Dynamical Synapses in the Retinal Network

Simone Ebert

Synapses dynamiques dans le réseau rétinien

Simone EBERT

équipe BIOVISION
Centre INRIA
d'Université Côte d'Azur

Présentée en vue de l'obtention du grade de docteur en Santé numérique **d'**Université Côte d'Azur

Dirigée par : Bruno Cessac, Directeur de recherche, Inria,Université Côte d'Azur

Soutenue le : 13/12/2023

Devant le jury, composé de :
Adrián Palacios, Full Professor, University of Valparaíso
Leon Lagnado, Full Professor, University of Sussex
Michael J. Berry, Professor, Princeton University
Katrin Franke, Assistant Professor, Baylor College of Medicine
Olivier Marre, Directeur de recherche, INSERM
Romain Veltz, Chargé de recherche, Inria,Université Côte d'Azur

NEURO
MOD

UNIVERSITE COTE D'AZUR

DOCTORAL THESIS

Dynamical Synapses in the Retinal Network

Author:
Simone EBERT

Supervisor:
Dr. Bruno CESSAC

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Biovision Lab
INRIA Research Center Sophia-Antipolis
Ecole Doctorale des Sciences et Technologies d'Iinformation et Communication

October 13, 2023

Synapses dynamiques dans le réseau rétinien

Jury:

Rapporteurs
Adrián Palacios, Full Professor, University of Valparaíso
Leon Lagnado, Full Professor, University of Sussex
Michael J. Berry, Professor, Princeton University

Examinateurs
Katrin Franke, Assistant Professor, Baylor College of Medicine
Olivier Marre, Directeur de recherche, INSERM
Romain Veltz, Chargé de recherche, Inria, Université Côte d'Azur
Bruno Cessac, Directeur de recherche, Inria, Université Côte d'Azur

Directeur de thèse
Bruno Cessac, Directeur de recherche, Inria, Université Côte d'Azur

Declaration of Authorship

I declare that this thesis titled, "Dynamical Synapses in the Retinal Network" and the work presented in it are my own. Where I have consulted the published work of others, this is always clearly attributed.

Simone EBERT
October 13, 2023

"There are these two young fish swimming along and they happen to meet an older fish swimming the other way, who nods at them and says - Morning boys. How's the water? - And the two young fish swim on for a bit, and then eventually one of them looks over at the other and goes - What the hell is water? "

David Foster Wallace

Abstract

The retina is the first stage of visual processing, filtering relevant visual information before sending it to the brain. Thereby, neural transmission induces a delay with which a signal reaches the brain, yet our visual system can efficiently detect changes in real time. To this end, a long-standing hypothesis is that retinal ganglion cells pursue a predictive coding strategy - they do not signal the visual scene per se, but rather form predictions about future inputs and signal only surprising events, e.g. mismatches between observation and predictions formed by previous inputs. In this thesis, we focused on two examples of predictive coding in the retina and propose mechanistic implementations of prediction formation and surprise signalling.

An example of purely temporal predictive coding in the retina is the Omitted Stimulus Response (OSR): when a regular sequence of flashes suddenly ends, the retina emits a large response signaling the missing stimulus. The latency of this response scales with the period of the flash sequence, thereby signaling the omitted flash with constant latency. This indicates a temporally precise expectation of when the next flash should have occurred. Here, we show that depressing inhibitory synapses may aid the retina to have a temporal expectation of an omitted stimulus in a flash sequence. First, we experimentally show that the latency scaling between stimulus frequency and response was lost when glycinergic amacrine cells (inhibitory interneurons) are pharmacologically inhibited in the mouse retina. We then present a simple circuit model and showed that depressing inhibitory synapses were a necessary component to reproduce our experimental findings. A new prediction of our model is that the accuracy of the constant latency requires enough flashes in the stimulus, which we could confirm experimentally.

The retina is also known to form predictions on a spatiotemporal scale, namely it anticipates the trajectory of a moving object and signals surprise when movement suddenly starts, changes direction or the object disappears. A fast decrease in gain was proposed to account for these responses. However, the underlying mechanisms behind gain control are not fully understood. We found that lateral inhibitory connectivity in the retinal network can enable retinal ganglion cells to anticipate the trajectory of a moving object, providing a possible network implementation for gain control. We then show how short-term plasticity can aid the retina to tune its anticipatory response timing to the speed of a moving object presented.

With this work, we propose how inhibitory interneurons can play a key role in prediction formation and error signaling to temporal and spatiotemporal patterns. We show that depressing synapses could be a key component for the temporal tuning of predictive and error responses in the retina, a concept which may generalize to other brain areas.

Keywords: retina, dynamical synapses, predictive coding, neuronal circuits

Résumé

La rétine est la première étape du traitement visuel, traitant les informations visuelles pertinentes avant de les envoyer au cerveau. La transmission neuronale induit un délai avant qu'un signal n'atteigne le cerveau, alors que notre système visuel est capable de détecter efficacement les changements en temps réel. À cette fin, une hypothèse importante est que les cellules ganglionnaires de la rétine utilisent une stratégie de codage prédictif - elles ne signalent pas la scène visuelle en soi, mais forment plutôt des prédictions sur les entrées futures et ne signalent que les événements surprenants, par exemple les discordances entre l'observation et les prédictions formées par les entrées précédentes.

Dans cette thèse, nous nous concentrons sur deux exemples de codage prédictif dans la rétine et proposons des implémentations mécanistiques de la formation de prédictions et de la signalisation de la surprise.

Nous étudions d'abord la formation de prédictions temporelles et la signalisation de la surprise dans la rétine avec l'exemple de la réponse au stimulus omis (OSR): lorsqu'une séquence régulière de flash lumineux s'interrompt soudainement, la rétine émet une réponse de forte amplitude, signalant le stimulus manquant, et la latence de cette réponse s'échelonne avec la période de la séquence de flash, signalant ainsi le flash omis avec une latence adaptée. Nous montrons que la dépression des synapses inhibitrices aide la rétine à avoir une attente temporelle d'un stimulus omis dans une séquence de flash. Tout d'abord, nous montrons expérimentalement que l'échelonnement de la latence entre la fréquence du stimulus et la réponse est perdu lorsque les cellules amacrines glycinergiques (interneurones inhibiteurs) sont inhibées pharmacologiquement dans la rétine de la souris. Avec un modèle de circuit simple nous montrons que la dépression des synapses inhibitrices ést un élément nécessaire pour reproduire nos résultats expérimentaux.

La rétine est également connue pour former des prédictions à l'échelle spatiotemporelle, c'est-à-dire qu'elle anticipe la trajectoire d'un objet en mouvement et signale sa surprise lorsque le mouvement commence soudainement, change de direction ou que l'objet disparaît. Une diminution rapide du gain (contrôle du gain) qui tronque le signal pour avancer le pic de réponse a été proposée pour expliquer ces réponses, mais les mécanismes sous-jacents du contrôle du gain ne sont pas connus. Nous avons découvert que la connectivité inhibitrice latérale dans le réseau rétinien peut permettre aux cellules ganglionnaires de la rétine d'anticiper la trajectoire d'un objet en mouvement, ce qui constitue une implémentation réseau possible pour le contrôle du gain. Nous montrons ensuite comment la plasticité à court terme peut aider la rétine à adapter sa réponse anticipative à la vitesse de l'objet présenté.

Avec ce travail, nous proposons comment les interneurones inhibiteurs peuvent jouer un rôle clé dans la formation des prédictions et la signalisation des erreurs dans les modèles temporels et spatiotemporels. Nous montrons que la dépression des synapses pourrait être un élément clé de l'ajustement temporel des réponses de prédiction et d'erreur dans la rétine, un concept qui pourrait être généralisé à d'autres zones du cerveau.

Mots-clés: rétine, synapses dynamiques, codage prédictif, circuits neuronaux

Acknowledgements

I would like to express my gratitude to everyone who contributed along the way of this journey. First of all, I would like to thank my supervisor Bruno Cessac for all his kindness, support and encouragement. For his endless patience in many very inspiring blackboard sessions and all the great discussions. A huge thanks to Olivier Marre as well, for all your advice and incredibly valuable feedback, and for hosting me in your team so kindly. This turned my PhD into a great interdisciplinary experience.

I want to thank the members of my thesis committee: Katrin Franke, Romain Veltz, Olivier Marre and especially the readers, Adrian Palacios, Michael J. Berry and Leon Lagnado for their insightful comments and constructive feedback.

Thanks to all members of the Biovision team: To my office mates Jerome and Erwan, to, Pierre, Clement, Franz, Johanna, Florent, Christos, Sebastian and Sebastian for making the atmosphere in the team so joyful, to Helen , and to Marie-Cecile for all your help. Also to all the colleagues at the Institut de la Vision, especially to Thomas! It was great to share this project with you. I would like to thank again Adrian the entire Ecovis team and especially to David, Daniela, Francisco and Jorge for the warm welcome in Valparaiso and great a collaboration. I would like to thank the Neuromod Institute, Ingrid, Patricia, Mathieu, Chloe and Alexandre, for the funding of this PhD and especially for all the opportunities you provided, for your enthusiasm and support.

Jenny, thank you for being my pillar in France, for all the sport and coffee breaks and amazing moments. To Alexandra, for countless weekend trips and evenings in Antibes that that mad it feel like home there. I am grateful to have met you all during this time, it would not have been the same without you. I would like to thank my family, especially my parents, for their endless support and all their visits. And finally to Silva, Valentina, Evi, Atessa, Alina, Jil, Lara, Paula, Angela und Rojin for their everpresent friendship.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	**A**macrine **C**ell
ACM	**A**daptive **C**ascade **M**odel
BC	**B**ipolar **C**ell
CMA-ES	**C**ovariance **M**atrix **A**daptation **E**volutionary **S**trategy
DSGC	**D**irection **S**elective **G**anglion **C**ell
GABA	**G**amma-**A**mino**B**utyric **A**cid
GC	**G**anglion **C**ell
HC	**H**orizontal **C**ell
INL	**I**nner (it) **N**uclear **L**ayer
IPL	**I**nner **P**lexiform **L**ayer
LN	**L**inear **N**onlinear Model
MEA	**M**ulti- **E**lectrode **A**rray
OMS	**O**bject **M**otion **S**ensitive Cell
ONL	**O**uter **N**uclear **L**ayer
OPL	**O**uter **P**lexiform **L**ayer
OSR	**O**mitted **S**timulus **R**esponse
RAM	**R**eciprocal **A**macrine **M**odel
RF	**R**eceptive **F**ield
RGC	**R**etinal **G**angion **C**ell
RRP	**R**eadily **R**eleasable Vesicle **P**ool
STA	**S**pike **T**riggered **A**verage
STD	**S**hort **T**erm **D**epression
STF	**S**hort **T**erm **F**acilitation
STP	**S**hort **T**erm **P**lasticity

Chapter 1

Introduction

The retina is the interface between our brain and the visual environment around us. As the first stage in the perception of visual information, it translates the sensory input into a neuronal code that is transmitted to the rest of the brain. It does not just record a visual scene like a camera but rapidly and efficiently extracts relevant features of the environment such that our brain can form a real-time representation of the world around us.

A very influential model of how sensory systems achieve such a fast and efficient encoding of information is *predictive coding*. This theory has been proposed as a unifying framework for the function of the brain - namely to maintain an internal model of the world based on context and past experience which continually predicts sensory input (Millidge, Seth, and Buckley, 2021). To use resources in a maximally efficient way, neurons minimally respond to sensory inputs that match the brains' internal predictions and transmit only what is not predictable, that is to say unexpected or surprising inputs.

The retina has been one of the first neuronal structures on which this idea has been applied (Barlow et al., 1961; Srinivasan, Laughlin, and Dubs, 1982; Rao and Ballard, 1999). Concretely, the predictive coding hypothesis suggests that the retina carries an expectation of statistics of the visual input and reduces redundancy in its output by removing predictable components of the environment. In this sense, the center–surround architecture with its spatially displaced excitatory and inhibitory inputs performs predictive coding in space, while temporally displaced excitation and inhibition create a temporal antagonism with which the retina removes stimulus correlations over time (Huang and Rao, 2011).

A visual scene rapidly evolves, is continually changing and may contain vastly differing stimulus statistics in different environments, for example during the day or at night. Important computations that enable the retina to adapt to stimulus statistics is short-term plasticity, where synaptic strengths are dynamically adjusted based on recent activity. It has been shown that dynamic adaption can mediate the dynamic predictive encoding of visual inputs in ever-changing visual environments. For example, retinal ganglion cells, the output neurons of the retina, adapt to predominant orientations in the stimulus, thereby reducing redundancy and responding more strongly to stimuli that do not fit into a previously observed pattern (Hosoya, Baccus, and Meister, 2005). Inhibition and synaptic plasticity thereby play a key role (Johnston et al., 2019), and have also been shown to mediate the prediction of nearby objects (Kastner and Baccus, 2013b).

There exist further experimental examples that strongly suggest the employment of a predictive coding strategy by the retina: When it is presented with a periodic stimulus of repeating flashes, some retinal ganglion cells, the output neurons of the retina, do not respond to the repetitive (and thereby predictable) stimulus, but only emit a large response when the flash train ends, signalling a missing stimulus which

was expected (Schwartz et al., 2007a). In addition, the latency of this "surprise signal" is adapted to the period between the previously observed flashes, indicating an internal prediction of the temporal structure of the stimulus. This response phenomenon has been termed the Omitted Stimulus Response (OSR). Even more remarkably, retinal ganglion cells are also able to predict the motion of an object in a visual scene (Berry et al., 1999), and signal surprise responses when the object starts moving, abruptly changes its trajectory or disappears (Chen et al., 2013; Chen et al., 2014; Ding et al., 2021). The mechanistic implementation of predicting complex spatial and temporal stimulus structures in the aforementioned examples in the retina are still not fully understood, and a possible explanation via short-term plasticity has to our knowledge not been deeper investigated so far.

In the course of thesis, we sought to gain further mechanistic insights into the neuronal implementation of predictive coding in the retina and the role of dynamical synapses in it. To investigate how the retina can dynamically encode unexpected inputs with temporal precision, we asked:

1. What are the potential mechanisms allowing the retinal network to form expectations and compute surprise ?
2. What is the role of dynamical synapses in surprise computation?

We looked at the two aforementioned examples of the OSR and motion anticipation and investigated how the retina can detect surprise and achieve temporal precision in complex spatio-temporal patterns via known concepts of retinal computations, namely antagonistic excitation-inhibition inputs and dynamic adaptation. We developed how dynamical adaptation in inhibitory synapses can contribute to predictive encoding by setting temporal expectations in internal representations of complex spatio-temporal patterns.

1.1 Thesis outline

Chapter 2 of this thesis reviews relevant background for the understanding of the rest of the thesis. It provides a brief overview of the anatomical and functional organization of the retina, then continues with a more detailed description of predictive coding ideas and short-term plasticity in the retina and finally outlines the computational methodology used in this study.
In Chapter 3, we study the underlying mechanism of the Omitted Stimulus Response in a jointly experimental and computational study that has been carried out in close collaboration with Olivier Marres' laboratory at the Vision Institute in Paris, France. Here, we experimentally show that specific inhibitory inputs are necessary to implement a temporal representation of a stimulus pattern and show with a computational model that dynamic adaptation of inhibition could be the underlying mechanism.
Chapter 4 tests the idea of temporal expectation formation via dynamical synapses in the context of a more complex, spatio-temporal stimulus pattern at the example of a moving bar. We show how a retinal network model with a spatial organization can adjust its prediction of the future position of the moving bar to the stimulus speed via dynamical synapses.
Finally, in Chapter 5 we give a a concluding overview of the results and insights from this thesis and discuss possible future directions to test the theoretical predictions developed in the previous chapters.

Chapter 2

Background

This chapter highlights the relevant background to follow the ideas elaborated in the chapters 3 and 4 of this thesis and relates them to state-of-the-art concepts. It starts with a description of the retinal architecture and its central processing motifs. It then presents the known functions and principles of short-term plasticity in the retina, followed by an overview of the predictive coding theory applied to the retina. It highlights the phenomena of retinal temporal pattern recognition and motion anticipation, which are studied in detail in the following two chapters. Finally, it provides an overview of the computational and mathematical tools used in this thesis.

Contents

2.1 The Retina - Structural and Functional Organization

The retina is a thin layer of neural tissue located at the back of the eyeball. It transforms the visual world around us into a neuronal code of spike trains that is understood by the rest of our brain. In contrast to a camera, it goes far beyond a simple capturing of a visual scene. At this very first stage of perception, it already performs a number of impressive tasks, from the basic transformation of sensory inputs to biochemical and electrical signals, over feature extraction (Silveira and Roska, 2011) to efficient encoding of information (Palmer et al., 2015).

2.1.1 Retinal Architecture

The retinal network consists of 5 broad cell types, organized in 3 cellular layers, interconnected in 2 synaptic layers (Masland, 2012a). The first cellular layer, the outer nuclear layer (ONL), is composed of photoreceptor cells (Rods and Cones) which translate a visual stimulus from the environment into an electrical signal. This signal is then transferred onto excitatory bipolar cells (BCs) in the second layer, the inner nuclear layer (INL), which in turn connect to retinal ganglion cells (RGCs) in the third layer, the ganglion cell layer. Importantly, the inner nuclear layer contains two major types of inhibitory inter-neurons which modulate the visual signal at two stages: Horizontal cells (HCs) modulate the visual signal in the fist synaptic layer, the outer plexiform layer (OPL). Amacrine cells (ACs) modulate signal transmission in the second layer, the inner plexiform layer (IPL), where they inhibit bipolar cell terminals and RGC dendrites (Demb and Singer, 2015).

Amacrine cells are known for their great diversity with at least 30 sub-types which exhibit various different connectivity patterns within the retinal circuitry (Vaney,

1990). While some relay feed-forward inhibition onto RGCs, others form feed-back or reciprocal connections to bipolar cells (Diamond, 2017) or inhibit yet other amacrine cells. Broadly, they can be divided in to major groups - "small-field“ amacrine cells locally modulate the visual signal and mainly release glycine as a neurotransmitter while "wide-field“ amacrine cells transmit signals over large lateral distances via GABAergic inhibition (Masland, 2012b). Further, many amacrine cells also release excitatory neurotransmitter or neuromodulators and form gap junctions (Franke et al., 2017).

The majority of retinal cells process and transmit signals in an analog manner, responding via graded changes in membrane potential (Baden et al., 2013), while ganglion cells transform the signal into a spike train that is further transported into the brain via the optic nerve.

