prior belief that staying in the same state ('0' or '1') is equally probable as changing the state. The confidence in that belief is what tells different cells apart. An open question remains how do the two types of surprise-encoding cells we found relate to functional cell types: we could ask whether there is a relationship between their type and how they respond to surprise. For instance, it would be beneficial to investigate if the two distinct subsets of cells have in common other traits apart from confidence in their expectation.

Lastly, the current analysis treats all neurons as individual processing units and does not consider the interactions of nearby neurons [Pillow et al., 2008]. Taking into account the joint activity might reveal if, and how, is encoding a feature like surprise influenced by the spatial proximity of cells of same or different type [Nirenberg and Latham, 1998, Ferrari et al., 2018, Roy et al., 2021].

4.4 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 raw electrode traces were 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 salamander. The flash duration of 40 ms, with 80 ms period between the flashes, ($\sim$12 Hz frequency) were taken from Schwartz et al. [Schwartz et al., 2007a].

Stimulus statistics

We generated the flash sequence using a statistical model where the each length of flashes, presented in a row, was given a certain probability. The number of consecutive flashes was drawn from three distributions, each shown for 20 minutes. All three distributions had the same mean number of 7 flashes presented in a row 7, and range of 1 to 16 flashes in a row. The length of each sequence of flashes was drawn from a negative binomial distribution $\mathrm{NB}(r, p)$, with parameter $p = 0.98$ (either one flash or very long sequences), 0.8 (equivalent to geometric distribution) and 0.01 (flash sequence length clustered around the mean), respectively. The second parameter, r was calculated to maintain a constant mean value m as: $r = \frac{(1-p)m}{p}$. We found no significant differences in neural responses to these 3 distributions (see Appendix). Therefore, we concatenated neural data from all three stimulus distributions for the remainder of our analysis.

Data analysis

To generate the spike raster plots shown in figure 4.2, we aligned the spiking responses of neurons to the same sequence of n flashes presented in a row (shown above). The peri-stimulus-time-histogram (PSTH) in the bottom row of Figure 4.2 was computed by averaging the spike count recording over all the stimulus repeats, before averaging over a 5 ms time bin.

For the remainder of the analysis, we discretised the neural responses and stimulus in 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).

Neural model

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

$$\lambda_t = f(gs_t + b) \tag{4.5}$$

where f is a non-linearity, and g 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:

$$s_t = -\log p(x_t|x_{t-1}, x_{t-2}, \ldots, \theta) \tag{4.6}$$

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 of stimulus statistics

The computed surprise depends on each cell's internal model of the stimulus statistics. We 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 model parameters

Next, we considered an 'adaptive belief model' where the transition probabilities, θ_i, are not known in advance, but must be inferred by combining each cell's prior belief with newly observations, using Bayes' law.

The likelihood of a binary stimulus x_t at time t, is described using a Bernoulli distribution:

$$p\left(x_t|x_{t-1}=i,\theta_i\right)=\theta_i^{x_t}\left(1-\theta_i\right)^{1-x_t} \tag{4.7}$$

The posterior distribution over a parameter θ_i is given by:

$$p\left(\theta_i|x_t,x_{t-1},\ldots\right)\propto\theta^{\alpha_t^i-1}\left(1-\theta\right)^{\beta_t^i-1} \tag{4.8}$$

where α^i and β^i are parameters, that are updated over time, as more evidence is accumulated.

We next explain how to update these parameters on each time-step. In effect, we can divide the stimulus into two parts: transitions from $x_{t-1}=0$, and transitions from $x_{t-1}=1$. Here, we will consider only transitions from $x_{t-1}=1$, although exactly the same arguments apply to the remaining transitions. For notational simplicity, we will neglect the subscript i, describing which type transition of transition we are referring to.

On each time-step we update $p\left(\theta|x_t,x_{t-1},\ldots\right)$ using Bayes' law, such that: $p\left(\theta_i|x_t,x_{t-1},\ldots\right)\propto p\left(\theta_i|x_{t-1},\ldots\right)p\left(x_t|x_{t-1}=i,\theta_i\right)$. This gives rise to the following update rules for the parameters of the posterior:

$$\alpha_{t+1} \leftarrow \alpha_t+x_t \tag{4.9}$$

$$\beta_{t+1} \leftarrow \beta_t+(1-x_t) \tag{4.10}$$

Finally, the probability of observing x_t being a flash given previous observations is given by:

$$\begin{aligned} p\left(x_t = 1|x_{1:t-1}\right) &= \int_\theta p(x_t = 1|\theta)\ p\left(\theta|x_{1:t-1}\right) \\ &= \int_\theta \theta\ p\left(\theta|x_{1:t-1}\right) \equiv \langle\theta\rangle_{p(\theta|x_{1:t-1})} = \frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}} \end{aligned} \tag{4.11}$$

Note that after a certain burn-out period, this model will be equivalent to a fixed Markov model described earlier with 2 parameters, $p(x_t = 0|x_{t-1} = 0)$ and $p(x_t = 1|x_{t-1} = 0)$.

Leaky integration of model parameters The statistics of the external world are not static, but change in time. An optimal Bayesian model would be the one assuming there is a non-zero probability of transition matrix changing between two observations, called 'dynamic belief model'. However, inferring these parameters would require numerical integration, which might be a too difficult requirement for early sensory neurons. Meyniel et al. found that the dynamic belief model can be approximated by a 'forgetful' model, which values recent observations more strongly than the ones in the past [Meyniel et al., 2016]. While achieving similar performance, this leaky integration model is thus more biologically realistic.

In case of the adaptive surprise model, we perform Bayesian prior update with leaky integration. The prior of each neuron is updated by likelihood, with a decay η kept constant for the whole population. This gives recurrence relation for parameters of beta distribution:

$$\begin{aligned} \alpha_{t+1} &\leftarrow (1-\eta)\alpha_t + x_t \\ \beta_{t+1} &\leftarrow (1-\eta)\beta_t + (1-x_t) \end{aligned} \tag{4.12}$$

When $\eta = 0$ we have the perfect integration model. When $\eta > 0$, α will eventually decay to 0 in the case of no observations.

We can expand out the recurrence relation for α as follows:

$$\begin{aligned}\alpha_{t+1} &= (1-\eta)\left((1-\eta)\alpha_{t-1} + x_{t-1}\right) + x_t \\ &= \sum_{k=0}^{\infty}(1-\eta)^k x_{t-k}\end{aligned} \tag{4.13}$$

If we have a prior assumption that $\alpha = \alpha_0$, and $\beta = \beta_0$, then we can write:

$$\begin{aligned}\alpha_{t+1} &\leftarrow (1-\eta)\alpha_t + \eta\alpha_0 + x_t \\ \beta_{t+1} &\leftarrow (1-\eta)\beta_t + \eta\beta_0 + (1-x_t)\end{aligned} \tag{4.14}$$

In this case, α will decay to α_0.

Unlike in [Meyniel et al., 2016] where the prior was uniform and same for all subjects, we also learned the parameters of the prior, i.e. α_0^i and β_0^i. Each cell had a prior described with two pairs of beta distribution parameters: $\alpha_i, \beta_i, i = 0/1$.

Model fitting

The model was fitted using maximum likelihood (ML) [Doya et al., 2007]. The parameters fitted to the data were the ones describing the internal model, as well as the gain and bias. In case of fixed model which is a Markov model of order 1, there was 2 parameters of the prior related to surprise: p_{01} and p_{10}. For the Markov model of order 2, there were 4 parameters. For the adaptive surprise model, there were 4 parameters, the effective number of observations describing the transitions: $\alpha_{0\to1}, \beta_{0\to0}, \alpha_{1\to1}, \beta_{1\to0}$. All three models were fitted using algorithms with multiple starting points (MultiStart in MATLAB, 50 starting points, random initial parameters).

The gain and the bias were estimated using Newton method, in order to have the optimal filter for each internal model candidate. Consider an LN model with $d \times 1$ inputs, $\boldsymbol{x}$, and linear weights, $\boldsymbol{w}$. The mean firing rate is $f\left(\boldsymbol{w}^T\boldsymbol{x}\right)$

where $f(\cdot)$ is some positive and convex function. The log-likelihood for a Poisson model, and it's gradient and hessian are as follows:

$$\mathcal{L} = \sum_t n_t \log f_t - f_t \tag{4.15}$$

$$\nabla_w \mathcal{L} = \sum_t \boldsymbol{g}_t \left(\frac{n_t}{f_t} - 1 \right) \tag{4.16}$$

$$\nabla_w^2 \mathcal{L} = \sum_t \boldsymbol{A} \left(\frac{n_t}{f_t} - 1 \right) - \boldsymbol{g}_t \boldsymbol{g}_t^T \left(\frac{n_t}{f_t^2} \right) \tag{4.17}$$

To maximise the log-likelihood, we update $\boldsymbol{w}$ using the Newton method as:

$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_t - \left(\nabla_w^2 \mathcal{L} \right)^{-1} \nabla_w \mathcal{L} \tag{4.18}$$

In this case, the weight was a vector of 2 elements, the gain and the bias.

Statistical tests. All reported p-values were computed using Wilcoxon signed-rank test, taking into account that the data is not normally distributed. **Programming tools.** The data analysis and model fitting were done in MATLAB R2021a. Code and data will be available upon paper acceptance.

Chapter 5

Discussion

The assumption of neuron's encoding surprising stimuli has been used to understand neural activity in various parts of the brain, across numerous animal species. We will mainly consider comparison of the retina to other sensory modalities, mostly because the notion of surprise on a cognitive level assumes longer timescales than the ones in the retina.

5.1 Surprise-related responses in the sensory cortex

When presented with a sequence of repeated stimuli followed by a novel one, neurons in both visual and auditory cortex respond stronger to an unexpected event than to a repeated one: this phenomena is known generally as the mismatch response, or mismatch negativity if recorded with electroencephalography (negativity is due to the difference between expected and unexpected event response being negative) [Näätänen et al., 1978]. There is a long history of studying mismatch negativity (MMN) in both auditory [Näätänen et al., 2007, Näätänen and Kreegipuu, 2012] and visual cortex [Tales et al., 1999, Pazo-Alvarez et al., 2003, Stefanics et al., 2014], with promising extensions to applications in neurological disorders [Keller and Mrsic-Flogel, 2018, Kremláček et al., 2016].

In the auditory cortex, it was found that the strength of responses to two different tone frequencies depend on which one is less common in the sequence. An oddball stimulus (i.e. the tone presented with a smaller probability) represents

a violation of the prediction and therefore elicits a stronger response [Goldstein et al., 2002, Ulanovsky et al., 2003]. In the case of a more complex and naturalistic stimuli, such as zebra finch songs, Gill and colleagues found that auditory neural responses correspond better to stimulus surprise than the intensity of the stimulus or its change [Gill et al., 2008]. In our work, we show similar claim seems to stand for the retina, since ganglion cells transmit information about surprise instead of stimulus luminosity. This is one of the most important conclusions of our work: there is a direct link between stimulus surprise and the responses.

Later work by [Rubin et al., 2016] has found that the responses of neurons in primary auditory cortex to the oddball sequences can be fitted using a model based on compressed representation of the past, formalized by Information Bottleneck. Rubin et al. show that neurons make predictions using an internal representation of the stimulus sequence.

Our research also found that the retina could have an internal representation of the stimulus statistics instead of being perfectly adapted to the stimulus. Thanks to a small number of interpretable parameters that describe the adaptive surprise model, it was possible to investigate each neuron's internal model of the stimulus statistics. What is especially interesting is that the prior knowledge is common for all cells in the recorded population, while the confidence in this prior is responsible for the distinction in responses across different cells. This might be interpreted in the context of development: all cells were exposed to the same natural statistics prior to the experiment, and therefore they all expect similar temporal regularities.

Analysing the fitted prior parameters also revealed there were two distinct groups of cells: one with a strong preference towards local experience (termed 'adaptive surprise cells'), and those with high confidence in their prior knowledge ('non-adaptive surprise cells'). It also provided an explanation for how the fixed surprise model was able to reproduce the behaviour of non-adaptive surprise cells.

In fact, the fixed surprise model can be framed as an adaptive surprise model in the limit of having an infinitely long memory (i.e. small leak).