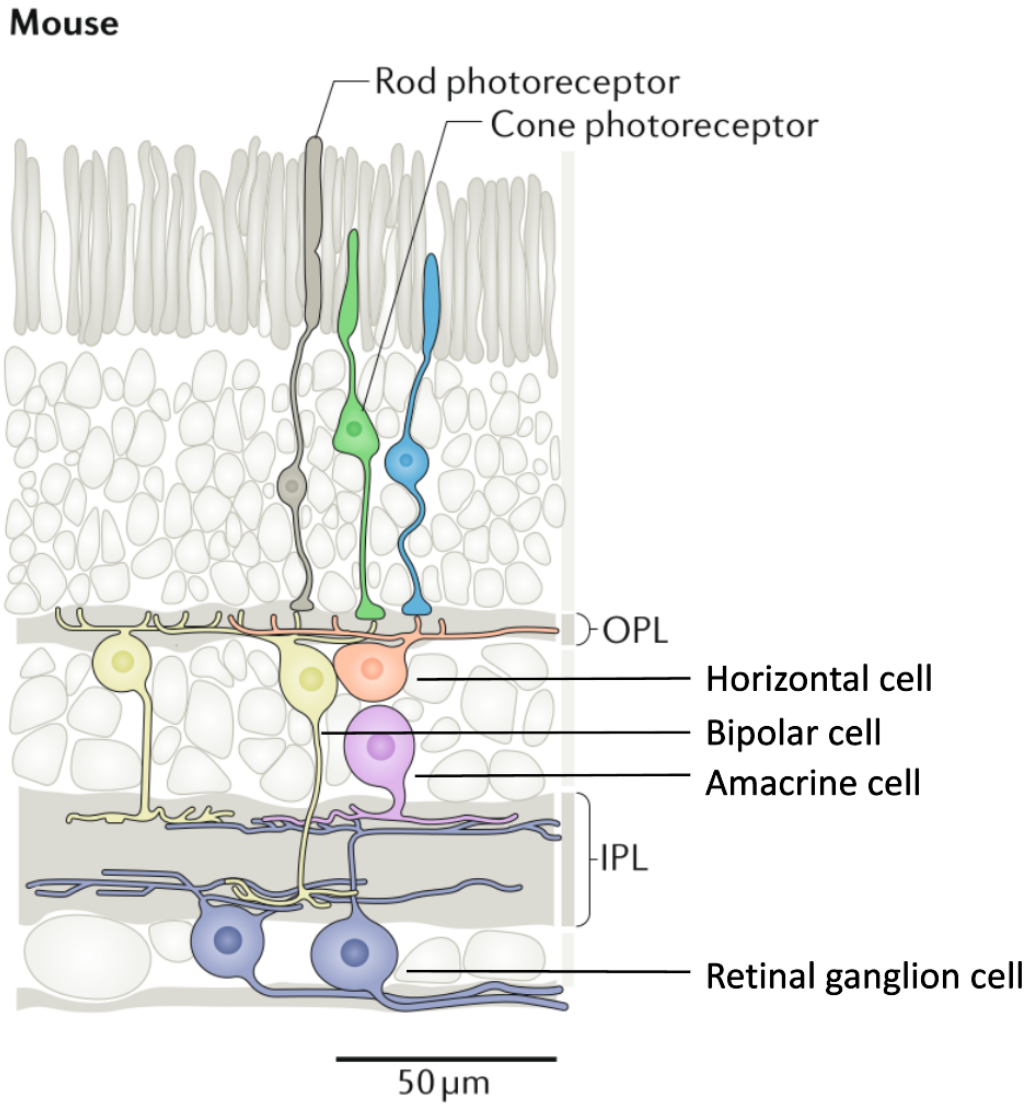

FIGURE 2.1: Celltypes and Connectvity in the Retinal Architecture.

The vertebrate retina comprises five classes of neurons that are organized into three nuclear and two synaptic layers. The first cellular layer consists of photoreceptor cells (rods and cones). In the first synaptic layer (outer plexiform layer = OPL), photoreceptors release glutamate onto the dendrites of bipolar cells (BCs) and horizontal cells (HCs), which comprise the second cellular layer. While horizontal cells connect laterally to provide feedback and feed-forward signals to both the first and second cellular layer, bipolar cells project to the second synaptic layer (inner plexiform layer = IPL) in the inner retina where they connect onto retinal ganglion cells (RGCs) and amacrine cells (ACs). Like horizontal cells, amacrine cells connect laterally and provide mostly inhibitory feed-back and feed-forward signals. Finally, RGCs integrate all synaptic inputs and send this information to the brain via the optic nerve. Image adapted from Baden, Euler, and Berens, 2020.

2.1.2 Receptive Fields

Retinal neurons are typically only responsive to stimuli in a restricted region of the visual space, their receptive field (Kuffler, 1953). In bipolar and ganglion cells, receptive fields are generally organized in an antagonistic center surround structure: a given cell depolarizes to a stimulus in its receptive field center but hyperpolarises in

response to the same stimulus in the surround region. In the center region, bipolar cells receive direct inputs from photoreceptors, while surround region is mediated via horizontal cells, suppressing photoreceptors and bipolar cells in the OPL (Demb and Singer, 2015). At the ganglion cell level, amacrine cell inhibition additionally shapes the receptive field structure. Wide-field amacrine cells with large dendritic trees are driven by bipolar cells in the surround region and shape the receptive field in a more complex and nonlinear manner, complementary to horizontal cells (Baccus et al., 2008; Zaghloul et al., 2007).

2.1.3 Feature Detection

The retina does not respond to an input image like a camera by reading out input intensity in a pixel-by-pixel manner, but already performs an extensive processing of the visual scene. This is generally thought to be implemented by decomposing the input into parallel channels, selective for distinct features of the stimulus (Baden et al., 2018). These parallel 'feature detectors' are formed by distinct types of retinal ganglion cells that are typically uniformly distributed, tiling the retinal surface in a "mosaic".

Their feature specificity is thought to arise from integrating inputs from distinct bipolar and amacrine cell types in within the IPL. Bipolar cells exhibit around 13-14 sub-types which differ in a number of properties (Euler et al., 2014), most important of which are response polarity and temporal processing. Bipolar cells can be broadly divided in ON and OFF type, which respond to the same visual signal in the opposite way: ON bipolar cells are depolarized by light increments whereas OFF bipolar cells are hyperpolarized and vice versa. This signal inversion originates in two different types of postsynaptic receptors in bipolars, one being activated (OFF) while the other being deactivated (ON) by glutamate release from photoreceptors (Dacheux and Miller, 1976). Two antagonistic information streams arise from this, signalling light increments and decrements in parallel .

Bipolar cells also differ vastly in their temporal kinetics: transient cells respond briefly to a stimulus which decays rapidly, while sustained cells respond for longer time and maintain a stable response rate throughout the stimulus. The underlying mechanisms that shape temporal kinetics are diverse, including intrinsic cell properties and regulation via amacrine cells (Awatramani and Slaughter, 2000; Dong and Werblin, 1998).

Additionally, lateral inhibition from over 40 different types of amacrine cells diversely impact the detection of more complex features. The center surround receptive field organization has early on been identified to render ganglion cells specifically responsive to edges, by supressing similar inputs in the center and surround (Hubel and Wiesel, 1962; Masland, 2012b). Orientation-selective ganglion cells selectively respond to certain orientations in a visual input, which has traditionally be thought to be implemented by suppression from other directions via amacrine inputs (Antinucci and Hindges, 2018) .

A visual scene *in vivo* is comprised of motion at many different levels, object motion, global movement of visual flow due to self motion and fast eye movements (saccades). It is thus not surprising that several retinal cell types have been identified which specifically detect motion features. Object motion sensitive (OMS) cells specifically detect the motion of objects in a moving environment (Baccus et al., 2008; Ölveczky, Baccus, and Meister, 2003) via amacrine suppression from the periphery. Thus cells only respond when the motion pattern in the receptive field center differs from the periphery. Another prominent example are direction selective ganglion

cells (DSGCs), which are selectively responsive to one direction of motion. One of the main causes of this selectivity are starburst amacrine cells, which suppress responses the opposite (null) direction (Wei, 2018) (for review see (Gollisch and Meister, 2010; Silveira and Roska, 2011)).

2.2 Short-term Plasticity in the Retina

Visual scenes can contain an enormous range of light intensities, from a few photons to billions, while saccades and self-motion introduce different levels of fluctuations into the visual input (Mostofi et al., 2020). Yet the visual system maintains sensitivity across all these different environmental conditions. To this end retinal neurons are known to continually adjust their sensitivity to the statistics of the visual input. Retinal neurons undergo luminance and contrast adaptation, thereby maintaining their dynamic response range via multiple complex mechanisms at the level of photoreceptors, bipolar cells and ganglion cells.

2.2.1 Luminance and Contrast Adaptation

To maintain sensitivity across a long range of luminescence, the retina adapts to the mean illumination level by shifting its dynamic response range to the mean input luminance (Dunn and Rieke, 2008). This process is generally thought to take place at the photo-receptor level (Shapley and Enroth-Cugell, 1984; Alexander, 2017), but intracelluar processing in bipolar cells seems to be involved as well (Jarsky et al., 2011; Singer and Diamond, 2006). At high light levels, photoreceptor adaptation dominates, whereas with decreasing light levels adaptation in bipolar cells overweights (Dunn, Lankheet, and Rieke, 2007).

Besides mean light intensity, the retinal circuit undergoes adaptation to contrast on a spatial and temporal scale. Mechanisms involved in contrast adaptation are thought to be located after photo-receptor release, within bipolar cell intracellular processing and synaptic release as well as within ganglion cells. Contrast adaptation probably arises from multiple mechanisms and adaptation sites and affects both sensitivity and kinetics of retinal ganglion cells on multiple timescales and differs between celltypes (Chander and Chichilnisky, 2001). Broadly, contrast adaptation can be divided into fast and slow contrast adaptation. Fast contrast adaptation is thought to be mediated by via response gain control within ganglion cells, while slow contrast adaptation arises from baseline potential decrements to high temporal contrasts (Baccus and Meister, 2002; Manookin and Demb, 2006).

A common site of adaptation to both luminance contrast is the ribbon synapse of bipolar cells (Oesch and Diamond, 2011). Instead of transmitting pulse-wise signals from arriving action potentials, ribbon synapses transmit 'analog signals', by continuously altering the rate of glutamate release based on intracellular calcium signals that in turn vary with membrane voltage. Additionally, ribbon synapses have a specialized organelle which holds synaptic vesicles close to the active zone where they can rapidly be released – the ribbon (Lagnado and Schmitz, 2015; Odermatt and Lagnado, 2009).

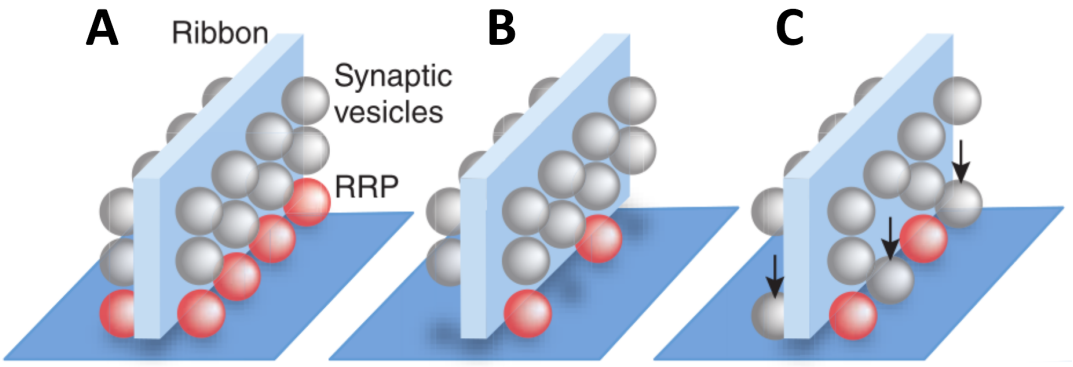

FIGURE 2.2: Vesicle dynamics at the synaptic ribbon

A The synaptic ribbon holds synaptic vesicles close to the active zone. Vesicles at the base of the ribbon, closest to the presynaptic membrane, constitute the readily releasable vesicle pool (RRP). **B** In response to membrane depolarization, vesicles from the RRP are released which leads to vesicle depletion in the RRP. **C** Vacated sites in the RRP are refilled from a reserve pool (RP), which is constituted by vesicles in the vicinity of the ribbon. The rates with which vesicles are released and replenished from the reserve pool determine the size of the RRP upon persistent depolarization. Image adapted from Diamond, 2011.

Vesicles in the synaptic terminal of the ribbon synapse can be divided in different pools (Figure 2.2). During synaptic activity, a readily releaseable vesicle pool (RRP) is located close to the active zone and can be partially or completely depleted upon activation on a very fast time scale. The RRP is then replenished by a reserve pool (RP) that is possibly formed by vesicles in the vicinity of the ribbon, which reloads the RRP on a second, slower time scale. The RP itself may be refilled from a large reservoir of vesicles in the synaptic terminal, which adds a third, very slow time scale to vesicle dynamics (Gersdorff and Matthews, 1997). During sustained stimulation the RRP gets completely depleted, the responses to consecutive stimuli are strongly depressed (Singer and Diamond, 2006).

The dynamics at the ribbon synapse contribute to both luminance and contrast encoding. Vesicle release at the ribbon synapse can be divided in two phases: A transient component, reflecting the rapid, partial depletion of the RRP due to a light step, whose discharge size is proportional to the contrast, and a sustained component, reflecting steady-state release from a partially filled RRP, where the RRP remaining occupancy encodes absolute luminance (Oesch and Diamond, 2011; Burrone and Lagnado, 2000).

2.2.2 Adaptation in amacrine cells

Amacrine cells can modify the contrast encoding at the bipolar ribbon synapse (Oesch and Diamond, 2019) and are known to be a common postsynaptic player to receive depressed inputs from the rod bipolar cell ribbon synapse (Dunn and Rieke, 2008; Singer and Diamond, 2006). However their precise role in retinal adaptation to luminance and contrast is not clear. Amacrine cells have been proposed to be essential contributors to adaptation to more complex visual features such as orientation. Retinal ganglion cells have been shown to become less sensitive to orientations abundantly present in a stimulus but strongly respond to rare orientations which have not been seen in the recent stimulus history (Gollisch and Meister, 2010). Amacrine cells

have been proposed to mediate this adaptation via 'network plasticity', by strengthening inhibition in response to prolonged stimulation via the abundant visual feature, rendering ganglion cell receptive fields more sensitive to the opposite stimulus properties where inhibition remains low (Hosoya, Baccus, and Meister, 2005). Alternatively, amacrine cells have been proposed to mediate adaptation to oriented features via fast feed-forward inhibition silencing inputs from bipolar cells signalling a perceived orientation, while not observed orientations remain excitable (Johnston et al., 2019).

In addition to shaping adaptation to complex features, GABAergic amacrine cells have been shown to undergo themselves short-term depression (Li, Vigh, and Gersdorff, 2007; Vickers et al., 2012), while depression in glycinergic amacrine cells has recently been proposed to play a role in integrating ON on OFF inputs (Huang et al., 2022). While plasticity in excitatory synapses is generally thought to more efficiently encode strong signals by depressing ganglion cell responsiveness, plasticity in inhibitory synapses has been shown to lead to sensitization in a separate population of ganglion cells, whose sensitivity is elevated after stimulation (Kastner and Baccus, 2011; Kastner et al., 2019; Nikolaev et al., 2013). Sensitization via depression of inhibition has been proposed to play a functional role for a predictive coding strategy implemented in the retina (Kastner and Baccus, 2013a; Kastner and Baccus, 2013b), which will be elaborated in the following section.

2.3 Predictive Coding in the Retina

Predictive coding theory generally posits that sensory neurons transmit only unpredicted portions of an incoming sensory signal. To this end sensory neurons are thought to maintain a prediction of the expected neuronal input and emit error signals only when the actual input does not match with the predicted one (Palmer et al., 2015; Huang and Rao, 2011).

From an information theory point of view it is thought that neurons aim to transmit as much information as possible using a minimal amount neural resources: They generate a maximally informative non-redundant "efficient code" by removing predictable parts from the raw stimulus to maximize information transmission (Barlow et al., 1961).

The retina has been one of the first neural systems in which the efficient coding hypothesis has been tested. Srinivasan, Laughlin, and Dubs, 1982 suggested that the center-surround structure of RGCs is built to act as to reduce spatial redundancy in the visual input by subtracting the input in neighboring cells from the output response via surround inhibition, rendering cells selectively responsive to spatial changes (edges). In a similar manner, temporally displaced excitation and inhibition perform change detection on a temporal axis: visual inputs are transmitted via excitation while inhibitory inputs mediate predictions based on previous inputs (Denève and Machens, 2016). As long as prediction and input match, a cell remains silent and only responds when predictions and actual input deviate. The expectation is "stored" in the membrane potential of a receiving cell, set to a certain level by the balance of excitatory and inhibitory inputs, representing visual input and prediction, respectively.

Visual scenes are highly variable, it is thus necessary that neural circuits refine predictions by adapting them to the current environment. As described in the previous section, neuronal circuits may store stimulus features such as light intensity, contrast and variance in their response gain (Oesch and Diamond, 2011; Ozuysal and

Baccus, 2012). The retina has also been shown to adapt to more complex stimulus features such spatio-temporal correlations in the input. By dynamically reshaping their receptive field via synaptic adaptation, retinal circuits become less sensitive to abundantly present stimulus features - such as vertical edges in a forest (Cui et al., 2016; Hosoya, Baccus, and Meister, 2005; Johnston et al., 2019). Synaptic adaptation is thus likely to be a core element of a predictive coding implementation in the retina.

While the aforementioned examples all aim to reduce the redundancy of the neuronal code in predictable environments, is has as well been suggested that at the same time, redundancies in visual scenes are important to form predictions or expectations about future inputs (Barlow, 2001). It has been shown that expectations in the retinal network might as well serve to facilitate the detection of likely inputs, as it would be advantageous in noisy environments where a visual scene is less predictable. When an object moves across the surface of the retina, Kastner and Baccus, 2013a showed that depressing inhibition from the surround region acts as a prediction of the object to nearby cells by providing an estimate of the objects' likely position in the future: Depression of inhibition sensitizes the ganglion cell to a feature present nearby, and this sensitization serves as a prediction of the presence of this feature at a certain location - and at a certain time point in the future.

Consistent with the idea that the retina carries a prediction of the visual environment and follows an efficient coding strategy, it has been shown to be able to recognize temporal patterns in its input and to signal omissions in predictable pattern repetitions (Schwartz et al., 2007a; Schwartz et al., 2007b). The retina is also known to form predictions on a spatio-temporal scale, namely it anticipates the trajectory of a moving object and signals surprise when movement suddenly starts, changes direction or the object disappears (Berry et al., 1999; Chen et al., 2013; Chen et al., 2014; Ding et al., 2021).