5.2 Future directions

How fast can the retina adapt to different types of visual stimuli is an open question. The time-scales vary greatly: from hundreds of milliseconds in contrast adaptation and motion reversal [Smirnakis et al., 1997, Schwartz et al., 2007b], to order of seconds or tens of seconds for intrinsically photosensitive RGCs to adapt to mean luminance levels [Allen et al., 2017]. In the case of contrast adaptation, two distinct types of adaptation have been observed: the fast one happening on the scales of 100 ms, and the slow one occurring over 1-10 s [Smirnakis et al., 1997, Baccus and Meister, 2002]. To certain extent, it is possible to draw a parallel between these two timescales of adaptation with what we observe in our data. In the adaptive surprise model presented here, the importance of the observed stimulus state is halved after seeing three new stimuli. This corresponds to $3 \cdot 120$ ms $= 360$ ms, which is on the same order as the fast adaptations to contrast. On the other hand, time needed for the ratio between different transitions to vary less than 20% is around 80 seconds. Whether the 'non-adaptive surprise' cells do not adapt to the stimulus statistics at all, or they do so at a longer time scale, is a question that can be addressed by varying the time scales of the stimulus. Since no cell typing was performed, it is not impossible that these cells are actually the intrinsically photosensitive RGCs, which are known to have a different set of functions than the 'classic' RGCs [Aranda and Schmidt, 2020].

Future work could try to find the relationship between the learned prior parameters with a stimulus statistics that is less hidden and less complex. Similar to the oddball paradigm, the ratio of '00' to '01' transitions can be manipulated

to explore two environments in which they are either equally probable, or where staying in the same state is more probable (the opposite case, where changes between states are more frequent, might give too noisy responses). Recording a larger set of cells might also provide some insights into how robust is this relationship.

5.3 General relevance

A long-term goal for studies of functioning of the healthy retina might be to inform the clinical treatment of various types of retinal diseases. Around 25 million people worldwide suffer from retinal degeneration, such as retinitis pigmentosa or macular degeneration, which damages the photoreceptors and progressively leads to blindness. A promising solution for regaining sight are the retinal prosthesis, which mimics the activity prior to degeneration by stimulating the remaining cells in the retina. Knowing more precisely how the retina encodes the external stimuli allows for devices that can help blind people regain all aspects of vision.

Finally, we would like to take a step back and consider a holistic point of view: why would an animal have an early visual system implementing this kind of computation? The choice of prioritising surprising stimuli could be of behavioural relevance: if the salamander spots a dark spot approaching from the skies, it might be a good time to run and hide. More generally, in a dynamical environment where both the organism and their external world might change at any point, focusing on the relevant information - which computations in sensory systems allow for - might be crucial for survival.

Appendix A

Appendix

A.1 Repetitions on the mouse retina

The experiments from Chapter 4 were reproduced in the retina of the mouse (Figure A.1). Due to a small number of cells showing OSR in the recording we had (18), these results can be treated as preliminary only. The recordings were performed by Berat Semihcan Sermet.

A.2 Details on the stimulus design

The initial questions we wanted to ask was, (i) whether retina could adapt to changing the temporal statistics of flash sequences, and (ii) if this could be explained by a normative theory such as efficient or predictive coding. For stimulus, we used sequences of randomly sampled number of dark flashes, presented interleaved with silences. Inspired by work from [Hosoya et al., 2005], we wanted to manipulate the underlying temporal statistics in order to have three environments (A, B and C) with same mean number of flashes, but different surprise trends. Here we define stimulus surprise s as in [Shannon, 1948]:

$$s = -\log(p(x_t = 0|x_{t-1} = 1)$$

where $p(x_t|x_{t-1})$ is the probability of transition, in this case, from dark flash (stim. $= 1$) to silence (stim. $= 0$). In each of the environments, stimulus surprise changes

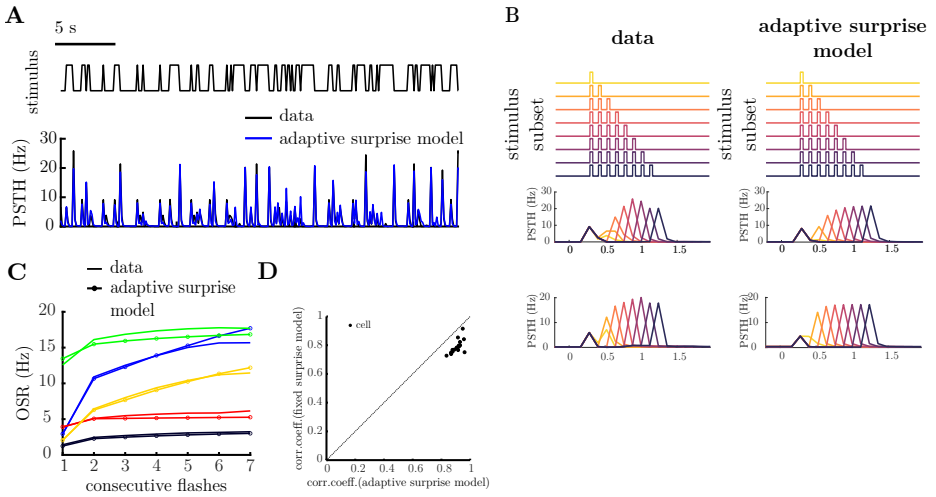

Fig. A.1 ***Repetition on mouse retina.*** ***A.*** *Stimulus excerpt (top, black) and corresponding response peri-stimulus time histogram (PSTH, bottom, black) compared to predicted firing rate for the adaptive surprise model (bottom, blue).* ***B.*** *Flash sequences of different length (top); PSTH of the responses (middle) and model prediction (bottom) when presented with a variable number of consecutive flashes. Each colour corresponds to a different length of the flash sequence. The fixed surprise model predicts same strength of OSR independent of the number of flashes, which is not what can be seen in the data.* ***C.*** *Average responses to an increasing number of flashes is not reproduced by the fixed surprise model, which can take into account only the previous state (flash or no flash). The flash history beyond that does not play a role in estimation of expectation. Different colours stand for individual cells.* ***D.*** *Comparison of Pearson correlation coefficient for fixed and adaptive surprise model. The latter outperforms the fixed one.*

differently as the number of consecutive flashes increase (Figure A.2A). While for B (red line) the length of the flash sequence does not play a role, A and C show either increase (blue) or decrease (green). The duration of each full-field flash (120 ms) and period between flashes (80 ms) is kept constant; each environment is presented for 20 minutes with two repetitions (ABC-ABC). The difference between these three behaviours is quite subtle, but if the retina adapts to stimulus