Fundamental to these pattern detection mechanisms is that that neural systems can detect temporal regularities (Barne et al., 2022) and form predictions which contain a precise temporal expectation of future sensory inputs. This temporal selectivity might as well be implemented via short-term plasticity (Moraitis, Sebastian, and Eleftheriou, 2018) - a theory which is at the core of this study.

2.3.1 Detection of Temporal Pattern Violations: The Omitted Stimulus Response

A simple yet sophisticated example of temporal pattern recognition in the retina is the Omitted Stimulus Response (OSR), which has been first described by Schwartz et al., 2007a. Inspired by the phenomenon of mismatch negativity, which is a cortical signature of novelty in regular stimulus patterns, they set out to test if such a novelty response - or error signal - can already originate in the retina. When presented with periodic sequences of spatially uniform dark flashes, most ganglion cells responded very weakly to each flash. But when the periodic sequence ended, they emitted a burst of activity signaling the missing stimulus - the OSR.

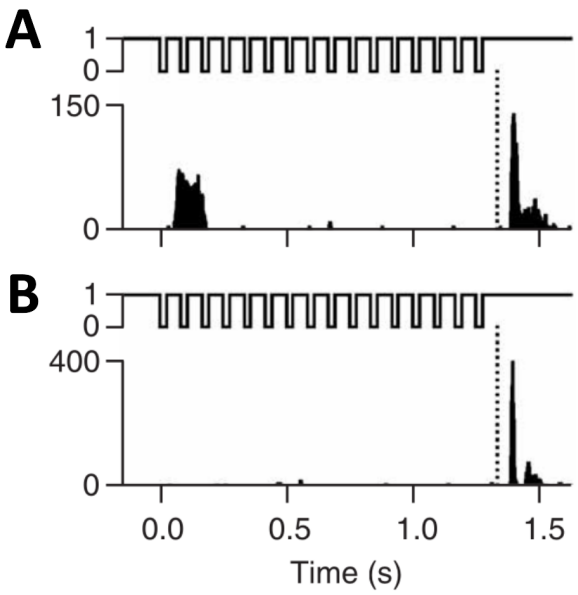

FIGURE 2.3: The omitted stimulus response signals temporal pattern violations.

A-**B.** Firing rates of two different retinal ganglion cells in the mouse retina in response to a repetitive train of dark flashes presented at 12 Hz. Upper panel shows the stimulus while the lower panel shows the cellular response. Cells emit a massive response peak at the end of the stimulus sequence, signalling the missing stimulus. Some cells also respond to the onset of the stimulus (**A**) while others do not (**B**). From Schwartz et al., 2007a.

Moreover, they found that not only could ganglion cells detect a missing flash after periodic stimulation, they also had a temporal prediction of the time when the next flash should have occurred. This predictive component was evident in the timing if the OSR: The latency to the response peak with respect to the time of the last flash increased as the period between flashes increased. When the stimulus period increased by 100 ms, the latency of the OSR shifted by approximately the same amount.

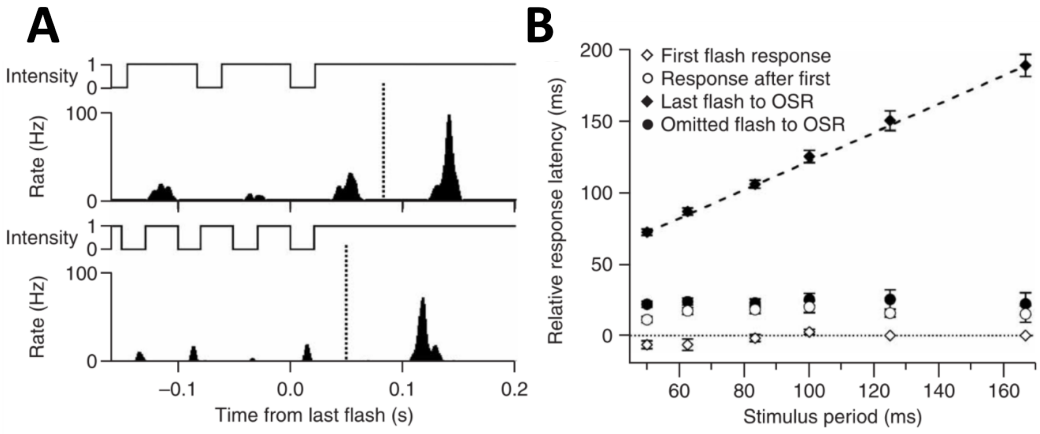

FIGURE 2.4: The omitted stimulus response is predictive

A Firing rate of a ganglion cell in response to flash trains presented at 12 Hz (upper) or 20 Hz (lower). Traces are aligned such that the last flash in each sequence is at 0, stimulation begins at different negative times ahead on the axis. Dotted lines indicate the time at which the next flash would have occurred. **B** Latency of ganglion cell responses plotted against stimulus period. The latency relative to the OSR increases by the same amount as the stimulus period (black diamonds). The dashed line has a slope of nearly 1. The latency relative to the time when the omitted flash should have occurred is nearly constant (black circles), as is the latency to the first flash (white diamonds) and consecutive flashes in the sequence (white circles). From Schwartz et al., 2007a.

The latency scaling of the OSR strongly implies that the retina has a temporal prediction of when the next flash in a periodic sequence should occur. Nevertheless, it remains debatable whether the OSR might simply be a response to a contrast step at the end of the stimulus, as the mean light level during the sequence of dark flashes is lower than between trials. To confirm that the OSR is indeed arising from a violation of the retina's prediction and not simply triggered by an overall change in light level, Schwartz et al., 2007b designed a stimulus where light intensity between flashes was higher than the background intensity such that the mean level remained constant. They found that the OSR indeed persists even when the average luminosity during flash sequences is kept at the baseline illumination in between trials.

Relevant retinal circuitry

In order to identify the required retinal circuitry to generate an OSR, Schwartz et al., 2007b performed a line of pharmacological experiments to systematically inhibit specific cell types in the salamander retina. They identified that ON bipolar cells are necessary for an OSR to dark flashes, as the OSR was completely abolished after application of APB, an agonist specific to glutamate receptors exclusively present in ON bipolar cells. On the contrary, the OSR to bright flashes remained when ABP was applied (Weidmann, 2009). This suggests that the OSR to dark and to bright flashes likely arises from distinct pathways, and that bipolar cells sensitive to the opposite polarity of the stimulus play an important role in its generation. Moreover, they used a variety of pharmacological agents to block GABAergic and glycinergic amacrine cell transmission but found that the OSR was largely unaffected. To our knowledge, all of the aforementioned experiments were carried out in the salamander retina and potential changes in the latency scaling were not investigated.

Models of the OSR

Based on the aforementioned experiments, several theoretical models have been proposed to elucidate the mechanisms behind the Omitted Stimulus Response. First, Werner, Cook, and Passaglia, 2008 proposed a dual LN-model with biphasic ON and OFF pathway interactions, which accurately captures the response peak after stimulus end via the rebound phase of the pathway selective to the opposite polarity of the stimulus. However, the latency shift with a slope of 1 is here only present in the onset of the OSR, but the model fails to shift the peak latency as a function of the stimulus period.

Gao and Berry (Gao et al., 2009) proposed that intrinsic oscillatory activity in ON bipolar cells can lead to prolonged periodic activity after the sequence of periodic stimuli ends. In their model, oscillatory activity is evoked via calcium-dependent-potassium channels. The frequency of the residual oscillations adapts to the stimulus frequency via a different calcium levels in the synaptic terminals. However, such oscillatory activity was not found in bipolar cells (Deshmukh and Berry, 2019) and this model could only account for a much smaller range of frequencies (with 1 single parameter set) than observed experimentally.

More recently, Tanaka et al. (Tanaka et al., 2019) proposed that the OSR with its latency shift can arise in a deep neural network model via summation of multiple excitatory inputs with different time constants. The explanation behind this model is that the OSR latency is determined by the sum of 2 ON bipolar cells which are activated only by certain stimulus frequencies due to different temporal filtering. It is difficult however to evaluate whether the model accurately captures the latency shift as observed in experiments, with the correct slope value, and this hypothesis has not been backed up by experimental evidence. Thus far, a verified mechanistic explanation for the OSR and its latency scaling is still lacking.

More recently the OSR phenomenon has been studied from a more functional perspective, where the amount of surprise in the OSR can be quantified and depends on a cells expectations or 'internal model' of the stimulus statistics (Despotović, 2022).

Temporal pattern recognition across neuronal systems

Responses to omissions in periodic stimuli that originate in the retina have first been reported in event-related potentials (ERP) of several fish species (Bullock et al., 1990), and have been shown to exist even in humans (Bullock et al., 1994; McAnany and Alexander, 2009). Consecutively, omitted stimulus responses have been confirmed to be present already in the output spikes of mouse and salamander retinas (Schwartz et al., 2007a), suggesting that predictions are formed even at very early stages of sensory processing. Examples of surprise signalling to unexpected stimuli have been observes across sensory modalities. A very prominent surprise response was discovered by (Näätänen, Gaillard, and Mäntysalo, 1978) in the human auditory cortex - the mismatch negativity response (MMN). When participants are presented with an "oddball" stimulus, where 'deviant' tone randomly occurs in a sequence of 'standard' tones, auditory ERPs to the deviant stimulus significantly differ from the standard ones (Näätänen et al., 1993; Garrido et al., 2009; Ulanovsky, Las, and Nelken, 2003; Li et al., 2017). A visual counterpart of the MMN has later on been discovered (Cammann, 1990), and is has been reported to be evoked by irregularities in a wide range of visual features, from line orientation (Astikainen, Lillstrang,

and Ruusuvirta, 2008) to facial expressions (Astikainen et al., 2013) (for review see (Kremláček et al., 2016)).

The anticipation of periodic events has even been observed at the behavioral level of simple organisms: When amoebas were periodically exposed to unfavorable conditions at constant intervals, they reduced their locomotive speed in response to each episode and continued to do so at the time when the next unfavorable episode would have occurred, even though experimental conditions were kept favorable (Saigusa et al., 2008).

Even though Omitted Stimulus Response is a response to a very artificial stimulus of high-contrast full field periodic flashes which are unlikely to occur in natural scenes, the ubiquity of similar phenomena in nervous systems implies that there might exist a general computational strategy for deviance detection. The underlying mechanisms of temporal processing in the rather artificial OSR may be thus embedded in more complex networks detecting temporal pattern violations in more realistic inputs.

2.3.2 Prediction and surprise in spatio-temporal patterns

Besides predicting the temporal structure of a spatially uniform stimulus, the retina has been shown to extract and predict the moving objects from a visual scene and to respond massively to sudden changes of motion trajectories. Many retinal cells (especially ganglion, but amacrine and bipolar as well) have been shown to extract motion in the visual scene (Baccus et al., 2008), predict the trajectory of a moving object (Berry et al., 1999) and track a moving object accurately (Leonardo and Meister, 2013). In addition to predicting motion in a visual scene, retinal ganglion cells have also been shown to detect violations in spatio-temporal patterns, such as motion onset of a previously static object (Chen et al., 2013), sudden reversal of the motion trajectory (Schwartz et al., 2007c; Chen et al., 2014) and the disappearance of a moving object behind an occluder (Ding et al., 2021).

Motion Anticipation

When the retina is presented with a static stimulus, retinal ganglion cells will respond to this stimulus with a delay of around 100 ms, due to the slow process of photo-transduction in photoreceptors and additional delays in the subsequent circuitry. If the retina represented a moving object simply by tracing its position across space, the signalled position would thus lag behind the actual position. Yet, the retina has been shown to compensate this delay (Berry et al., 1999): When a bar is moving across the receptive field of a retinal ganglion cell (RGC), the peak-firing rate of the cell occurs earlier as if the bar is flashed stimulus above the receptive field center. This phenomenon has been coined "motion anticipation" and implies that already at the inner retina, cells form a prediction of the future position of a moving object.

Higher visual areas estimate the location of an object in visual space by reading out which retinal ganglion cells respond to it. As a consequence, a population of ganglion cells forms an accurate representation - or neural image - of the position of the bar in the visual field. The neural image of a ganglion cell can be visualized by plotting its firing rate as a function of the distance between the bar position and the its receptive field center. When this response peaks before the bar is centered above the receptive field, the cell anticipates the object (Figure 2.5 B).

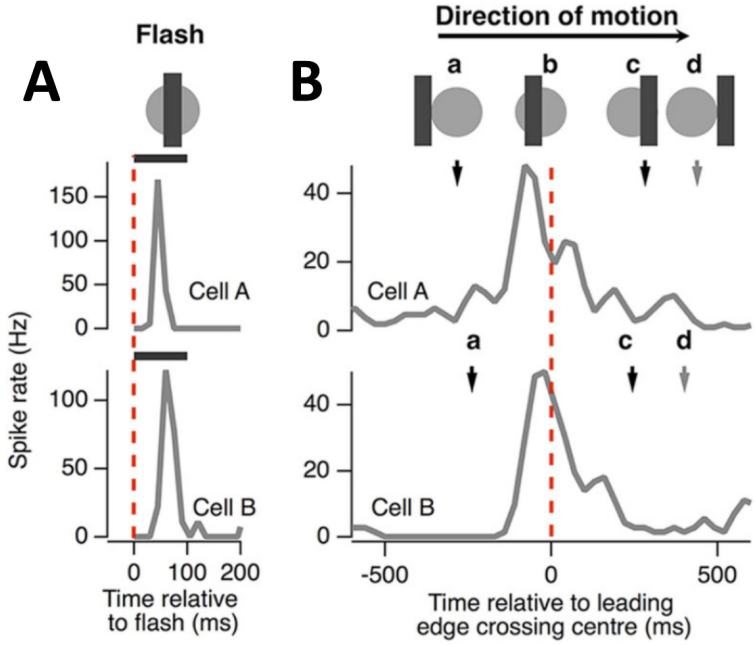

FIGURE 2.5: Motion anticipation of a moving bar

A Firing rate responses of two ganglion cells to a dark bar (width 160 μm) flashed over their receptive field centres for 100 ms (dark bar over firing rate). The response peaks occur around 50-80 ms after the bar appeared.

B Firing rate response of the same cells to a bar moving at 500 $\mu m/s$. The position of the bar relative to a cell's receptive field is shown above for the different time points indicated with a-d. The red line indicates the time point at which the leading edge of the bar is aligned with the receptive field center (0 on the temporal axis). Both cells emit their peak response 50-100 ms before the bar reaches their RF-center. From Johnston and Lagnado, 2015.

Importantly, in order to give an accurate estimation of a moving objects position, a cell needs to have a temporal representation of the speed with which the object is crossing a visual field. The distance an object has travelled within phototransduction and consecutive delays depends on the object speed. To accurately signal an objects' real position, a cell would thus need to adjust its anticipation to the object speed. Berry et al., 1999 showed that up to speeds of about 1 mm/s, the firing profile among ganglion cells indeed remained essentially unchanged, with a peak near or ahead of the leading edge. However, cells began to show a lag in their neural image at speeds faster than 1 mm/s.

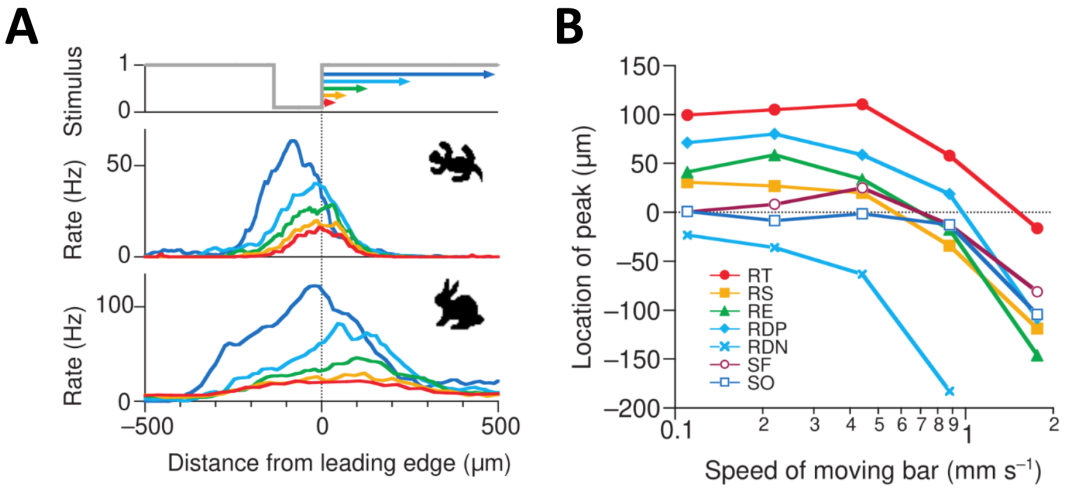

FIGURE 2.6: Scaling between motion anticipation bar speed

A Stimulation with moving dark bars of varying speed (0.11, 0.22, 0.44, 0.88 and 1.76 mm/s, upper panel) and response profiles of a ganglion cell in the salamander (middle) and rabbit retina (lower). **B** Motion Anticipation as a function of bar speed for various types ganglion cells. The distance between the peak in response profile and leading edge of the moving bar is plotted as a function of speed; positive numbers indicate anticipation. The functional types are: salamander, fast OFF (SF); salamander, other OFF (SO); rabbit, brisk-transient OFF (RT); rabbit, brisk-sustained OFF (RS); rabbit, local edge detectors (RE); rabbit, ON/OFF direction-selective cells in preferred (RDP) and the null direction (RDN). From Berry et al., 1999.

Both computational and experimental studies identified several mechanisms that contribute to motion anticipation, such as gain control in bipolar and ganglion cells (Berry et al., 1999; Chen et al., 2013), inhibition via amacrine cells on a cellular (Johnston and Lagnado, 2015) and network level (Souihel and Cessac, 2021) and lateral gap junction coupling (Souihel and Cessac, 2021; Liu et al., 2021).

In a simple network where ganglion cells pool over several bipolar cells that respond to the moving object, the peak response will lag behind the actual position of the object due to the temporal integration into bipolar (and ganglion) cell voltage. The following mechanisms have been proposed to counteract this delay:

(1) As first described by (Berry et al., 1999), gain control can correct for this delay in the following manner: As soon as the bar enters the receptive field of a cell, it starts to be activated, but then gets suppressed rather rapidly via a gain factor that quickly drops to 0. Gain control thus truncates the response end yields a peak response that occurs early on, before the object reached the receptive field center of the respective cell.

(2) Johnston and Lagnado, 2015 proposed that motion anticipation can be mediated via feed-forward inhibition from amacrine cell inputs that specifically suppress the response to the moving object in the latter half of the receptive field. This mechanism as well truncates the response and yields an early response peak.

(3) Lee and Menz, 2020 observed that motion anticipation can come in two flavors, either by truncation responses as described above, or by advancing the onset of the response, yielding a strong anticipation. They suggest that hyper-polarizing amacrine cells provide dis-inhibitory input to ganglion cells prior to the object crossing the receptive field center.