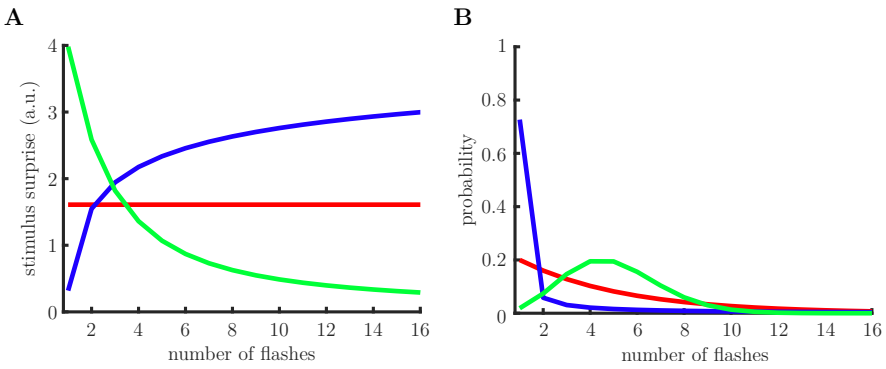

Fig. A.2 *Target surprise in function of the number of consecutive flashes. Three stimulus environments follow different surprise statistics (one colour for each environment).*

statistics, it should be possible to distinguish between them.

Each of these environment corresponds to flashes being drawn from a different probability distribution (Figure A.2B). These distributions are derived from the negative binomial distribution with a fixed mean number of flashes, equal to 7. In case of environment A (heavy-tailed distribution), the retina would be presented with either one flash only, or very long flash sequences. Flash sequences drawn from B (clustered distribution) are centered around the average number of flashes. Third distribution, C, is picked to reproduce a Markovian stimulus, where at each time point the current state is determined only based on the previous one. Therefore, the surprise is constant with the number of flashes, since nothing beyond the previous state (flash or no flash) influences the present state.

After confirming the presence of OSR (more in Results section of Chapter 4), we asked whether the RGCs adapted to the presented stimulus statistics. We compute the population mean for each environment across different numbers of flashes to check whether it matches the stimulus trends in Figure A.2A. For each of N flashes, we compute the average response to the stimulus sequence 11...10 of length $N + 1$. The minimal number of repetitions of a stimulus sequence is 50. The first thing to note is that for each of the three environments, the response gets stronger with the number of flashes (as previously seen in Figure 4.3C-D).

Except for the environment A (heavy-tailed distribution), this is not expected behaviour if the retina adapts to the underlying temporal statistics.

To observe more closely whether this is a systematic difference, we analyze the distribution of average responses (i) across flashes, and (ii) across repetitions Figure A.3. First thing to note is how the difference between two repetitions of the same environment is more pronounced than difference between two repetitions. Secondly, all distributions have a number of high firing rate responses that could be described as outliers, which might be inflating the difference between distributions on the population level despite being present only in some cells.

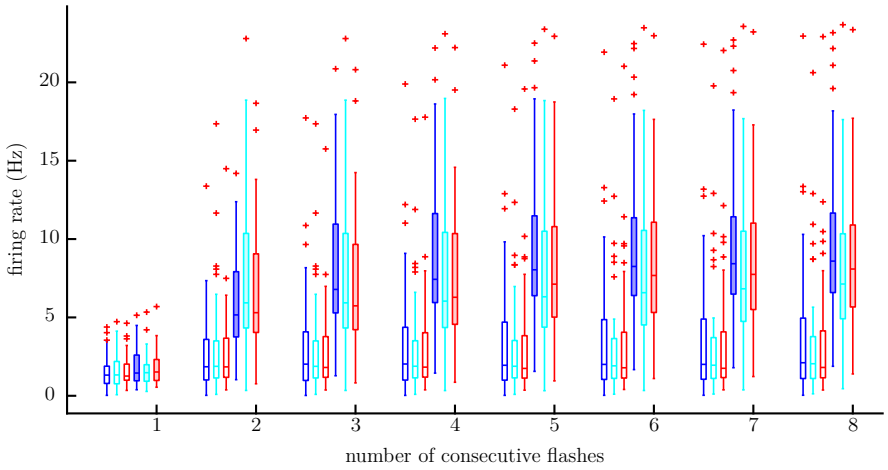

Fig. A.3 *Comparison of first (transparent) and second (shaded) repetitions across different length of flash sequences. All three environment are shown in chronological order (A - blue, B - cyan, C - red). The difference between two repetitions of the same distribution is higher than inter-environment differences.*

The ability of the retina to learn and adapt to presented stimulus statistics is well-documented [Gollisch and Meister, 2010]. Apart from variations in relatively simple features like the mean light level [Shapley and Enroth-Cugell, 1984] and contrast [Baccus and Meister, 2002], the RGCs also adapted to some less obvious, more abstract stimulus features, such as positive and negative temporal correlations, as well as spatial correlations [Hosoya et al., 2005]. On the

other hand, so far only a slight adaptation was found to occur when the 3rd and 4th order correlations (stimulus skewness and kurtosis) is manipulated [Tkačik et al., 2014]. Here we show a stimulus with three different temporal statistics, for which we did not find any conclusive evidence of adaptation to different stimulus environments.

One possible explanation for not finding any differences might be the complexity of the stimulus feature which is changing. The assumption we made was that the retina can learn beyond simple mean number of consecutive flashes and 'memorize' at least several flash sequences to successfully infer the the underlying distribution. This would require information retention on level of at least several seconds, which is more likely to be present in later visual processing areas than in the retina [Wark et al., 2007]. Possible future work could explore the responses of LGN and visual cortex to similar stimulus in order to test this reasoning.

There are several other ideas how to better probe whether the retina adapts to the stimulus surprise. One possible approach would be to switch between two environments with different transition probabilities, and record whether the receptive field changes (similar to [Hosoya et al., 2005]). Even more basic test would be if the flash sequences were kept constant in one environment, with different number of flashes in each.

Bibliography

[Allen et al., 2017] Allen, A. E., Storchi, R., Martial, F. P., Bedford, R. A., and Lucas, R. J. (2017). Melanopsin contributions to the representation of images in the early visual system. *Current Biology*, 27(11):1623–1632. 61

[Aranda and Schmidt, 2020] Aranda, M. L. and Schmidt, T. M. (2020). Diversity of intrinsically photosensitive retinal ganglion cells: circuits and functions. *Cellular and Molecular Life Sciences*, pages 1–19. 61

[Atick, 1992] Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251. 23

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3):183. 1, 21, 22

[Babadi et al., 2010] Babadi, B., Casti, A., Xiao, Y., Kaplan, E., and Paninski, L. (2010). A generalized linear model of the impact of direct and indirect inputs to the lateral geniculate nucleus. *Journal of Vision*, 10(10):22–22. 17

[Baccus and Meister, 2002] Baccus, S. A. and Meister, M. (2002). Fast and slow contrast adaptation in retinal circuitry. *Neuron*, 36(5):909–919. 61, 66

[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. 8

[Baden et al., 2020] Baden, T., Euler, T., and Berens, P. (2020). Understanding the retinal basis of vision across species. *Nature Reviews Neuroscience*, 21(1):5–20. 6, 14

[Baldi, 2002] Baldi, P. (2002). A computational theory of surprise. In *Information, Coding and Mathematics*, pages 1–25. Springer. 36

[Barlow, 2001] Barlow, H. (2001). Redundancy reduction revisited. *Network:*

computation in neural systems, 12(3):241. 22, 25, 26

[Barlow et al., 1964] Barlow, H., Hill, R., and Levick, W. (1964). Retinal ganglion cells responding selectively to direction and speed of image motion in the rabbit. *The Journal of physiology*, 173(3):377–407. 9

[Barlow and Levick, 1969] Barlow, H. and Levick, W. (1969). Changes in the maintained discharge with adaptation level in the cat retina. *The Journal of Physiology*, 202(3):699–718. 50

[Barlow et al., 1961] Barlow, H. B. et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1:217–234. 1, 22

[Berkes and Wiskott, 2005] Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of vision*, 5(6):9–9. 28

[Berry et al., 1999] Berry, M. J., Brivanlou, I. H., Jordan, T. A., and Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature*, 398(6725):334–338. 10, 11, 17, 30

[Bialek et al., 2006] Bialek, W., Van Steveninck, R. R. D. R., and Tishby, N. (2006). Efficient representation as a design principle for neural coding and computation. In *2006 IEEE international symposium on information theory*, pages 659–663. IEEE. 26

[Borst and Euler, 2011] Borst, A. and Euler, T. (2011). Seeing things in motion: models, circuits, and mechanisms. *Neuron*, 71(6):974–994. 50

[Bullock et al., 1990a] Bullock, T. H., Hofmann, M. H., Nahm, F. K., New, J. G., and Prechtl, J. C. (1990a). Event-related potentials in the retina and optic tectum of fish. *Journal of neurophysiology*, 64(3):903–914. 12, 13

[Bullock et al., 1990b] Bullock, T. H., Hofmann, M. H., New, J. G., and Nahm, F. K. (1990b). Dynamic properties of visual evoked potentials in the tectum of cartilaginous and bony fishes, with neuroethological implications. *Journal of*

Experimental Zoology, 256(S5):142–155. 13

[Bullock et al., 1994] Bullock, T. H., Karamürsel, S., Achimowicz, J. Z., McClune, M. C., and Başar-Eroglu, C. (1994). Dynamic properties of human visual evoked and omitted stimulus potentials. *Electroencephalography and clinical neurophysiology*, 91(1):42–53. 13

[Bullock et al., 1993] Bullock, T. H., Karamürsel, S., and Hofmann, M. H. (1993). Interval-specific event related potentials to omitted stimuli in the electrosensory pathway in elasmobranchs: an elementary form of expectation. *Journal of Comparative Physiology A*, 172(4):501–510. 13

[Cadena et al., 2019] Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4):e1006897. 18

[Chalk et al., 2018] Chalk, M., Marre, O., and Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191. 28, 29

[Chen et al., 2014] Chen, E. Y., Chou, J., Park, J., Schwartz, G., and Berry, M. J. (2014). The neural circuit mechanisms underlying the retinal response to motion reversal. *Journal of Neuroscience*, 34(47):15557–15575. 13

[Chen et al., 2013] Chen, E. Y., Marre, O., Fisher, C., Schwartz, G., Levy, J., da Silveira, R. A., and Berry, M. J. (2013). Alert response to motion onset in the retina. *Journal of Neuroscience*, 33(1):120–132. 12

[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. 32

[Chichilnisky, 2001] Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: computation in neural systems*, 12(2):199. 17

[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. 50, 51

[Creutzig and Sprekeler, 2008] Creutzig, F. and Sprekeler, H. (2008). Predictive coding and the slowness principle: An information-theoretic approach. *Neural Computation*, 20(4):1026–1041. 28

[da Silveira and Roska, 2011] da Silveira, R. A. and Roska, B. (2011). Cell types, circuits, computation. *Current opinion in neurobiology*, 21(5):664–671. 8

[Dacey, 2004a] Dacey, D. (2004a). 20 origins of perception: Retinal ganglion cell diversity and the creation of parallel visual pathways. 9

[Dacey, 2004b] Dacey, D. (2004b). Origins of perception: retinal ganglion cell diversity and the creation of parallel visual pathways. *The cognitive neurosciences*, 3:281–301. 8

[Dayan and Abbott, 2001] Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series. 16

[De Valois and De Valois, 1991] De Valois, R. L. and De Valois, K. K. (1991). Vernier acuity with stationary moving gabors. *Vision research*, 31(9):1619–1626. 10

[Demb, 2007] Demb, J. B. (2007). Cellular mechanisms for direction selectivity in the retina. *Neuron*, 55(2):179–186. 9