(4) Souihel and Cessac, 2021 explore yet another potential effect that amacrine

cells can have on motion anticipation: in a mathematical study of a retinal network with lateral inhibitory connectivity between bipolar cells, they show that complex indirect effects can arise from the network connectivity, inducing a wave of activity that amplifies the bipolar cell response and enhances the effect of gain control.

(5) Finally, gap-junction coupling between bipolar (Liu et al., 2021) or ganglion cells (Souihel and Cessac, 2021) may pass excitation laterally to neighboring cells that do not yet perceive the object in their receptive field. This lateral excitation has been shown to increase the area bipolar cell excitability way beyond the receptive field inputs from the OPL (Sidney P. Kuo and Rieke, 2016). This could lead to an early onset in motion anticipation as well.

Especially the gain-control mechanism has received a lot of attention in the consecutive literature and has been shown to successfully account for more complex motion, such as tracking the position of a moving object on a 2D-plane (Leonardo and Meister, 2013). However, the principle of gain control is a rather broad and phenomenological description and could have many underlying biophysical mechanisms.

Motion onset and reversal detection

If the retina implements a predictive coding strategy to efficiently encode the position of a moving object, one expects that its emits an error signal if the tracked object moves in an unpredicted way. To this end, Schwartz and Berry, 2008 investigated how retinal ganglion cells respond to "motion reversal", that is a moving bar that suddenly changes is trajectory by 180 degrees. While RGCs anticipated the moving bar during smooth motion in one direction, they also generated a large response when the bar was suddenly reversing its motion near the ganglion cells' receptive field. This reversal response occurs with a constant latency regardless of the exact location of the reversal in a cell's receptive field, resulting in a large synchronised response within nearby ganglion cells, signalling the unexpected change in trajectory. A static bar centered above a cells' receptive field evokes a transient response at appearance, but yield little response while it stays still. When it suddenly starts moving, ganglion cells strongly respond - even more than to smooth motion across the entire receptive field (Chen et al., 2013).

Both motion onset and reversal responses have been accounted for by a phenomenological Adaptive-Cascade Model (ACM), where a similar gain control mechanism as the one responsible for motion anticipation acts at two stages in the retinal network, at the level of non-linear receptive field sub-units (bipolar cells) and ganglion cells (see Appendix A).

In this architecture, the response to smooth motion in each sub-unit is suppressed due to preceding activation by an earlier segment of the bar due to reduced gain. However, gain control takes a certain amount of time to start, which is why the gain is not affected early after motion onset phase. The response amplitude is thus bigger than to smooth motion. Since bipolar sub-units in the model have a comparatively small receptive field, gain control in bipolar cells can only produce a small spatial shift. Therefore a second gain control at the ganglion cell level is needed to sufficiently shift the response peak to match experimentally observed anticipation levels (Chen et al., 2013).

This model can also account for the motion reversal response with constant latency in the following manner: When the moving bar passes over a bipolar cell it activates it and initiates the initial response to smooth motion. At the same time,

that cell has its gain strongly suppressed, rendering it unresponsive for a characteristic time τ until the gain recovers. Therefore, after motion reversal, the bar will not generate responses in bipolar cells that it just passed until after τ has elapsed and their gain recovered. This principle holds for any reversal location within the receptive field of a ganglion cell, yielding a response time τ after reversal regardless of the exact reversal location. The ganglion cell gain control also plays a role and stabilizes the constant latency in a similar manner (Chen et al., 2014).

More recently, Ding et al., 2021 showed that direction-selective ganglion cells, which classically only respond to one direction of moment, respond to motion in their null-direction when the object is occluded while moving over their receptive field center. This signalling helps the population of retinal ganglion cells to signal the spatial location of moving objects when their smooth trajectory is unexpectedly interrupted. To our knowledge, a mechanistic explanation for this phenomenon yet needs to be identified.

All-together, these examples indicate that the retina employs a predictive coding strategy, detects and signals surprise about spatio-temporal patterns and that the timing of these surprise signals is coordinated and adapted to the stimulus to precisely signal spatial and temporal changes.

2.4 Modelling retinal responses

We have seen so far that the retina is a neuronal network with structural and functional complexity that preforms a number of extraordinary computations for efficient information processing. While experimental studies have provided a vast amount insight into the physiology and biochemistry of the retina, a detailed understanding of how the retinal architecture and physiology leads to sophisticated response mechanisms such as prediction and surprise remains unclear and likely cannot be answered using experimental methods alone. Therefore, computational neuroscience is a theoretical approach to understanding the underlying mechanisms of neural signals by modeling the potentially relevant processes in nervous systems and simulating their contribution to neuronal responses (Roberts et al., 2016).

Neuronal systems can be modelled at many different scales, ranging from detailed biophysical process in single neurons to large networks where thousands of cells are represented in a simplified way. Depending of the retinal property under investigation, computational models of retinal responses thus range form the single cell to network level (Guo et al., 2014). The level of abstraction used for the description of neuronal responses thereby lies on a continuum between two extreme categories, phenomenological and mechanistic (Ooyen, 2011). On one end, phenomenological models aim to replicate the experimental data with minimal complexity, where model components do not need to correspond to the underlying biology. On the other extreme, mechanistic models are comprised of mathematical equations which directly represent biological elements and their actions, with the aim to understand how individual components influence the behavior of the system.

To find the right level of complexity for a computational model, it is necessary to trade-off between biological plausibility and modelling simplification, as to interpret the models behavior in a meaningful way.

In this thesis, we employ computational neuroscience as a tool to investigate cellular and circuit mechanisms underlying retinal responses by implementing biophysical mechanisms in phenomenological descriptions of neuronal responses. The

parameterization of our simulations are guided by fits to experimental data and mathematical analysis. This section will therefore review relevant principles of theoretical neuroscience to understand the computational methodology used in subsequent chapters.

2.4.1 Linear-Nonlinear Models

One of the most widely spread models used to describe retinal responses is the linear-nonlinear model (LN-model) (Chichilnisky, 2001; Baccus and Meister, 2002). This model typically consist of two processing steps, a linear filter and a non-linearity (Figure 2.7). An incoming light stimulus $s(t)$ is convoluted with a linear filter $F(t)$ representing the receptive field of the neuron, describing how it integrates the stimulus over space and time. This convoluted signal can be interpreted as the cells voltage response to the stimulus, $g(t)$. It is then passed through a nonlinear function $N(t)$ to predict a firing rate $r'(t)$. The non-linearity $N(t)$ is typically piece-wise linear function with a threshold that needs to be passed to evoke a spiking response. LN-models provide good approximations of recorded average firing rates in response to simple stimuli (Pillow et al., 2008). However, they fail to accurately capture retinal responses to more complex stimuli. Therefore, the general structure of the LN model has been extend to cascades of LN models with multiple stages and linear-nonlinear sub-units (Schwartz et al., 2012; Gollisch, 2013) and can for example account for differential motion (Baccus et al., 2008) or approaching motion (Münch et al., 2009). The ACM described in the previous section is another example of such a multi-layered model with LN-subunits, which can account for motion anticipation, motion onset and motion reversal via an added contrast gain control mechanism (Chen et al., 2013; Chen et al., 2014).

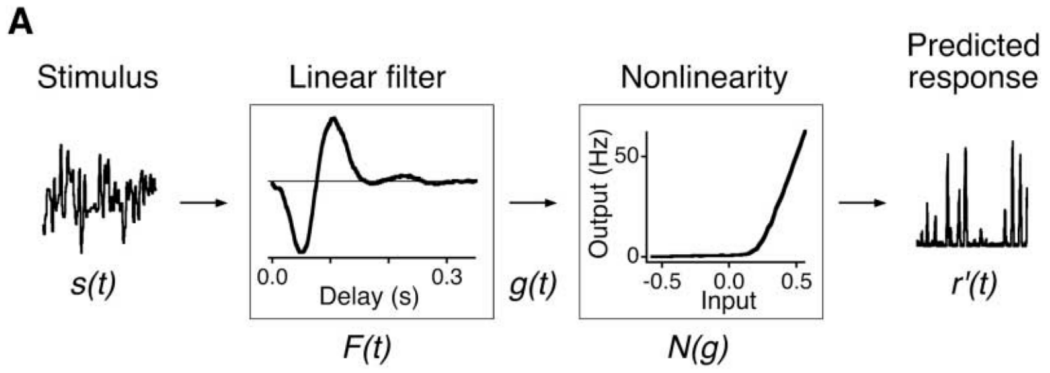

FIGURE 2.7: The Linear-Nonlinear (LN) model.

A stimulus $s(t)$ is first convolved with a linear filter $F(t)$, yielding the linear response prediction $g(t)$, and then passed through a nonlinearity $N(g)$ to produce the predicted firing rate $r'(t)$.

The spatial and temporal shapes of the linear filter can be estimated from experiments where spiking responses to a white-noise stimulus, typically a 2D checkerboard with random pixels fluctuating over time, are recorded. By averaging over all stimuli preceding a spike, a 3-dimensional spike-triggered average (STA) can be obtained (Chichilnisky, 2001), which gives and approximation of the receptive field of the recorded cell. It represents the spatial and temporal stimulus properties a cell is responsive for. The STA can be decomposed onto a spatial and a temporal profile (Figure 2.8). The temporal STA describes the light intensity change that most excites

the cell, while the spatial profile reveals the region in space to which the neuron is responsive and exhibits the typical center surround structure.

For modeling purposes, the (whitened) STA can be explicitly used as linear weighting of the stimulus over space and recent time. To reduce the parameter space of the linear filter, it is often represented by parametric fits of smooth function to the spatial and temporal profile, allowing for spatial and kinetic measures of the receptive field (Chichilnisky and Kalmar, 2002).

The temporal profile can be approximated by a difference of α-kernels (Figure 2.8 **A**), which can be interpreted as a cascade of two low-pass filter :

$$k_t(t) = \frac{t}{{\tau_{RF}^+}^2} e^{-\frac{t}{\tau_{RF}^+}} - w_t \frac{t}{{\tau_{RF}^-}^2} e^{-\frac{t}{\tau_{RF}^-}} \tag{2.1}$$

The time constants τ_{RF}^+ and τ_{RF}^- give a measure of the cells' characteristic integration times of bright and dark stimulus intensities and w_t is a measure for the relative strength of between the two. RGCs differ in the temporal speed with which they integrate signals and are often selective for bright or dark stimulus intensities, yielding a temporal filter with either positive of negative response phases. In addition, RGCs might have monophasic filter (only one positive or negative phase), which leads to sustained response behavior, or biphasic filter (both positive and negative phases), resulting in transient responses (Suh and Baccus, 2014).

The spatial profile can be approximated by a Difference-of-Gaussian profile (Figure 2.8 **B**):

$$k_x(x) = e^{-\frac{(x-x_c)^2}{2\sigma_s^2}} - w_s e^{-\frac{(x-x_c)^2}{2\sigma_c^2}}, \tag{2.2}$$

where x_c is the position of the receptive field, and σ_c and σ_s are the the sizes of the receptive field center and surround. The parameter w_s is a scalar for the relative strength of the surround.

Depending on the dimension of stimulus to which the neuronal response is to be predicted, one may use only the relevant STA dimensions. For example, if the model is used to predict neuronal responses to full field stimulation, only the temporal profile may be used as a linear filter. If the stimulus has a 1- or 2- dimensional spatial structure, a corresponding spatial filter can be added (Figure 2.8 **B,C**).

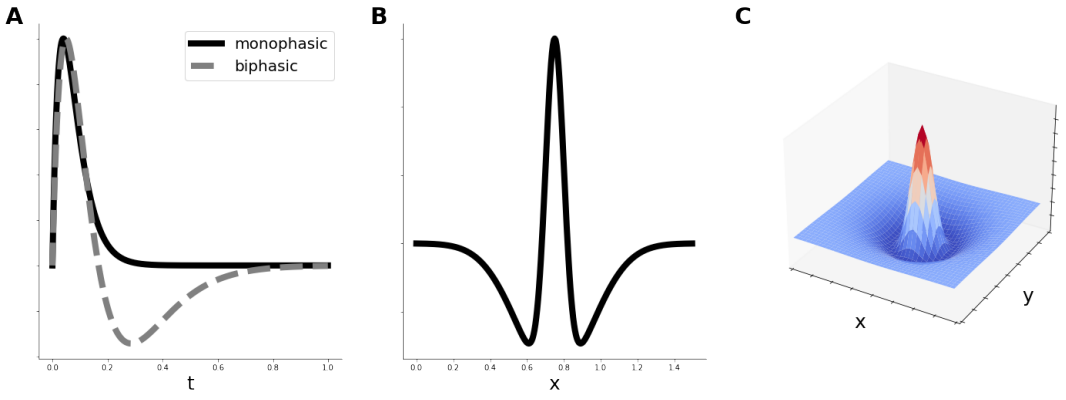

FIGURE 2.8: Spatial and temporal kernels for linear receptive field approximation

A Temporal kernel of the receptive field $k(t)$, from eq. 4.4 can have a monophasic (black) or biphasic (grey) shape. **B** 1-dimensional spatial kernel $k(x)$ from 4.5 with a postive center and a negative surround. **C** 2-dimensional spatial kernel $k(x, y)$.

Even though LN-based models perform very well at predicting neuronal responses to a wide range of stimuli, they provide little insight into biophysical mechanisms underlying the response generation.

2.4.2 The retina as a dynamical system

The previously described LN model approach employs a phenomenological description of neuronal activity, treating intrinsic neuronal processes as a black box and only considering its input-output function. On the other hand, mechanistic neuron models aim to shed light on the mechanisms with which neurons process information. These models generally rely on the biophysical mechanisms which underlie neuronal activity and highlight their role in generating a certain neuronal response. The level of detail in the theoretical description of a neuron thereby depends on the physiological properties that are investigated and thought to be relevant to evoke a certain response behavior (Guo et al., 2014). By describing each neuron in a large network by a single compartment model with one differential equation, one can study dynamical circuit interactions in a biologically plausible set-up (Sterratt et al., 2011). This enables mathematical analysis for an in-depth understanding of how physiologically relevant parameter act on the system (Ermentrout and Terman, 2010; Wohrer and Kornprobst, 2009; Cessac, 2022).

Neuronal Representation

In this setup, the characterizing quantity of the neuron is its membrane potential V, described by one differential equation. The complexity of the neuron is reduced to two components, its impermeable hydrophobic membrane and its ion channels which allow ionic currents to flow across the membrane. The neuron can then be described by an electrical Resistor-Capacitor (RC) circuit, where the capacitor represents the membrane and resistor elements represent ionic channels. In a conductance-based modelling approach ion channels are characterized by their conductance g. With these quantities we can define a neuron with n different types of ion channels via the general equation (Dayan and Abbott, 2001):

$$C_m \frac{dV}{dt} = -\sum_{i=1}^{n} g_i(V - E_i). \tag{2.3}$$

This equation describes how the membrane potential of a neuron with capacitance C_m changes due to currents flowing across the membrane. These currents are generated by ions passing across the membrane trough ionic channels in the membrane. The amount of current flow depends on the membrane conductance for the specific ion, g_i, and its specific reversal potential E_i, such that the current vanishes when the membrane potential satisfies $V = E_i$. Since most retinal neurons are not firing, we are neglecting spike generating ion channels such as voltage gated sodium or potassium channels. Constant currents flowing across the neuron in absence of any simulation (eg. via ionic pumps and voltage-independent potassium channels) are summarized by a passive leak current $I_L = g_L(V - E_L)$. The reversal potential of the leak current E_L is assumed to be equal to the rest potential of the cell. In addition, the neuron can be excited by synaptic inputs from a pre-synaptic neuron X via the synaptic current $I_X = g_X(V_X)(V - E_X)$, where E_X is the reversal potential of the ionic species of the synapse and

$$g_X(V_X) = \lambda_X N_X(V_X) \tag{2.4}$$

is the synaptic conductance that generally varies as a function of the pre-synaptic voltage V_X. The pre-synaptic voltage is rectified and has to pass a certain threshold θ_X to evoke a postsynaptic response, simulated via the function:

$$N_X(V) = \begin{cases} V - \theta_X, & \text{if } V \geq \theta_X \text{ ;} \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

Note that this dependence of post-synaptic conductance on pre-synaptic voltage is a simplification that sums up all biological steps of synaptic release, from calcium influx over transmitter release to ion-channel kinetics.

By convention, the membrane potential is defined as intracellular - extracellular potential, $V \equiv V_{in} - V_{out}$, while the total membrane current $i_m = -\sum_i g_i(V - E_i)$ is defined as positive outward. As positive ions leaving the intracellular medium leads to a more negative membrane potential V, the membrane current i_m therefore needs to have a negative fore sign.

Retinal neurons, especially bipolar cells can be activated by a visual stimulus that evokes a current input from photo-receptors, which can be simulated as an external (positive or negative) current input I_{ext}. Altogether, this yields the equation:

$$C_m \frac{dV}{dt} = -g_L(V - E_L) - \sum_{X,syn} g_X(V_X)(V - E_X) + I_{ext}(t). \tag{2.6}$$

This equation can be simplified in 3 steps. First, a characteristic time scale for each neuron is defined as:

$$\tau = \frac{C_m}{g_L + \sum_X g_X(V_X)}. \tag{2.7}$$

Note that this time constant depends on presynaptic voltages V_X. For simplicity, we neglect this effect of synaptic activity on the cellular time-constant and replace $g(V_X)$ by a static approximation $\bar{g}_X$:

$$\tau = \frac{C_m}{g_L + \sum_X \bar{g}_X}. \tag{2.8}$$

By introducing this simplification, any effects of synaptic activity on the time scale of the neuron are neglected. Second, a synaptic weight for each type of synapse X is introduced:

$$w_x = \frac{\lambda_x E_x}{C_m}. \tag{2.9}$$

Finally, the resting potential of a cell is referenced to 0, and $E_L = 0$. The general equation 2.6 can then be re-written as:

$$\frac{dV}{dt} = -\frac{V}{\tau} + \sum_X w_x N(V_X, \theta_X) + I_{ext}(t). \tag{2.10}$$

To simulate the retinal network with multiple cells, equation (2.10) can be implemented for each cell in the network. Based on the retinal organization, a typical retina model is structured in 3 different cell types, corresponding to bipolar cells (BCs), amacrine cells (ACs) and retinal ganglion cells (RGCs), respectively. Photorececeptor and horizontal cells are implicitly represented in the input the network

model receives. If a spatial organization is taken into account, each cell type forms a layer with N cells located on a 1- or 2-dimensional grid (Figure 2.9.