[Demb, 2008] Demb, J. B. (2008). Functional circuitry of visual adaptation in the retina. *The Journal of physiology*, 586(18):4377–4384. 50

[Deny, 2016] Deny, S. (2016). *Local and non-local processing in the retina*. PhD thesis, Paris 6. 17, 18, 50

[Deny et al., 2017] Deny, S., Ferrari, U., Mace, E., Yger, P., Caplette, R., Picaud, S., Tkačik, G., and Marre, O. (2017). Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1):1–17. 9

[Deshmukh, 2015] Deshmukh, N. R. (2015). *Complex computation in the retina*. PhD thesis, Princeton University. 12, 15

[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. 25

[Doi and Lewicki, 2007] Doi, E. and Lewicki, M. S. (2007). A theory of retinal population coding. *Advances in neural information processing systems*, 19:353. 24, 25

[Doi and Lewicki, 2014] Doi, E. and Lewicki, M. S. (2014). A simple model of optimal population coding for sensory systems. *PLoS computational biology*, 10(8):e1003761. 25

[Doya et al., 2007] Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press. 57

[Enroth-Cugell and Robson, 1966] Enroth-Cugell, C. and Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of physiology*, 187(3):517–552. 9

[Ferrari et al., 2018] Ferrari, U., Deny, S., Chalk, M., Tkačik, G., Marre, O., and Mora, T. (2018). Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons. *Physical Review E*, 98(4):042410. 52

[Field, 1987] Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394. 23

[Foster et al., 1991] Foster, R. G., Provencio, I., Hudson, D., Fiske, S., De Grip, W., and Menaker, M. (1991). Circadian photoreception in the retinally degenerate mouse (rd/rd). *Journal of Comparative Physiology A*, 169(1):39–50. 6

[Fradkin, 2020] Fradkin, S. I. (2020). *Predictive Processing in the Retina Through Evaluation of the Omitted-Stimulus Response*. PhD thesis, Rutgers The State

University of New Jersey, School of Graduate Studies. 13

[Gao et al., 2009] Gao, J., Schwartz, G., Berry, M. J., and Holmes, P. (2009). An oscillatory circuit underlying the detection of disruptions in temporally-periodic patterns. *Network: Computation in Neural Systems*, 20(2):106–135. 2, 15

[Geisler, 1989] Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological review*, 96(2):267. 50

[Geisler, 2003] Geisler, W. S. (2003). Ideal observer analysis. *The visual neurosciences*, 10(7):12–12. 50

[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. 30, 60

[Gjorgjieva et al., 2014] Gjorgjieva, J., Sompolinsky, H., and Meister, M. (2014). Benefits of pathway splitting in sensory coding. *Journal of Neuroscience*, 34(36):12127–12144. 25

[Goldin et al., 2021] Goldin, M. A., Lefebvre, B., Virgili, S., Ecker, A., Mora, T., Ferrari, U., and Marre, O. (2021). Context-dependent selectivity to natural scenes in the retina. *bioRxiv*. 9, 18

[Goldstein et al., 2002] Goldstein, A., Spencer, K. M., and Donchin, E. (2002). The influence of stimulus deviance and novelty on the P300 and novelty P3. *Psychophysiology*, 39(6):781–790. 60

[Gollisch and Meister, 2008] Gollisch, T. and Meister, M. (2008). Modeling convergent on and off pathways in the early visual system. *Biological cybernetics*, 99(4):263–278. 17

[Gollisch and Meister, 2010] Gollisch, T. and Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164. 9, 10, 50, 66

[Grünert and Martin, 2020] Grünert, U. and Martin, P. R. (2020). Cell types and

cell circuits in human and non-human primate retina. *Progress in Retinal and Eye Research*, page 100844. 8

[Hartline, 1938] Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415. 7

[Hartline, 1940] Hartline, H. K. (1940). The receptive fields of optic nerve fibers. *American Journal of Physiology-Legacy Content*, 130(4):690–699. 7

[Hattar et al., 2002] Hattar, S., Liao, H.-W., Takao, M., Berson, D. M., and Yau, K.-W. (2002). Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science*, 295(5557):1065–1070. 6

[Heitman et al., 2016] Heitman, A., Brackbill, N., Greschner, M., Sher, A., Litke, A. M., and Chichilnisky, E. (2016). Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv*, page 045336. 17, 18

[Hosoya et al., 2005] Hosoya, T., Baccus, S. A., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77. 2, 31, 51, 63, 66, 67

[Kaplan and Shapley, 1986] Kaplan, E. and Shapley, R. M. (1986). The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proceedings of the National Academy of Sciences*, 83(8):2755–2757. 9

[Karklin and Simoncelli, 2011] Karklin, Y. and Simoncelli, E. P. (2011). Efficient coding of natural images with a population of noisy linear-nonlinear neurons. *Advances in neural information processing systems*, 24:999. 25

[Kastner and Baccus, 2011] Kastner, D. B. and Baccus, S. A. (2011). Coordinated dynamic encoding in the retina using opposing forms of plasticity. *Nature neuroscience*, 14(10):1317. 8, 50

[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. 52

[Keller and Mrsic-Flogel, 2018] Keller, G. B. and Mrsic-Flogel, T. D. (2018). Pre-

dictive processing: a canonical cortical computation. *Neuron*, 100(2):424–435. 59

[Kersten, 1987] Kersten, D. (1987). Predictability and redundancy of natural images. *JOSA A*, 4(12):2395–2400. 21

[Kim and Rieke, 2001] Kim, K. J. and Rieke, F. (2001). Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. *Journal of Neuroscience*, 21(1):287–299. 1, 2

[Kotekal and MacLean, 2020] Kotekal, S. and MacLean, J. N. (2020). Recurrent interactions can explain the variance in single trial responses. *PLoS computational biology*, 16(1):e1007591. 17

[Kremláček et al., 2016] Kremláček, J., Kreegipuu, K., Tales, A., Astikainen, P., Poldver, N., Näätänen, R., and Stefanics, G. (2016). Visual mismatch negativity (vMMN): A review and meta-analysis of studies in psychiatric and neurological disorders. *Cortex*, 80:76–112. 59

[Kuffler, 1953] Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68. 7

[Kühn and Gollisch, 2016] Kühn, N. K. and Gollisch, T. (2016). Joint encoding of object motion and motion direction in the salamander retina. *Journal of Neuroscience*, 36(48):12203–12216. 8, 9

[Latimer et al., 2014] Latimer, K. W., Chichilnisky, E., Rieke, F., and Pillow, J. W. (2014). Inferring synaptic conductances from spike trains with a biophysically inspired point process model. *Advances in neural information processing systems*, 27:954–962. 17

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 18

[Levick, 1967] Levick, W. (1967). Receptive fields and trigger features of ganglion cells in the visual streak of the rabbit's retina. *The Journal of physiology*,

188(3):285–307. 7

[Lindsay, 2021] Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031. 18

[Liu et al., 2009] Liu, Y. S., Stevens, C. F., and Sharpee, T. O. (2009). Predictable irregularities in retinal receptive fields. *Proceedings of the National Academy of Sciences*, 106(38):16499–16504. 7

[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. 26

[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. 18, 19

[Mahuas et al., 2020] Mahuas, G., Isacchini, G., Marre, O., Ferrari, U., and Mora, T. (2020). A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons. *arXiv preprint arXiv:2006.06497*. 17

[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. 8, 52

[Masland, 2012] Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, 76(2):266–280. 7

[McAnany and Alexander, 2009] McAnany, J. J. and Alexander, K. R. (2009). Is there an omitted stimulus response in the human cone flicker electroretinogram? *Visual neuroscience*, 26(2):189–194. 13

[McIntosh et al., 2016] McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, 29:1369–1377. 15,

17, 18, 19

[Meister and Berry, 1999] Meister, M. and Berry, M. J. (1999). The neural code of the retina. *Neuron*, 22(3):435–450. 8

[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. 51, 56, 57

[Milosavljevic et al., 2018] Milosavljevic, N., Storchi, R., Eleftheriou, C. G., Colins, A., Petersen, R. S., and Lucas, R. J. (2018). Photoreceptive retinal ganglion cells control the information rate of the optic nerve. *Proceedings of the National Academy of Sciences*, 115(50):E11817–E11826. 17

[Młynarski et al., 2021] Młynarski, W., Hledík, M., Sokolowski, T. R., and Tkačik, G. (2021). Statistical analysis and optimality of neural systems. *Neuron*, 109(7):1227–1241. 28

[Morgan and Kamp, 1980] Morgan, W. W. and Kamp, C. W. (1980). Dopaminergic amacrine neurons of rat retinas with photoreceptor degeneration continue to respond to light. *Life sciences*, 26(19):1619–1626. 6

[Münch et al., 2009] Münch, T. A., Da Silveira, R. A., Siegert, S., Viney, T. J., Awatramani, G. B., and Roska, B. (2009). Approach sensitivity in the retina processed by a multifunctional neural circuit. *Nature neuroscience*, 12(10):1308–1316. 9, 50

[Näätänen et al., 1978] Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, 42(4):313–329. 59

[Näätänen and Kreegipuu, 2012] Näätänen, R. and Kreegipuu, K. (2012). The mismatch negativity (MMN). 59

[Näätänen et al., 2007] Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118(12):2544–2590. 59

[Nijhawan, 1994] Nijhawan, R. (1994). Motion extrapolation in catching. *Nature*. 10

[Nijhawan, 2002] Nijhawan, R. (2002). Neural delays, visual motion and the flash-lag effect. *Trends in cognitive sciences*, 6(9):387–393. 10

[Nirenberg and Latham, 1998] Nirenberg, S. and Latham, P. E. (1998). Population coding in the retina. *Current Opinion in Neurobiology*, 8(4):488–493. 52

[Ocko et al., 2018] Ocko, S., Lindsey, J., Ganguli, S., and Deny, S. (2018). The emergence of multiple retinal cell types through efficient coding of natural movies. In *Advances in Neural Information Processing Systems*, pages 9389–9400. 25

[Ölveczky et al., 2003] Ölveczky, B. P., Baccus, S. A., and Meister, M. (2003). Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408. 9, 50

[Oyster, 1968] Oyster, C. (1968). The analysis of image motion by the rabbit retina. *The Journal of Physiology*, 199(3):613–635. 9

[Palmer et al., 2015] Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913. 27, 28

[Paninski, 2004] Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243. 17

[Park et al., 2014] Park, I. M., Meister, M. L., Huk, A. C., and Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature neuroscience*, 17(10):1395–1403. 17

[Pazo-Alvarez et al., 2003] Pazo-Alvarez, P., Cadaveira, F., and Amenedo, E. (2003). MMN in the visual modality: a review. *Biological psychology*, 63(3):199–236. 59

[Petrusca et al., 2007] Petrusca, D., Grivich, M. I., Sher, A., Field, G. D., Gau-

thier, J. L., Greschner, M., Shlens, J., Chichilnisky, E., and Litke, A. M. (2007). Identification and characterization of a Y-like primate retinal ganglion cell type. *Journal of Neuroscience*, 27(41):11019–11027. 9

[Pillow, 2007] Pillow, J. (2007). Likelihood-based approaches to modeling the neural code. *Bayesian brain: Probabilistic approaches to neural coding*, 70:53–70. 16, 17

[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. 16, 42

[Pillow et al., 2008] Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999. 17, 52

[Prechtl and Bullock, 1994] Prechtl, J. C. and Bullock, T. H. (1994). Event-related potentials to omitted visual stimuli in a reptile. *Electroencephalography and Clinical Neurophysiology*, 91(1):54–66. 13

[Ramón et al., 2001] Ramón, F., Hernández, O. H., and Bullock, T. H. (2001). Event-related potentials in an invertebrate: crayfish emit 'omitted stimulus potentials'. *Journal of experimental biology*, 204(24):4291–4300. 13

[Ratliff et al., 2010] Ratliff, C. P., Borghuis, B. G., Kao, Y.-H., Sterling, P., and Balasubramanian, V. (2010). Retina is structured to process an excess of darkness in natural scenes. *Proceedings of the National Academy of Sciences*, 107(40):17368–17373. 25

[Rieke et al., 1995] Rieke, F., Bodnar, D., and Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):259–265. 23