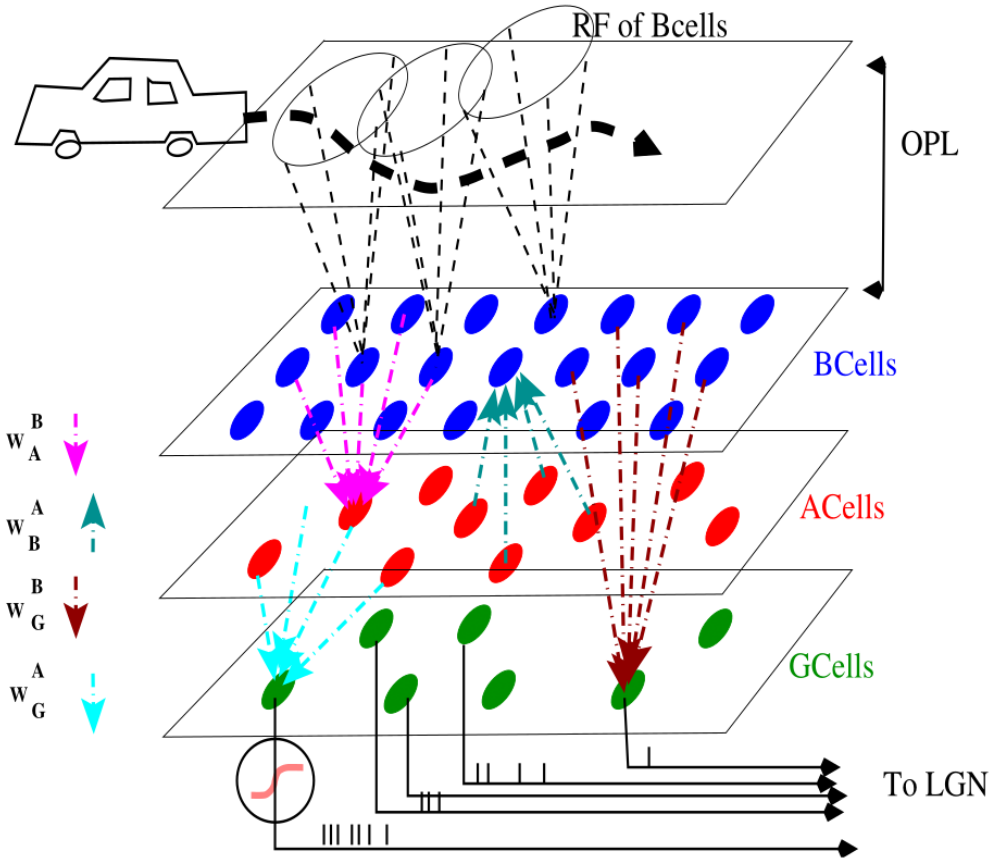

FIGURE 2.9: Structure of the retinal network model.

A stimulus with spatiotemporal structure (e.g. a moving car) is projected onto the outer retina, where it is transformed into a current input to BCs via processing by photoreceptor and horizontal cells. This process is simulated via a convolution of the stimulus with the Receptive Field (RF) of BCs. The network then consists of 3 cellular layers representing BCs (blue), ACs (red) and GCs (green). BCs synapse onto ACs in the vicinity (pink arrows) and ACs synapse back onto bipolar cells (green arrows). GCs receive synaptic inputs from the BC layer (brown arrows) and/or the AC layer (cyan arrows). The resulting GC voltage is passed through a non linearity as to produce spike trains conveyed to the LGN. From Cessac, 2022.

Connectivity

Assuming that synapses of the same type share the same connectivity weight, connections from BCs to ACs are weighted by an excitatory (positive) weight w_A^B while connections from ACs to BCs share an inhibitory (negative) weight w_B^A. In the same manner, connections from BCs and ACs to GCs are weighted by w_G^B and w_G^A respectively.

If all layers have the same number of cells N, the connectivity between two layers of the network is defined through NxN connectivity matrices Γ^{pre}_{post}, where $\Gamma^{pre_i}_{post_j} = 1$ if cell i on the pre-synaptic layer connects onto cell j in the postsynaptic one.

Assembling the full model

Finally, assuming that only bipolar cells receive direct inputs from photo-receptors, the joint dynamics of all cells in the network can be described by the dynamical system:

$$\begin{cases} \frac{dV_{Bi}}{dt} = -\frac{1}{\tau_B} V_{Bi} + w_B^A \sum_{j=1}^{N} \Gamma_{Bi}^{Aj} N(V_{Aj}, \theta_A) + I_{ext_i}(t), \\ \frac{dV_{Aj}}{dt} = -\frac{1}{\tau_A} V_{Aj} + w_A^B \sum_{i=1}^{N} \Gamma_{Aj}^{Bi} N(V_{Bi}, \theta_B) \\ \frac{dV_{Gk}}{dt} = -\frac{1}{\tau_G} V_{Gk} + w_G^B \sum_{i=1}^{N} \Gamma_{Gk}^{Bi} N(V_{Bi}, \theta_B) + w_G^A \sum_{j=1}^{N} \Gamma_{Gk}^{Aj} N(V_{Aj}, \theta_A). \end{cases} \tag{2.11}$$

For a full account of the retinal processing from the photoreceptor integration to RGC spiking response, this network description of retinal interactions by the system (2.11) can be combined with the LN- model approach. The dynamical system is then placed in between a linear filter and an output non-linearity. In this sense, the process of stimulus integration via photoreceptors and horizontal cells is simulated by spatio-temporal filtering (with equations 2.1 and 2.2). The filtered signal is then used as external input current I_{ext_i} into bipolar cells in the system (2.11). The retinal network then processes the signal, and the voltage responses of ganglion cells, V_{Gk}, serve as output of the system. This output is passed through the non-linearity function (2.5) to predict a firing rate $R_G(t)$.

Mathematical Analysis

The simplifications described above allow the dynamical system to be mathematically analyzed in order to gain insights into the effect of circuit interactions in the network based on interpretable parameters such as neuronal time constants and connectivity weights. If one assumes that all cells in the network operate in their linear range, that is to say above their rectification threshold, the system in (2.11) can be linearized to enable mathematical analysis (Souihel and Cessac, 2021; Cessac, 2022; Evgenia et al., 2023). It can be written in an alternative form:

$$\frac{\vec{X}}{dt} = L.\vec{X} + \vec{F}(t) \tag{2.12}$$

where $\vec{X}$ is the state vector of all membrane potentials in the network :

$$\vec{X}_\alpha = \begin{cases} V_{B_i}, \alpha = i, i = 1...N; \\ V_{A_j}, \alpha = N + j, j = 1...N; \\ V_{G_k}, \alpha = 2N + k, k = 1...N. \end{cases} \tag{2.13}$$

Here, V_{Bi} is the voltage response of BC i in the network, and V_{Aj} and V_{Gk} describe the voltage responses of AC j and GC k, respectively. The external inputs are introduced in $\vec{F}$ as:

$$\vec{F}_\alpha = \begin{cases} F_{B_i}, \alpha = i, i = 1..N; \\ 0, \alpha = N + j, j = 1..N; \\ 0, \alpha = 2N + k, k = 1..N, \end{cases} \tag{2.14}$$

where F_{B_i} is the filtered stimulus that serves as input into the system. Finally L is the transport operator of the system, which is an 3N x 3N matrix of the following form:

$$L = \begin{pmatrix} \frac{-I_{NN}}{\tau_B} & w_B^A \Gamma_B^A & 0_{NN} \\ w_A^B \Gamma_A^B & \frac{-I_{NN}}{\tau_A} & 0_{NN} \\ w_G^B \Gamma_G^B & w_G^A \Gamma_G^A & \frac{-I_{NN}}{\tau_G} \end{pmatrix} \tag{2.15}$$

Its diagonal entries are the inverse time constants of the system and the remaining entries are equal to w_{post}^{pre} if there is a synapse at the corresponding connection and 0 otherwise. The spectrum of L, that is to say its eigenvalues λ_β for $\beta = 1 \dots N$, control the behavior of the solution of the system in (2.13). Souihel and Cessac, 2021 showed that the impact of lateral connectivity between BCs and ACs on the spectrum of L is controlled by a dimensionless parameter μ:

$$\mu = -w_B^A w_A^B \tau^2, \tag{2.16}$$

which is governed by lateral connectivity weights w_B^A and w_A^B as well as the time constants τ_B and τ_A as :

$$\frac{1}{\tau} = \frac{1}{\tau_A} - \frac{1}{\tau_B}. \tag{2.17}$$

The parameter μ thus summarizes the network effect of lateral connectivity on the output response of the system and has been shown to determine the anticipatory effect of lateral connectivity to a moving bar (Souihel and Cessac, 2021) as well as the variability of GC responses, more specifically their characteristic response profile (mono- or bi-phasic) (Evgenia et al., 2023). This analysis holds only if the connectivity matrices Γ_B^A and Γ_A^B commute.

2.4.3 Modeling Short-Term Plasticity

Short-Term Plasticity is a broadly observed phenomenon at synaptic connections all across the nervous system and describes a rapid, activity-dependent modulation of synaptic efficacy. Several complex mechanisms at pre- and post-synaptic terminals can cause both depression (weakening) and facilitation (strengthening) of postsynaptic activity in response to persistent pre-synaptic activity. Mathematical models have been extensively used to study the mechanisms and roles of short-term plasticity in cortical information processing and temporal filtering (Lindner et al., 2009; Anwar et al., 2017; Hennig, 2013). The most prominent model for STP in the cortex, the Tsodyks - Markram Model, relies on a simplified phenomenological description of vesicle dynamics in the pre-synaptic terminal (Abbott et al., 1997; Tsodyks, Pawelzik, and Markram, 1998).

Here, the synaptic strength is governed by two dynamic variables - the pre-synaptic vesicle occupancy n (or fraction of vesicles available for release) and the release probability p. If the synapse has been active in a given previous time window, the synaptic strength typically decreases due to depletion of available resources, leading to short-term depression (STD) (Figure 2.10 **A**). This effect can be described by the equation (Hennig, 2013):

$$\frac{dn}{dt} = \frac{1 - n(t)}{\tau_n} - \sum_j \delta(t - t_j) p n(t), \tag{2.18}$$

where vesicle occupancy gets depleted based on pre-synaptic inputs $\sum_j \delta(t - t_j)$. The pre-synaptic input is as a spike train with spikes occurring at times t_j, simulated by a Dirac delta function δ, which is 0 for all values except when $t = t_j$, where it

integrates to 1 (Sterratt et al., 2011). The vesicle occupancy is replenished with a time constant τ_n and release is scaled by the release probability p, which may be a fixed parameter or a dynamic variable if joint facilitation occurs. In this case, persistent activity increases calcium influx to the synaptic terminal, which increases the release probability, resulting in synaptic facilitation (STF), described by (Hennig, 2013):

$$\frac{dp}{dt} = \frac{p_0 - p(t)}{\tau_p} + \sum_j \delta(t - t_j) a_f (1 - p(t)), \tag{2.19}$$

where release probability increases proportional to a_f, the amount of facilitation per presynaptic spike, and decreases back to baseline release probability p_0 with a time constant τ_p. Both n and p scale the post-synaptic response. Synapses can be either dominated by STD or by STF, or show a mixture of both. Figure 2.10 **A** shows two examples of post-synaptic responses (here synaptic currents) which are either dominated by facilitation or depression. The same pre-synaptic input can cause massively different responses depending on synaptic dynamics. Short-term changes in synaptic efficacy have profound effects on temporal filtering properties in postsynaptic responses (Figure 2.10 **B**). For example, STD suppresses synaptic efficacy in a frequency-dependent manner. The steady state of synaptic strength exponentially decays with increasing frequency, creating a low pass filter for pre-synaptic signals (Tsodyks and Markram, 1997). STF has an opposite filtering effect, letting preferentially faster frequencies pass, while synapses with a mixture of depression and facilitation band-pass filter pre-synaptic signals.

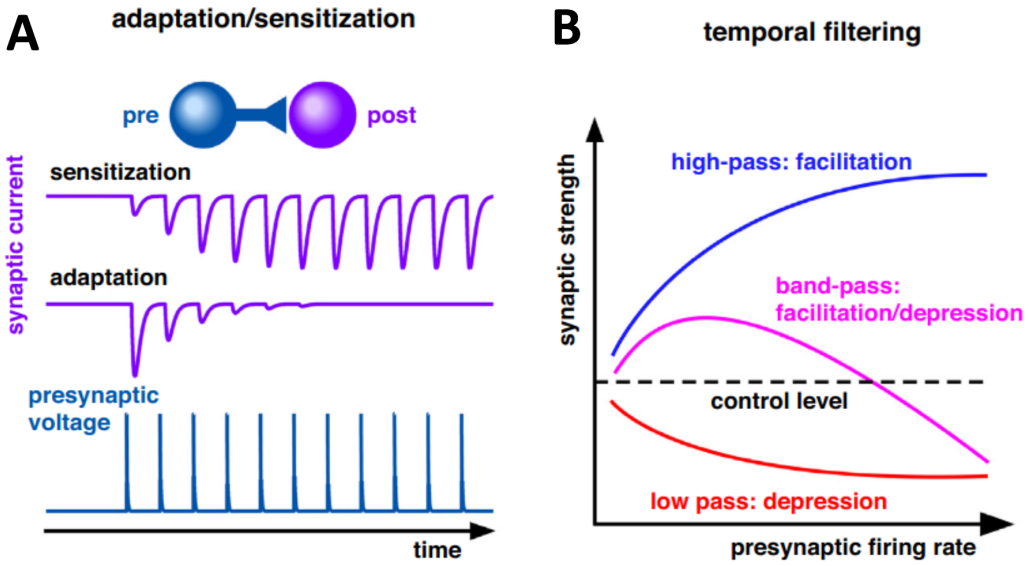

FIGURE 2.10: Short-term plasticity and its effect on temporal filtering.

A Short-term facilitation and depression have opposite effects on postsynaptic currents. A persisting pre-synaptic input (bottom row, blue) can have different effect on postsynaptic responses (upper rows, purple): Post-synaptic currents can either be facilitated (sensitization) or depressed (adaptation).
B STP conveys frequency-filtering properties to synapses. Facilitating synapses are high-pass filter pre-synaptic inputs by amplifying high frequencies (blue) while depressing synapses act as low-pass filters by reducing responses to high frequencies (red). Synapses exhibiting both facilitation and depression act as band-pass filter (pink). From Anwar et al., 2017.

A simple model for short-term plasticity in the retina

As stated above, cortical models of STP are based on spiking inputs, as neurons in the cortex mostly communicate via spikes. In the retina however, neurons mainly communicate via graded potentials. Short-term plasticity in retinal synapses thus needs to be based on continuous changes in pre-synaptic voltage.

The aforementioned mechanism of gain control (Berry et al., 1999; Chen et al., 2013; Chen et al., 2014) can be seen as a phenomenological description of short-term depression, where the reduction in response gain after continuous activation could come from depletion of synaptic vesicles. More detailed models of STP at retinal synapses consist of dynamical systems representing the dynamics of multiple synaptic vesicle pools in the ribbon synapse, where the value of one of the variables directly serves as input to the postsynaptic cell (Ozuysal and Baccus, 2012; Schröder et al., 2020).

For the work of this thesis we sought to implement STP in the previously described dynamical system network model. We therefore aim at a simple yet realistic description of the main properties of STP, which allows mathematical analysis. To this end, we chose to focus on synaptic depression, which we simulate via one kinetic equation describing synaptic vesicle occupancy, similar to cortical models of short-term plasticity (Tsodyks and Markram, 1997; Zhou et al., 2012; Hennig, 2013), described by equation (2.18). Considering non-spiking retinal neurons, we replace spiking inputs by continuous pre-synaptic voltage and introduce short-term plasticity into the network model in the following way: if a synaptic output weight from cell $X = A, B$ undergoes short-term depression, it is modulated by a dimensionless variable n, which interprets as a vesicle occupancy in the synaptic terminal. It obeys the kinetic equation:

$$\frac{dn^X}{dt} = (1 - n^X)k^X_{rec} - \beta^X k^X_{rel} N(V_X, \theta_X) n^X. \tag{2.20}$$

Here, k^X_{rel} is the release rate of synaptic vesicles, which determines how fast the vesicle pool get depleted. With the increase of pre-synaptic voltage V_X above threshold, vesicle depletion is activated and the synaptic vesicle occupancy n decreases. The pre-synaptic voltage is scaled by the scaling factor β^X. At the same time k^X_{rec}, the recovery rate, determines how fast the vesicle pool is replenished and n recovers from depression. To include this mechanism in the network, the synaptic weight from a pre-synaptic cell X to its post-synaptic partner, w^X_{post}, is scaled by plasticity:

$$w^X_{post}(t) = w^X_{post} n^X(t). \tag{2.21}$$

The evolution of vesicle occupancy can be characterized by a characteristic time of synaptic depression which is derived from equation (2.20):

$$\tau_{n^X} = \frac{1}{k^X_{rec} + \beta k^X_{rel} N(V_X)}, \tag{2.22}$$

and a steady state :

$$n^*_X = \frac{k^X_{rec}}{k^X_{rec} + \beta k^X_{rel} N(V^*_X)}. \tag{2.23}$$

From equation 2.23 we see that n^*_X increases with k^X_{rec} and decreases with $\beta^X k^X_{rel}$. Note that V^*_X depends on its pre-synaptic inputs, which are themselves scaled by

n_X from neighboring cells, introducing complex nonlinear effects. In general, the evolution of n^X is slow compared to the other neuronal dynamics.

2.5 Summary

Taken all together, the retina is a sophisticated neural tissue which enables the perception of light and allows the brain to form a representation of a visual environment. Its complex structural and functional organization selectively efficiently relays important information to the brain by predicting future input patterns and detecting unpredicted changes. At the same time, short-term plasticity allows the retina to adapt to a continually changing visual environment.

So far, the role of STP in predictive coding strategies in the retina has been mostly studied at the level of individual cells or local circuits (Hosoya, Baccus, and Meister, 2005; Kastner and Baccus, 2013a; Johnston et al., 2019). However, due to experimental inaccessibility, the effects of STP on the retinal output at the network level have been difficult to address so far. The integration of short-term depression via a simple yet meaningful approximation into a retinal network model allows us to study the effect of STP on the retinal network from a computational and mathematical perspective. Using a combination of theoretical modeling and electrophysiological experiments, we explore in the next chapters how STP in the retinal network may contribute to the formation of predictions at the examples of the Omitted Stimulus Response and motion anticipation.

Chapter 3

Temporal pattern recognition in retinal ganglion cells is mediated by dynamical inhibitory synapses

In this chapter, we study the underling mechanisms of temporal pattern recognition in the retina at the example of the Omitted Stimulus Response. The results of this chapter consist of a experimental and a theoretical part. First, we present experimental findings that inhibitory amacrine cells are required for the latency shift in the OSR, which is an important indicator of the temporal expectations the retina carries about its input. We then propose a computational model that is guided by these experimental results, with which we elaborate the hypothesis dynamic adaptation of inhibitory synaptic weights can serve as a biological implementation for temporal expectations.

This project is carried out in collaboration with Olivier Marre's team at the Vision Institute and all experiments have been carried out by Thomas Buffet, who is a PhD student in the lab, and Semichan Beret, who is now a PostDoc at the Pasteur Institute. The work of this chapter is available on bioRxiv and has been submitted to Nature Communications and is currently under revision.

Contents

3.1 Introduction

A long standing hypothesis is that visual neurons do not signal the visual scene *per se*, but rather surprising events, eg. mismatches between observation and expectation formed by previous inputs (Barlow et al., 1961). It has been observed in a number of sensory modalities that neurons strongly respond when a sequence of repetitive stimuli is unexpectedly interrupted (Ulanovsky, Las, and Nelken, 2003; Bullock et al., 1994; McAnany and Alexander, 2009). In the retina, this phenomenon has been coined the Omitted Stimulus Response (OSR) (Schwartz et al., 2007a). When a periodic sequence of flashes suddenly ends, some ganglion cells emit a large response. Interestingly, the latency of this response shifts with the period of the flash sequence, so that the ganglion cell responds to the omitted flash with a constant latency. This suggests that the retina forms predictions of observed patterns, and responds to a violation of its internal expectation. Several studies have proposed different mechanisms that may underlie this phenomenon, such as oscillatory activity in bipolar cells (Gao et al., 2009) or summation of parallel pathways with different response polarities (Werner, Cook, and Passaglia, 2008) and kinetics (Tanaka et al., 2019). However, none of these mechanism could be experimentally proven or fail to accurately predict the constant latency across stimulus frequencies. Thus, the mechanisms by which the retina achieves this phenomenon remain unclear and debated.

Here we investigated how inhibitory amacrine cells affect the OSR and showed that depression in inhibitory synapses can account for this characteristic latency shift. To this end, we performed electrophysiological recordings of retinal ganglion cells and found that blocking inhibitory transmission from glycinergic amacrine cells selectively abolished the predictive latency shift of the OSR. To better understand how glycinergic inhibition impacts the latency of the OSR, we developed a circuit model equipped with a glycinergic amacrine cell. This model could reproduce the latency shift of the OSR when the glycinergic synapse showed short-term depression, thereby adjusting its weight to the stimulus frequency. Our model generated several predictions about the OSR, which we could confirm in experiments. The latency shift that is characteristic of the OSR is thus due to a depressing inhibitory synapse whose weight is changed by the stimulus frequency. For low frequency sequence, the synaptic weight is large and this increases the latency of the response, while for high frequency stimuli, the weight is low due to depression, and the latency is only shifted by a small amount. Our results suggest a generic circuit to generate responses to surprise that could be implemented in several brain areas.

3.2 Results

3.2.1 ON biphasic ganglion cells exhibit an Omitted Stimulus Response to dark flashes

Using a multi-electrode array of 252 electrodes, we extracellularly recorded the spiking activity of ganglion cells from the mouse retina (Marre et al., 2012, Figure 3.1 A). We presented sequences of 12 full-field dark flashes of 40 ms duration each, at frequencies of 6, 8, 10, 12 and 16 Hz, with a grey baseline illumination. We estimated the receptive fields of the same retinal ganglion cells with a checkerboard-like noise. We defined ON cells based on their receptive fields (responding to a light increase, 3.1 B). We focused on the response of ON cells to sequences of dark flashes.

Many ON cells responded with a broad peak of activity after the stimulus stopped for all frequencies tested ($n = 74$ cells). Some cells showed a response to the onset of

the flash sequence and/or showed small responses to flashes for the 6Hz stimulus. However, the majority of cells did not respond at all to the flashed stimuli and we chose to focus on the omitted stimulus response peak only for further analysis.

A subset of cells with response after stimulus end shifted the response latency according to the stimulus frequency (36 %, n = 26, fig. 3.1 D, see Methods). They exhibit an "Omitted Stimulus Response" (OSR), as it was first described in Schwartz et al., 2007a (see Chapter 2 Figure 2.4): when the period of the flash train increases, the latency of the response to the last flash in the stimulus shifts by the same amount (fig. 3.1C,D), such that the latency of the response to the omitted stimulus is constant (fig. 3.1E).

This indicates that the retina has a precise temporal expectation of when the next flash should have occurred and shifts the latency of its response accordingly. This is illustrated in fig. 3.1F, where the relation between the latency of the response and the period of the flash sequence is linear, with a slope of nearly 1 (fig. 3.1G). Thus, a slope value of 1 indicates that the cell responds with a constant latency to the omitted stimulus, signalling the missing flash, while a slope value of 0 indicates a response with a constant latency relative to the last flash, signalling the offset of a response. A slope value of nearly 1 is thus a defining feature of the OSR. In the following we will refer to this specific relation as "latency shift".

Clustering of response to chirp stimuli did not yield a distinguishable cluster for cells exhibiting an OSR (Baden et al., 2016), it is thus likely that the OSR is exhibited in multiple retinal cell types. Interestingly, the ON cells showing this latency shift mostly had a biphasic response profile (Figure 3.1 B).

3.2.2 Amacrine cells are required for the latency shift in the Omitted Stimulus Response

It remains unclear how the retinal circuit generates the OSR. Previous studies have shown that the ON bipolar cell pathway is necessary to have a response *per se* (Schwartz et al., 2007a), but the components of the retinal circuit needed for the latency shift are yet to be determined. We hypothesized that the inhibitory cells of the retinal circuit are responsible for the shift in latency, as inhibition has been shown to shift latency in various circuits (Tsodyks, Pawelzik, and Markram, 1998; Wehr and Zador, 2003; Wehr and Zador, 2005; Fontaine, Peña, and Brette, 2014). Amacrine cells are the main class of inhibitory interneurons in the mouse retina. To investigate this further, we blocked glycinergic transmission using strychnine ($2\mu M$) and recorded the spike responses of retinal ganglion cells to flash trains of varying frequencies (see Methods). As glycinergic transmission is only employed by certain classes of inhibitory amacrine cells in the mouse retina (Wässle et al., 2009) this blocks only a subset of amacrine cells.

While the response after the sequence end remained after strychnine application, we observed that the slope between response latency and stimulus period decreased from an average of 1.13 ± 0.08 in the control condition to 0.07 ± 0.18 after strychnine was added (Figure 3.1 F and G, $n = 12$, mean $\pm$ SEM, see Methods). While the OSR occurred at roughly the same time in the highest frequency tested, the peak was significantly advanced after low frequency flashes compared to the control condition (Figure 3.1 D). As a consequence, the OSR did not have a constant latency relative to the omitted stimulus after strychnine was added (Figure 3.1 E). These results demonstrate that glycinergic amacrine cells are a key contributor to the OSR. Although they do not generate the response alone, they are crucial for the latency shift, which is a hallmark of the OSR.

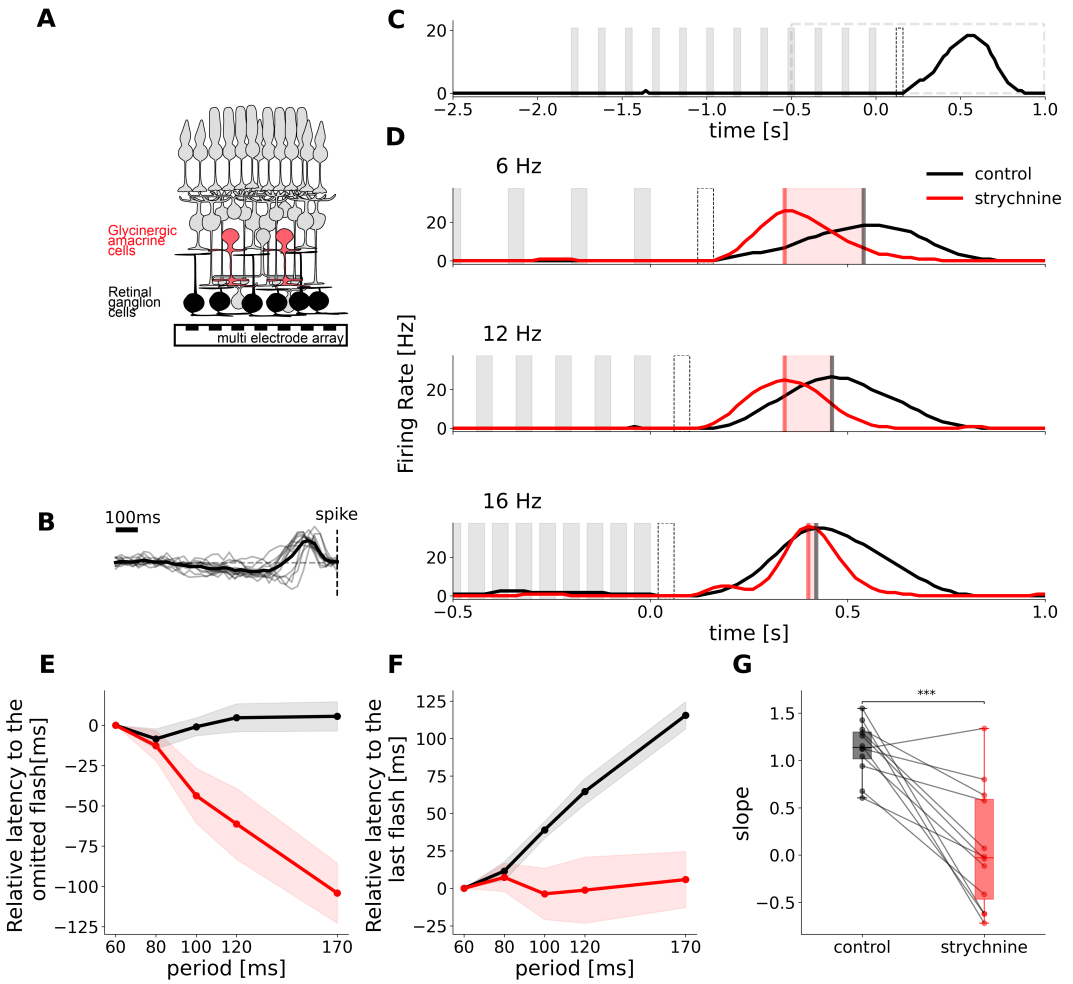

FIGURE 3.1: Glycinergic Amacrine cells are necessary for predictive timing of the OSR.

A. Schematic representation of the retina, with the activity of retinal ganglion cells being recorded with a multi electrode array. **B.** Temporal traces of the receptive fields of the cells that exhibit an OSR. **C.** Example of OSR. The cell responds to the end on the stimulation, after the last flash. The times of dark flashes are represented by grey shaded rectangles. The black dotted rectangle shows the timing of the omitted flash. The grey dotted rectangle shows the time period of focus in panel D. **D.** Experimental recording of the OSR in one cell in control condition (black) and with strychnine to block glycinergic amacrine cell transmission (red). Firing rate responses to flash trains of 3 different frequencies are aligned to the last flash of each sequence. Flashes are represented by gray patches, vertical lines indicate the maximum of the response peak, red shaded areas indicate the temporal discrepancy between control and strychnine conditions. **E.** Mean latency between OSR and the omitted flash plotted against the period of the stimulus for a population of $n = 12$ cells. Latency is expressed relatively to the latency of the response to the 16Hz stimulus in the control condition. **F.** Mean $\pm$ SEM latency between OSR and last flash in the stimulus plotted against stimulus period. Control latencies shift with the period of the stimulus wit a slope 1.13 ± 0.08. With strychnine this shift is abolished (slope = 0.07 ± 0.18). Latency is expressed relatively to the latency of the response to the 16Hz stimulus in the control condition. **G.** The slope of the latency shift decreases significantly when strychnine is added, p-value = 0.0001 (Welch t-test).

3.2.3 A circuit model with depressing synapses in inhibitory glycinergic amacrine cells explains the latency shift

Our experimental results provided compelling evidence that glycinergic amacrine cells provide inputs which are required to achieve the latency shift of the OSR. However, it remained unclear how the retinal circuit exhibits an OSR in a glycinergic dependent way. We developed a mechanistic model in which we explicitly simulated evoked inputs from glycinergic amacrine cells to understand their role in the latency shift of the OSR. Notably, bath application of strychine acts on the whole retinal network and likely has side-effects such as changes (increases) in the baseline activity across bipolar, other amacrine and ganglion cells due to a reduction of maintained inhibition. For simplicity, we do not take these side effects into account in our modeling approach and only simulate the minimal amount of direct inputs we found necessary for an OSR with latency shift.

Since we focused on biphasic ON ganglion cells, we equipped our model with two ON inputs, one being excitatory E^{ON} and mimicking ON bipolar cell input, and one being inhibitory I^{ON}, conveying broad delayed inhibition. This delayed inhibitory input summarizes the influence of various inhibitory pathways (horizontal cells, GABAergic amacrine cells, ON glycinergic amacrine cells) that generate the biphasic response profile (see discussion). In addition, we explicitly included a glycinergic amacrine cell with an OFF polarity, I_{Gly}^{OFF}, in order to provide inhibition to dark stimuli. All three units receive the visual stimulus as input, and connect onto a ganglion cell G (Fig. 3.2 A, see Methods for details).

A characteristic feature in the Omitted Stimulus Response is that the latency shifts by the same amount as the stimulus period. In other contexts, it has been shown that the relative strengths of excitation and inhibition can determine the latency of the response (Tsodyks, Pawelzik, and Markram, 1998; Wehr and Zador, 2003; Wehr and Zador, 2005; Fontaine, Peña, and Brette, 2014). We reasoned that the effective strength of the inhibitory input should thus depend on the frequency of the flash sequence. This can be achieved with a dynamic synapse, i.e. a synapse whose strength varies with the frequency of the stimulation.

Several previous reports have shown that inhibitory synapses can be depressing, i.e. have a decreasing weight depending on previous inputs (Li, Vigh, and Gersdorff, 2007; Vickers et al., 2012; Nikolaev et al., 2013; Kastner et al., 2019; Huang et al., 2022). It has been hypothesized that the variation in synaptic strength results from varying availability of vesicles in the readily releasable vesicle pool, which gets gradually depleted upon persistent inputs (Singer and Diamond, 2006; Burrone and Lagnado, 2000; Oesch and Diamond, 2011). In the retina, this has been modelled via dynamical systems of vesicle pools, where the value of one of the variables directly serves as input to the postsynaptic cell (Ozuysal and Baccus, 2012; Schröder et al., 2020). To keep our model as simple as possible, we modelled the glycinergic synapse as a depressing synapse using one kinetic equation to simulate synaptic vesicle occupancy, similar to cortical models of short-term plasticity (Tsodyks and Markram, 1997; Zhou et al., 2012; Hennig, 2013), but replaced spiking inputs by continous presynaptic voltage. The vesicle occupancy then scales the output of the cell (see Methods).

This mechanistic model allowed us to reproduce the main properties of the recorded ganglion cells. Thanks to excitatory and inhibitory ON inputs, it has an ON biphasic impulse response. It also responds with a peak at the end of dark periodic flash stimuli, due to delayed disinhibition, because inhibition has a slower temporal filter than excitation (Figure 3.2 B). Our model also successfully simulated an OSR whose

latency with respect to the last flash increased with the stimulus period with a slope around 1, and thus with a constant latency with respect to the omitted stimulus (Figure 3.2 B and C, black line).

We next simulated the experimental effect of strychnine with the model by removing the glycinergic amacrine cell input. Since the ON inhibitory cell of our model also includes the effect of ON glycinergic amacrine cells, we simultaneously decreased the weight of the ON inhibitory input (note however that our results did not depend on that additional modification, see discussion). Figure 3.2 B and C (red) show that the model replicates the experimental effect of strychnine. The slope of the latency shift decreases from 1.16 to 0.34 when I_{Gly}^{OFF} is removed from the circuit.

To test whether the model achieves the latency shift of the response peak thanks to the dynamical synapse, we simulated the response of the model while keeping the occupancy of the glycinergic amacrine cell synapse constant at 1 (Figure 3.3 A). The latency increased in all frequencies simulated, but the slope of the latency shift decreased to 0.32 (see figure 3.3 B-D). Without STP, our model is identical to a LN-model. The linear filters in the distinct pathways can be summed together to one ON biphasic filter profile, followed by the nonlinearity placed in the ganglion cell. This corresponds to an LN-model similar to what was proposed by Werner, Cook, and Passaglia, 2008, but is not able to predict the latency shift with a slope of 1. In our model, the dynamical synapse is thus essential to achieve the latency shift of the response with a slope of 1 observed experimentally. Note that there was also an overall increase in latency which can be explained by the fact that the fixed synapse is stronger than the dynamic counterpart, since it never has a full vesicle occupancy when stimulated.

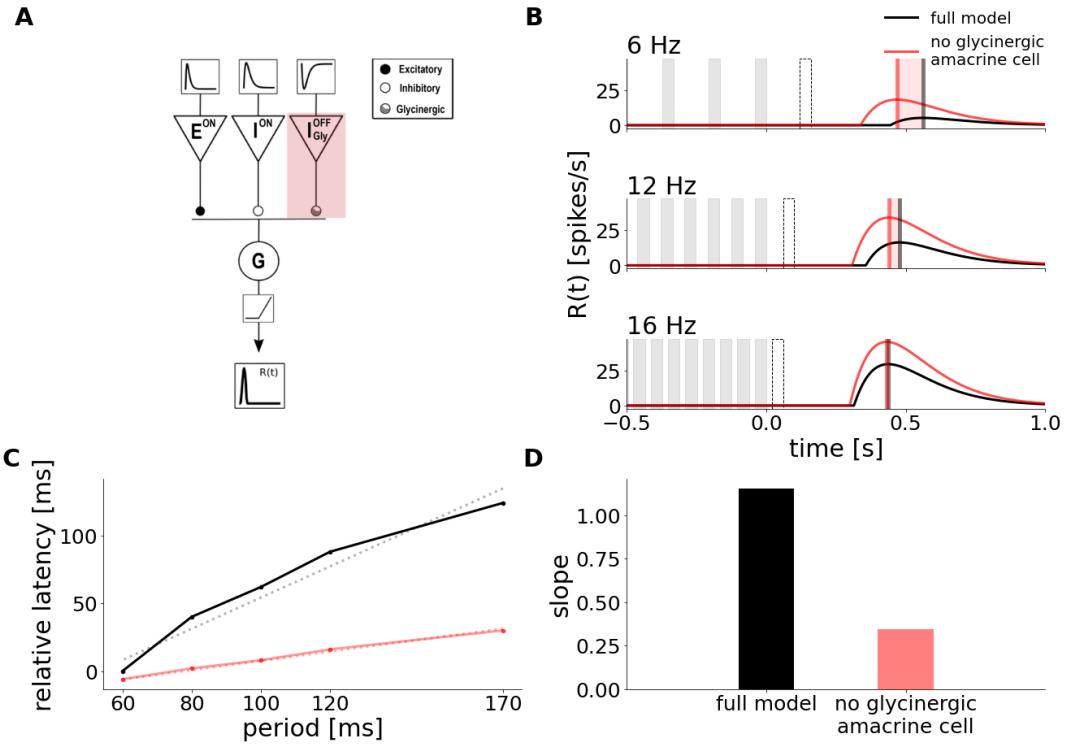

FIGURE 3.2: Mechanistic model replicates latency shift and strychnine experiment.