[Roy et al., 2021] Roy, S., Jun, N. Y., Davis, E. L., Pearson, J., and Field, G. D. (2021). Inter-mosaic coordination of retinal receptive fields. *Nature*, 592(7854):409–413. 52

[Rozenblit and Gollisch, 2020] Rozenblit, F. and Gollisch, T. (2020). What the salamander eye has been telling the vision scientist's brain. In *Seminars in cell & developmental biology*, volume 106, pages 61–71. Elsevier. 8

[Rubin et al., 2016] Rubin, J., Ulanovsky, N., Nelken, I., and Tishby, N. (2016). The representation of prediction error in auditory cortex. *PLoS computational biology*, 12(8):e1005058. 60

[Rust and Movshon, 2005] Rust, N. C. and Movshon, J. A. (2005). In praise of artifice. *Nature neuroscience*, 8(12):1647–1650. 51

[Rust and Palmer, 2021] Rust, N. C. and Palmer, S. E. (2021). Remembering the past to see the future. *Annual Review of Vision Science*, 7:349–365. 10

[Salisbury and Palmer, 2015] Salisbury, J. and Palmer, S. E. (2015). Optimal prediction and natural scene statistics in the retina. *arXiv preprint arXiv:1507.00125*. 26

[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. 12, 13, 14, 15, 31, 32, 33

[Schwartz et al., 2007a] Schwartz, G., Harris, R., Shrom, D., and Berry, M. J. (2007a). Detection and prediction of periodic patterns by the retina. *Nature neuroscience*, 10(5):552–554. 2, 12, 13, 14, 31, 33, 34, 52

[Schwartz et al., 2007b] Schwartz, G., Taylor, S., Fisher, C., Harris, R., and Berry II, M. J. (2007b). Synchronized firing among retinal ganglion cells signals motion reversal. *Neuron*, 55(6):958–969. 11, 30, 33, 61

[Sederberg et al., 2018] Sederberg, A. J., MacLean, J. N., and Palmer, S. E. (2018). Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proceedings of the National Academy of*

Sciences, 115(5):1105–1110. 27

[Segev et al., 2006] Segev, R., Puchalla, J., and Berry, M. J. (2006). Functional organization of ganglion cells in the salamander retina. *Journal of neurophysiology*, 95(4):2277–2292. 8

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423. 22, 63

[Shannon, 1951] Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal*, 30(1):50–64. 21

[Shapley and Enroth-Cugell, 1984] Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Progress in retinal research*, 3:263–346. 17, 50, 66

[Shapley and Victor, 1978] Shapley, R. M. and Victor, J. D. (1978). The effect of contrast on the transfer properties of cat retinal ganglion cells. *The Journal of physiology*, 285(1):275–298. 1, 2, 9

[Simoncelli and Olshausen, 2001] Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216. 1, 23, 24, 26

[Simoncelli et al., 2004] Simoncelli, E. P., Paninski, L., Pillow, J., Schwartz, O., et al. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, 3(327-338):1. 16

[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. 50

[Smirnakis et al., 1997] Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., and Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386(6620):69–73. 1, 9, 61

[Srinivasan et al., 1982] Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982).

Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459. 22

[Stefanics et al., 2014] Stefanics, G., Kremláček, J., and Czigler, I. (2014). Visual mismatch negativity: a predictive coding view. *Frontiers in human neuroscience*, 8:666. 59

[Subramaniyan et al., 2018] Subramaniyan, M., Ecker, A. S., Patel, S. S., Cotton, R. J., Bethge, M., Pitkow, X., Berens, P., and Tolias, A. S. (2018). Faster processing of moving compared with flashed bars in awake macaque v1 provides a neural correlate of the flash lag illusion. *Journal of neurophysiology*, 120(5):2430–2452. 10

[Sumbre et al., 2008] Sumbre, G., Muto, A., Baier, H., and Poo, M.-m. (2008). Entrained rhythmic activities of neuronal ensembles as perceptual memory of time interval. *Nature*, 456(7218):102–106. 13

[Switkes et al., 1978] Switkes, E., Mayer, M. J., and Sloan, J. A. (1978). Spatial frequency analysis of the visual environment: Anisotropy and the carpentered environment hypothesis. *Vision research*, 18(10):1393–1399. 23

[Tales et al., 1999] Tales, A., Newton, P., Troscianko, T., and Butler, S. (1999). Mismatch negativity in the visual modality. *Neuroreport*, 10(16):3363–3367. 59

[Tanaka et al., 2019] Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S. A., and Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *arXiv preprint arXiv:1912.06207*. 15, 18, 19

[Tessier-Lavigne, 2000] Tessier-Lavigne, M. (2000). Visual processing by the retina. *Principles of neural science*, pages 507–522. 5

[Tishby et al., 2000] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*. 27

[Tkačik et al., 2014] Tkačik, G., Ghosh, A., Schneidman, E., and Segev, R. (2014). Adaptation to changes in higher-order stimulus statistics in the salamander

retina. *PLoS one*, 9(1):e85841. 67

[Truccolo et al., 2005] Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089. 17

[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. 30, 60

[Vaney et al., 2012] Vaney, D. I., Sivyer, B., and Taylor, W. R. (2012). Direction selectivity in the retina: symmetry and asymmetry in structure and function. *Nature Reviews Neuroscience*, 13(3):194–208. 50

[Wark et al., 2007] Wark, B., Lundstrom, B. N., and Fairhall, A. (2007). Sensory adaptation. *Current opinion in neurobiology*, 17(4):423–429. 1, 67

[Wässle, 2004] Wässle, H. (2004). Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757. 9, 12

[Weidmann, 2009] Weidmann, M. D. (2009). *Exploring the Cellular Mechanism of the Omitted Stimulus Response*. PhD thesis, Princeton University. 14

[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. 2, 13, 15

[Wiskott and Sejnowski, 2002] Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770. 28

[Yamins et al., 2014] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624. 18

[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. 34, 52

[Young et al., 2021] Young, B. K., Ramakrishnan, C., Ganjawala, T., Wang, P., Deisseroth, K., and Tian, N. (2021). An uncommon neuronal class conveys visual signals from rods and cones to retinal ganglion cells. *Proceedings of the National Academy of Sciences*, 118(44). 6