A. Schematic description of the Model. It is composed of an ON excitatory input E^{ON}, an ON inhibitory input I^{ON}, and an OFF inhibitory input I^{OFF}_{Gly} representing a glycinergic amacrine cell. Each of those units receives as input the convolution of the stimulus with a monophasic temporal kernel, determining the cells polarity, and connects onto a ganglion cell G. The response of G is then passed through a nonlinearity to simulate the cells' firing rate. The synapse from I^{OFF}_{Gly} to G can adapt its strength to the stimulus via short-term depression. The shaded red area represents the weight of this glycinergic amacrine cell being set to zero to simulate the effect of strychnine. **B.** Simulation of the model responses to flash trains of 3 different frequencies with the full model (black, control), and the weight of I^{OFF}_{Gly} set to 0 (red, strychine). The weight of I^{ON} was decreased from -70 to -30 Hz in this simulation, accounting for the broad effect of strychnine, which likely reduces inhibition overall. The times of dark flashes are represented by grey shaded rectangles, while the black dotted rectangle shows the timing of the omitted flash. **C.** Latency of the OSR plotted against stimulus period in control and strychnine simulation. The latency is expressed relatively to the latency of the response of the full model to the 16Hz stimulus. The slope of the latency shift decreased from 1.16 to 0.34 when I^{OFF}_{Gly} is set to 0 (dotted lines). **D.** Value of the slope fitted to latency shift in the full model and strychnine simulation.

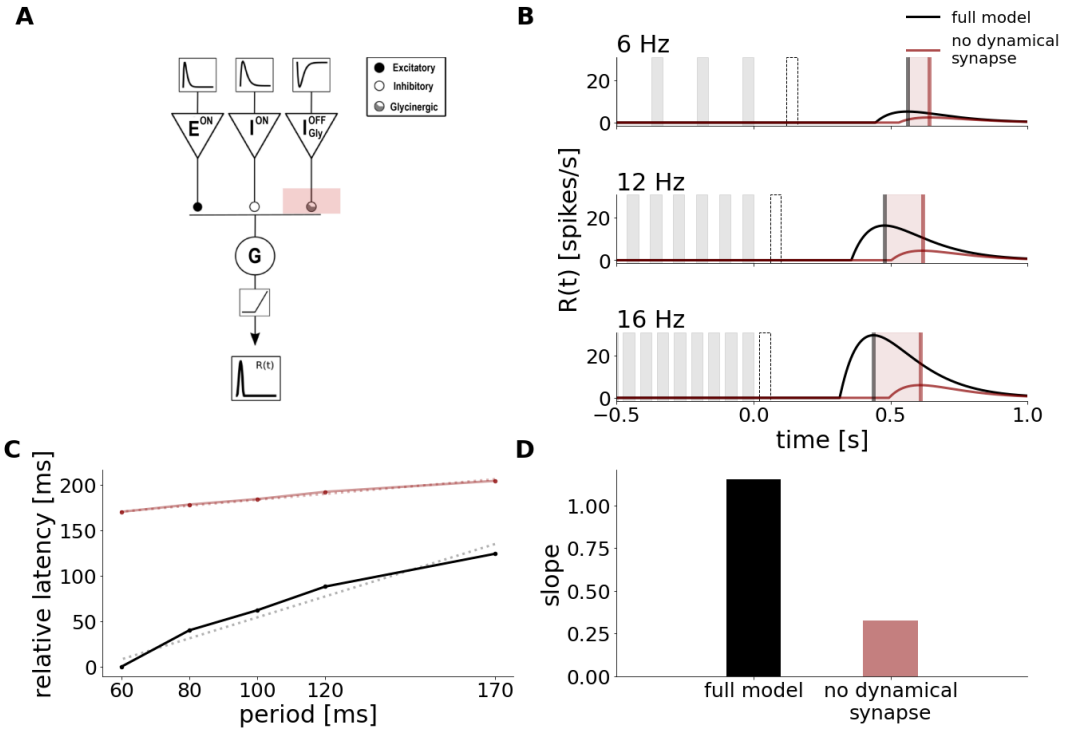

FIGURE 3.3: Short-term plasticity is the crucial component for latency shift.

A. Schematic description of the Model, as in Figure 3.2. The shaded red area now represents removing the dynamic characteristic of the glycinergic synapse. **B.** Simulation of the Model responses to flash trains of 3 different frequencies with dynamic occupancy (black) and the weight of I_{Gly}^{OFF} held constant (red). The times of dark flashes are represented by grey shaded rectangles, while the black dotted rectangle shows the timing of the omitted flash. **C.** Latency of the OSR plotted against stimulus period in control and strychnine simulation. The latency is expressed relatively to the latency of the response to the 16Hz stimulus in the control condition. The slope of the latency shift decreased from 1.16 to 0.32 when the weight of I_{Gly}^{OFF} was held constant (dotted lines). **D.** Value of the slope fitted to latency shift in the full model and without the adaptive property of the glycinergic synapse.

3.2.4 The depressing inhibitory synapse induces a latency shift

To understand how the model can account for the latency shift in the OSR, it is helpful to look at the temporal evolution of its internal variables. The ON excitatory input evokes a hyperpolarization in response to dark flashes, and cancels the depolarization evoked by the slightly slower ON inhibitory cells during the flash sequence. At the end of the flash sequence, due to differing time constants, there is a time window where depolarization due to the inhibitory delayed cell exceeds the hyperpolarization due to the bipolar cell, and triggers a spiking response (Figure 3.4 A). This is similar to many classical rebound responses recorded experimentally, and it can be predicted with a biphasic filter. But by itself, this biphasic filter would not predict the latency shift as observed experimentally. As we will describe in the following paragraph, this latency shift can be explained thanks to the specific effect of the glycinergic amacrine cell equipped with a depressing synapse.

Since this glycinergic amacrine cell is an OFF cell, it inhibits the ganglion cell in response to dark flashes. This has the effect of delaying the spiking response at the end of the flash sequence and to increase its latency (Figure 3.4 B, black compared to grey). The latency of the response is then shifted for different stimulus frequencies because the depressing synapse changes the strength of the glycinergic inhibition. In this synapse, the vesicle occupancy represents the amount of synaptic depression and decreases when stimulation starts (Figure 3.5 A, 3rd row). It then reduces the current input from I_{Gly}^{OFF} to the ganglion cell (Figure 3.5 A, 4th row). This reduction of inhibition shifts the OSR towards an earlier response, reducing its latency (Figure 3.5 A, 5th row).

Fast frequency stimuli cause stronger depression, reducing the I_{Gly}^{OFF} current input by about 30 %. This has a strong impact on the latency, which is more than 100 ms shorter when the synapse is depressed. In contrast, slow frequency stimuli cause only weak depression, reducing the A_{Gly} by about only 10 %. The latency was thus only slightly reduced in that case. (compare 3.5 A and B).

In summary, the steady state vesicle occupancy of the synapse is determined by the stimulus frequency (fig. 3.5 C and D). The vesicle occupancy can then reduce the inhibitory current, yielding a large reduction for fast inputs and a small reduction for slow inputs (3.5 E). As a result, vesicle occupancy acts like a scaling factor to tune the inhibitory current input and thereby shifts the response latency based on stimulus frequency This explains how the latency shift observed experimentally is achieved via glycinergic inputs.

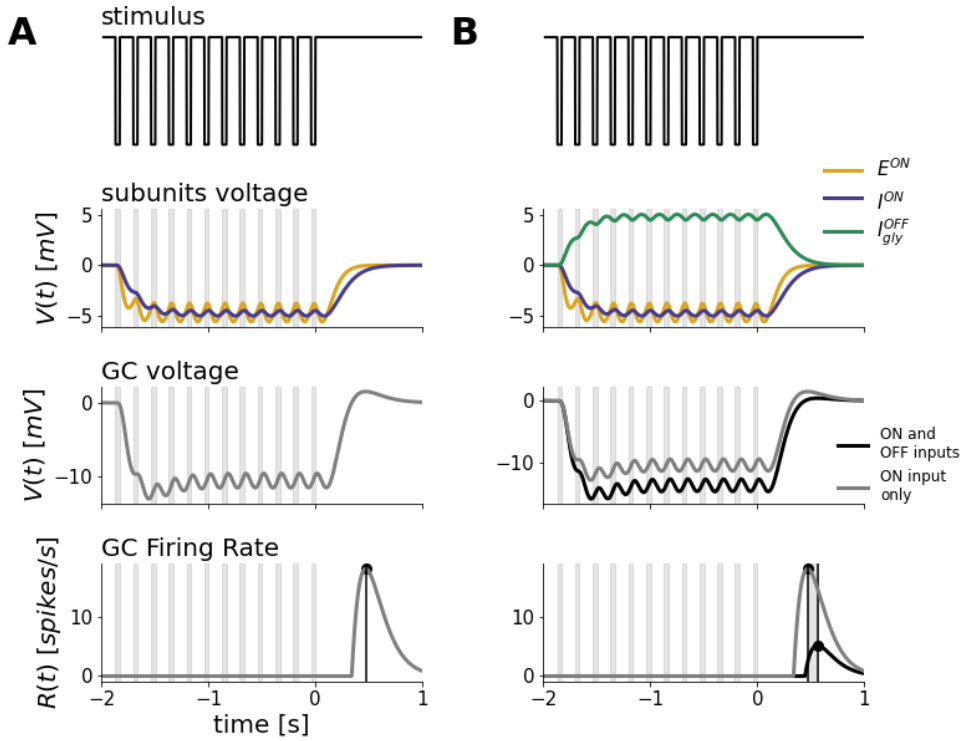

FIGURE 3.4: ON components of the model produce a peak after stimulus end while the glycinergic OFF input shifts the latency.

A. - B. Model responses at each processing step to a 6 Hz dark flash stimulus. From top to bottom: Stimulus Intensity, Bipolar and Amacrine voltage responses, ganglion cell voltage and firing rate. The glycinergic OFF synapse is dynamic. **A.** ON excitation and inhibition hyperpolarize in response to dark flashes. When both inputs are substracted in the ganglion cell, its voltage sum hyperpolarizes during flash presentation, followed by an overshoot of disinhibition due to the slower response profile of the inhibitory input. After passing the voltage through a rectification function, only the disinhibitory peak after stimulus end remains in the firing rate. **B.** Same as **A** but with additional OFF inhibition. The voltage of the glycinergic OFF input depolarizes in response to dark flashes, passing additional inhibition onto the ganglion cell. This lowers the GC voltage response and increases the latency between peak and stimulus end in the firing rate. Last two panels compare the models' simulation and peak time-point with (black) and without (grey) the input from I^{OFF}_{Gly}.

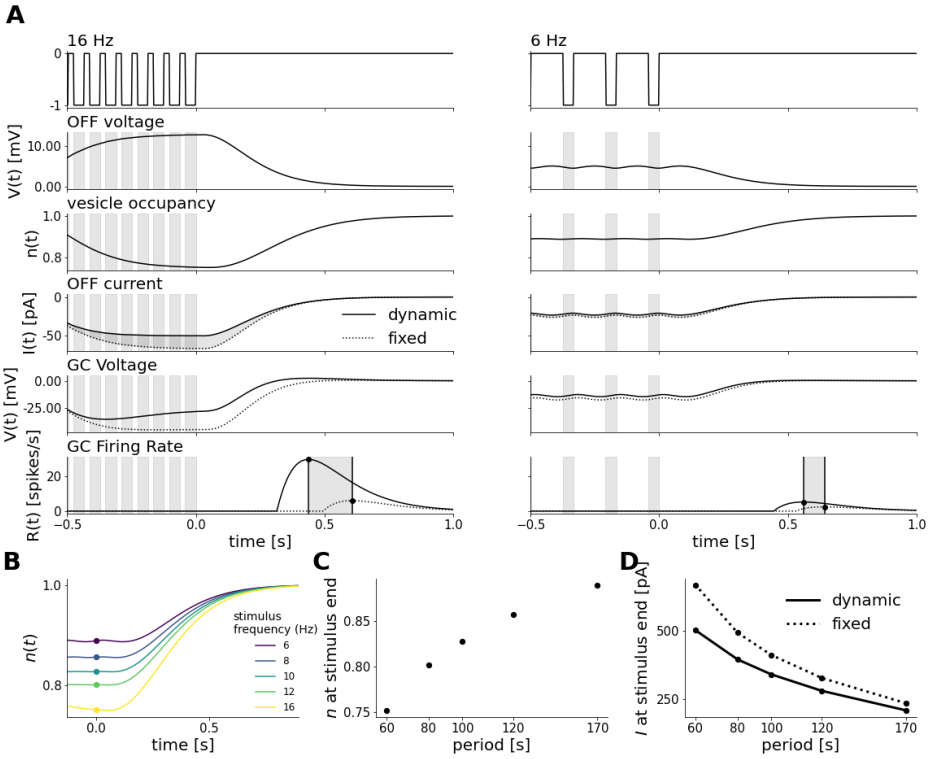

FIGURE 3.5: Synaptic depression scales OFF glycinergic input to stimulus frequency and thereby shifts the latency of the response.

A. Impact of occupancy scaling on I_{Gly}^{OFF} current input to G for a fast (16 Hz, left) and a slow (6 Hz, right) stimulus. From top to bottom: Stimulus Intensity, I_{Gly}^{OFF} voltage, vesicle occupancy, current input, G voltage and firing rate. Last 3 panels compare simulations with dynamic occupancy (solid lines) to when the occupancy is held constant (dotted lines). Depression has the effect to advance the OSR peak, more so for fast than slow frequencies. Traces are aligned to the time-point of the last flash, only flashes within 0.5 s betofre stimulus end are shown. **B.** Occupancy traces to flash stimuli of different frequencies aligned to the last flash. Dots indicate the occupancy level at stimulus end. **C.** Level of occupancy after stimulus end scales with the period of the stimulus. **D.** I_{Gly}^{OFF} current input is decreased by short-term depression, more so for fast than slow frequencies. Dotted line shows current with fixed occupancy, solid line with dynamic occupancy.

3.2.5 The depressing inhibitory synapse predicts other features of the Omitted Stimulus Response

Can our model give other predictions about the OSR ? Since the glycinergic inhibitory synapse is more depressed at high frequency, the OSR is less inhibited, and its amplitude is thus stronger compared to low frequencies (compare Figure 3.5 A and B, 5th row). This trend was also observed in our experiments. Both simulated and experimental amplitudes showed a negative correlation between the response amplitude and the stimulus period (-0.87 ± 0.02, mean$\pm$ SEM $n = 14$) in data and as well -0.87 in simulations).

An important consequence of our depressing synapse model is that it takes several flashes to reach the steady state in the vesicle occupancy. If we shorten the flash sequence, the vesicle occupancy will not reach that steady state, and this should have predictable consequences on response amplitude and latency. We simulated the response to long flash trains consisting of 12 flashes (as in the experiments and simulations above) and shorter sequences of only 5 flashes.

Our simulations predicted that the amplitude of the OSR decreases when the stimulus contains only 5 flashes. We tested that in experiments and found the same tendency: the OSR amplitude was significantly smaller in all but the lowest frequency tested (see Figure 3.6 A)

Another model prediction was that the slope of the relation between OSR latency and stimulus period should decrease for shorter flash trains, reaching only a value of 0.67 for 5 flashes, compared to 1.16 when 12 flashes were presented (see Figure 3.6 B, left) In our model, this is a consequence of the dynamics of the depressing synapse.

In the 5-flashes scenario, our model predicts that the stimulus is too short for the synapse to reach a steady state occupancy when the stimulus frequency is high (Figure 3.6 C). I_{Gly}^{OFF} hence provides a larger inhibitory input than for a longer sequence. This increases the response latency and changes the slope of the relation between latency shift and stimulus period (Figure 3.6 D).

In our experiments, while there was no difference for low frequency stimuli, the absolute latency of the OSR was much larger after 5 flashes than after 12 when the stimulus frequency was high (see Figure 3.6 B). This change in latency led to a reduction of the mean $\pm$ SEM slope value from 0.84 ± 0.02 to 0.69 ± 0.04 in experiments, consistent with the model prediction. These agreements provide further evidence for the validity of our model, and for the key role of a depressing inhibitory synapse.

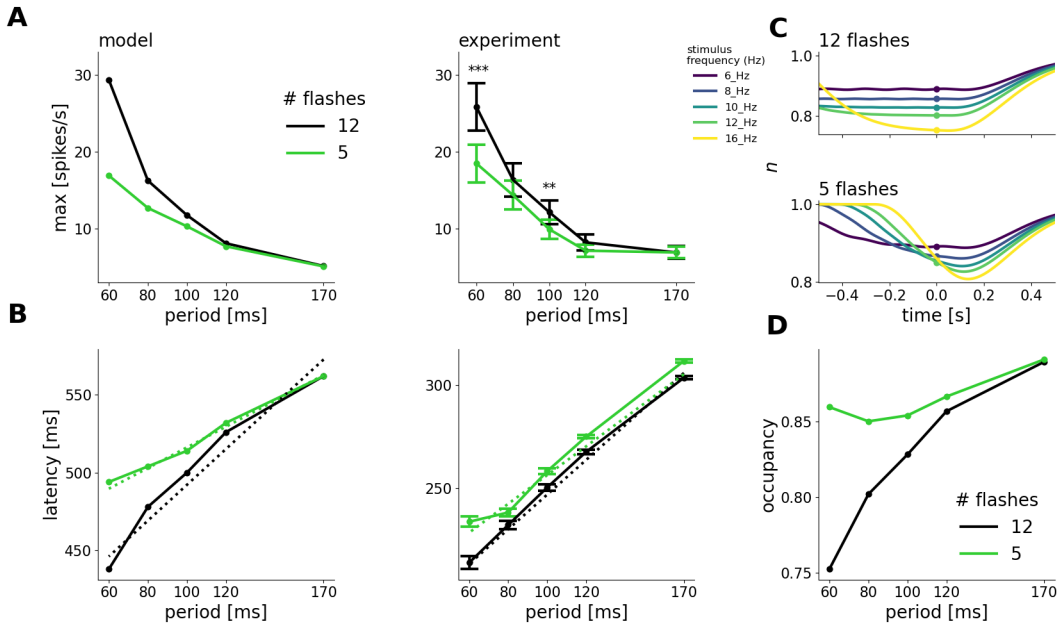

FIGURE 3.6: Latency shift decreases for shorter stimuli because of lacking steady state occupancy.

A. Amplitude of the OSR against stimulus period for 12 and 5 flashes in the stimulus in simulations (left) and experiments (right). Amplitudes to 6 Hz and 10 Hz stimuli were significantly different after Bonferroni-Holm correction Holm, 1979 (6 Hz: $p = 0.000006$, 10 Hz: $p = 0.001$). **B.** Latency against stimulus period in simulations (left) and experiments (right). Simulated slopes decreased from 1.16 after 12 flashes to 0.67 after 5 flashes. Experimental slopes decreased from 0.84 ± 0.02 to 0.69 ± 0.04, mean$\pm$ SEM. The latency after 12 and 5 flashes was significantly different after Bonferroni-Holm correction Holm, 1979 for 10 Hz ($p = 0.03$), 12 Hz ($p = 0.04$) and 16 Hz stimuli ($p = 0.02$). **C.** Temporal traces of vesicle occupancy to all frequencies simulated, for 12 flashes (upper) and 5 flashes (lower). Dots indicate occupancy at stimulus end. Traces do not reach a steady state for 5 flashes. **D.** Scaling of occupancy with stimulus period in 5- and 12-flashes scenario.

3.2.6 Peak timing carries predictive information but not response onset or amplitude

Next, we looked at further features of the OSR response to examine whether they contain relevant information about stimulus frequency, and how they might be impacted by glycinergic inhibition. Therefore, we analyzed the peak amplitude and the latency of the response onset of the OSR to different frequencies, and compared experimental findings to the predictions of our model.

A general trend we observed is that the OSR amplitude decreased when stimulus frequency increased (Figure 3.7 A), both in experimental data and in our model simulations. The experimental amplitude decrease seems to be slightly flattened with strychnine, however there are no significant differences in peak amplitudes between control and strychnine conditions. In our model, the amplitude frequency relation is flattened in the strychnine simulation as well, but our model predicts overall higher amplitudes with strychnine than in control, which we did not find in experiments. This discrepancy might be due to the fact that we are only simulating direct effects of glycinergic inhibition, such that when this aspect is removed, the response amplitudes are higher. The fact that the overall amplitude seems to be maintained in the retinal network may come from indirect effects of strychnine inhibition, such as baseline changes or disinhibition of other amacrine cells, which we do not include

in our model.

We next examined the relation between spiking onset and stimulus frequency. Our model predicts that spiking onset scales with frequency in a similar manner as the peak response, with a slope of 1 in control which decreases drastically when dynamic glycinergic inhibition is removed (Figure 3.7 B, right). Experimental results reveal a much lower slope of 0.39 in the control spiking onset, which on average decreases to 0.15 when glycinergic inhibition is blocked. However, the effect of strychnine on spiking onset in individual cells was highly variable, resulting in a non-significant difference in the latency slope. Overall, these results suggest that spiking onset does not contain predictive information about the arrival of the next flash as its timing does not shift in a 1-to-1 manner with the period of the flash train.

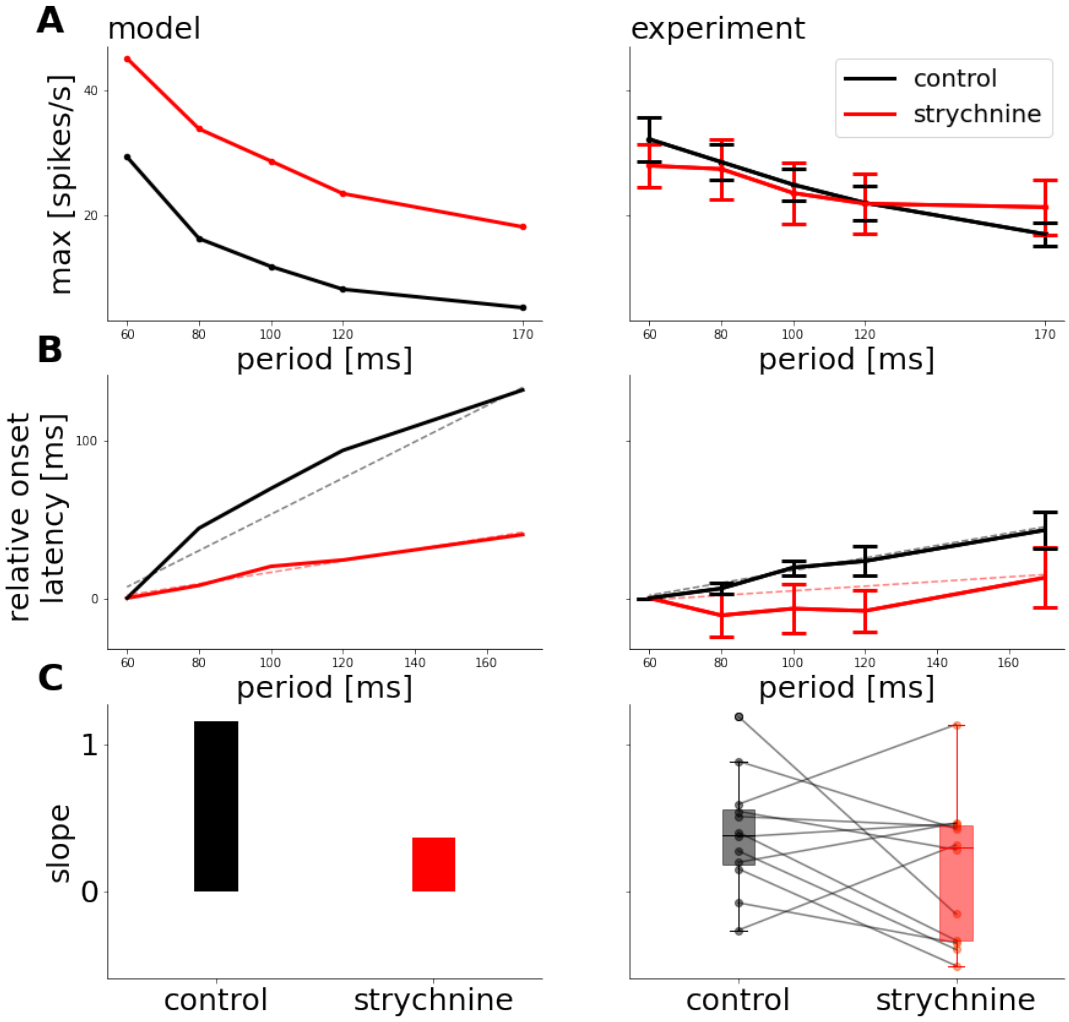

FIGURE 3.7: Peak amplitude and onset latency to different stimulus frequencies are not systematically impacted by strychnine

A. Amplitude of the OSR against stimulus period for control and strychnine conditions in simulations (left) and experiments (right). Experimentally measured amplitudes are not significantly different after Bonferroni-Holm correction (Holm, 1979) (6 H $p = 0.26$, 16 Hz: $p = 0.975$). **B.** Response onset latency against stimulus period in simulations (left) and experiments (right). In simulations, the control latencies scale with stimulus frequencies with a slope of 1.15, but not in strychnine simulations (slope = 0.36). In experiments, there are no significant differences between latencies in control and strychnine conditions, (6 H $p = 0.657$, 16 Hz: $p = 0.262$), and both control and strychnine conditions scale with frequency with a low slope (control = 0.39 ± 0.11, strychnine = 0.15 ± 0.36, mean$\pm$ SEM).
C. Quantification of onset latency slopes in simulation (right) and experiments (left) in panel **B**. The slope of the onset latency shift between control an strychnine is not significantly different in experiments after Bonferroni-Holm correction (Holm, 1979) (p = 0.186).

3.3 Discussion

The Omitted Stimulus Response is an example of sophisticated feature detection that takes place already in the retina. This phenomenon implies that retinal ganglion cells can carry a dynamic prediction of their future visual input with high temporal precision, and selectively respond when this prediction is not matched. Although high-contrast full field periodic flashes are artificial stimuli which are unlikely to

occur in natural scenes, they isolate temporal aspects of visual input patterns. The underlying mechanisms of temporal processing in the rather artificial OSR may be embedded in more complex networks detecting spatio-temporal patterns in more realistic scenes.

With this work, we provide evidence that the latency shift of the OSR, which allows a constant latency relative to the omitted stimulus, is generated by inhibition from glycinergic amacrine cells. Using computational modelling, we show how inhibition can enable retinal ganglion cells to respond to the missing flash at the end of a sequence. Short-term depression in inhibitory synapses allows shifting the latency of this response.

3.3.1 Short-term plasticity as a novel mechanism in the OSR

In contrast to those previous models of the OSR, we explicitly included an inhibitory input whose contribution to the peak latency is dependent on the stimulus frequency via short-term plasticity. By doing so, we can propose a mechanistic explanation and match the latency shift of the OSR as well as various other response properties of the experimentally observed OSR.

We will briefly discuss how our findings relate to previous models of the OSR. To recap, Werner, Cook, and Passaglia, 2008 proposed a dual LN-model with biphasic ON-OFF pathway interactions, which accurately captures the response peak after stimulus end via the rebound but fails to shift the peak latency as a function of the stimulus period with a slope of 1. When removing the depressing synapse, our model is amenable to a biphasic LN model, since the responses of our intermediate units are then linear and could be represented by a single linear filter. Our model then simulates the OSR in the same manner as this previous study. Thus, adding but the depressing inhibitory synapse was necessary to obtain the slope of 1, which is the signature of a predictive latency shift.

Gao and Berry (Gao et al., 2009) proposed intrinsic oscillatory activity in ON bipolar cells that evoked a latency shift via resonance tuned to the stimulus frequency. However, such oscillatory activity was not found in bipolar cells (Deshmukh and Berry, 2019). Further, our experiments show that glycinergic amacrine inhibition is necessary for the latency scaling of the OSR, which cannot be explained by this model.

Tanaka et al. (Tanaka et al., 2019) proposed that the OSR with its latency shift can arise in a deep neural network model via summation of multiple excitatory inputs with different time constants. It is difficult to evaluate whether the model accurately captures the latency shift as observed in experiments, with the correct slope value. The explanation behind this model is that the OSR latency is determined by the sum of 2 ON bipolar cells which are activated only by certain stimulus frequencies due to different temporal filtering. This purely excitatory mechanism of latency scaling is not in line with our experimental findings, suggesting that amacrine cells likely contribute to temporal filtering as well. Our hypothesis is thus that the components they isolated correspond to a mix of bipolar and amacrine cell properties.

3.3.2 Limitations of the study

Previous experimental studies (Schwartz et al., 2007a; Schwartz and Berry, 2008; Werner, Cook, and Passaglia, 2008) reported that the OSR is found in a higher proportion of retinal ganglion cells than what we observed in this study. This difference

could come from the fact that we define OSR as a response with a latency shift having a slope of at least 0.7, while it is not clear whether previous studies took multiple frequencies into account when classifying the OSR. Schwartz et al. (Schwartz and Berry, 2008) also showed that blocking inhibition from amacrine cells had no effect on the OSR. But again, this study only investigated the presence or absence of the OSR under amacrine blockade, and did not investigate if the latency of the OSR shifted with the stimulus frequency. In addition, previous studies were mostly carried out in salamander, where the underlying mechanisms may be different from the mouse.

In order to realistically simulate the model's response with glycinergic amacrine cells blocked, we had to decrease the weight of the inhibitory ON input I_{ON} in our simulations. Leaving the weight of this input untouched while setting $w_{I_{Gly}^{OFF}}$ to 0, we still obtain a decrease in latency shift but this configuration generated a response to each flash of the sequence, something we did not observe experimentally. We therefore deemed this configuration as less realistic than decreasing also the weight of the ON inhibitory cell, since strychnine is likely to affect glycinergic ON inhibition as well.

While our model predicts that the spike onset timing also scales with the period of the stimulus, this was not observed experimentally. Out model was specifically designed to implement a potential mechanism for the latency scaling of the peak response and contains only the minimal amount of components necessary for that. It seems likely that retinal ganglion cells receive inputs from more bipolar cells than the ones we included, some of which respond earlier and do not scale with frequency, specifically affecting spiking onset. Alternatively, dynamic glycinergic inhibition may act in a divisive instead of a substractive manner, which would impact peak timing much more than spiking onset.

Additionally, we are only simulating direct effects of glycinergic inhibition. It is certainly possible that bipolar cell processing is indirectly affected by glycine, eg. via a change in baseline membrane potential or by altered release dynamics.

3.3.3 Short-term plasticity is not evident in RGC spike trains

Dynamical synapses have previously been proposed to enable neuronal circuits in the retina to form expectations of future inputs (Hosoya, Baccus, and Meister, 2005; Johnston et al., 2019) and are thus a plausible candidate to play an important role in the OSR. Previous works have shown that inhibitory synapses can be depressing (Kastner and Baccus, 2011; Kastner and Baccus, 2013b; Nikolaev et al., 2013). In particular, glycinergic synapses that input to bipolar cells can be depressing (Huang et al., 2022).

However, we could not show experimentally that the depressing nature of inhibitory synapses is necessary for the latency shift in the OSR. Adaptation in excitatory retinal synapses is a well studied phenomenon to reliably encode luminance and contrast. Even though it has been shown that the OSR remains also when the overall luminance is held constant during the flash stimulus, it remains unclear how the latency scaling is affected and would need to be determined in future experiments.

Nevertheless, our model predicts that the depressing inhibitory synapse should have several functional consequences, that we verified in the data. In particular, a key prediction of the depressing synapse is that the OSR requires a long enough flash sequence to accurately shift the latency, which coincides with the time needed

to reach a steady state in the synaptic weights, and we confirmed this prediction experimentally.

Overall, we chose to only include the minimal components necessary to specifically explain the peak latency shift in the OSR and its abolishment via strychnine and arrived at a model with 15 parameter. For example, we chose to simulate synaptic depression via a modified version of cortical STP-models with only 2 parameter rather than the more complex systems used in the retina previously (Schröder et al., 2020). This model is surely too simple to represent the full retinal circuitry, which is why is is not surprising that it cannot capture the full response spectrum observed experimentally.

3.3.4 Relevance for cortical processing

Ultimately, our results might be of relevance to understand neuronal mechanisms of predictive coding beyond the retina. Very similar surprise responses exist in other sensory domains, such as the mismatch negativity response in the auditory cortex (Näätänen et al., 1993; Garrido et al., 2009; Ulanovsky, Las, and Nelken, 2003; Li et al., 2017), where neural activity is enhanced following a 'deviant' tone in a sequence of 'standard' tones. A recent study suggested that synaptic adaptation could be a key contributor to this phenomenon (Amsalem et al., 2020). Following the predictive coding theory, one possible explanation is that this response emerges from an interaction between feed-forward and feedback connectivity (Rao and Ballard, 1999; Millidge, Seth, and Buckley, 2021). Here we show that a purely feed-forward micro-circuit can generate this response to a violation of prediction via an interplay of excitation and inhibition, where synaptic depression takes place in inhibitory connections. All the components used in this micro-circuit are generic and can be found in other sensory areas (Denève and Machens, 2016; Tsodyks and Markram, 1997; Abbott et al., 1997), and it is thus likely that a similar circuit could be at work at the cortical level, for more complex pattern recognition than full-field flashes.

3.4 Methods

3.4.1 Experimental Setup

Recordings

Recordings were performed on C57BL6/J adult mice of either sex. Animals were killed according to institutional animal care standards. The retina was isolated from the eye under dim illumination and transferred as quickly as possible into oxygenated Ames' medium (Merck, A1420). The retina was extracted from the eye cup and lowered with the ganglion cell side against a multi-electrode array whose electrodes were spaced by 30μm, as previously described elsewhere Marre et al., 2012. During the recordings, the Ames' medium temperature was maintained at 37°C. Raw voltage traces were digitized and stored for off-line analysis using a 252-channel preamplifier (MultiChannel Systems, Germany) at a sampling frequency of 20kHz. The activity of single neurons was obtained using Spyking Circus, a custom spike sorting software developed specifically for these arrays Yger et al., 2018.

Visual stimulation

Visual stimuli were presented using a white LED and a Digital Mirror Device (DMD). Flash sequences contained 5 or 12 flashes of 5 different frequencies (6Hz, 8Hz, 10Hz,

12Hz, 16Hz). Polarities were either switched from grey to black (dark flashes) or from grey to white (bright flashes). 60 trials were conducted for each stimulus, with 2-4 s between each trial. The order of magnitude of the background illumination was 10^6 R*.

Spike Triggered Average

We displayed a random binary checkerboard during 40 min to 1h at 40 Hz to map the receptive fields of ganglion cells. A three dimensional STA (x, y and time) was sampled averaging over the stimulus preceding each spikefor a time window of 1 s, divided into N = 40 time bins. Temporal and spatial components were isolated via Singular Value Decomposition and reconstructed form eigenvectors. STA analysis was carried out in Python.

Pharmacology

To block glycinergic transmission, we dissolved strychnine (Sigma-Aldrich, S8753) in Ames' medium at a concentration of $2\mu M$, and perfused the retina with the solution at least 15 minutes before the recording.

Latency Analysis

To determine slope of latency shift, we measured the latency between the peak firing rate and the end of the last flash in the stimulus for all frequencies tested. We plotted these latencies against the respective period of the stimulus and fitted a straight line to determine the slope of the latency shift using SciPy's optimization package (Virtanen et al., 2020). Cells where classified as having an OSR in the control condition when the slope what at least 0.7 or higher. All cells where the peak time point could not be unambiguously determined in any condition were excluded from the analysis.

Statistical Analysis

Statistical testing was performed using Scipy's stats package (Virtanen et al., 2020) for t-test for independent samples For Bonferroni- Holm correction, significance levels α were adjusted to $\frac{\alpha}{m+1-k}$ for the k^{th} p-value of m total comparisons.

3.4.2 Modeling

Model Implementation

The 3 pathways of Fig. 3.2 receive an input from the Outer Plexiform Layer (OPL) written as a temporal convolution of the OSR stimulus, $s(t)$ with a linear filter of the form:

$$\alpha_X(t) = \frac{t}{\tau_X^2} exp(-\frac{t}{\tau_X}) H(t), \quad X = E^{ON}, I^{ON}, I^{OFF}_{Gly}, \tag{3.1}$$

where τ_X is cell X characteristic time of integration (in s) and $H(t)$ the Heaviside function.

Thus, the inputs read:

$$F_X(t) = S_X[\alpha_X * s](t) \tag{3.2}$$

where S_X is a scale factor and $*$ the space-time convolution. If S_X is negative, X is an OFF cell. Note that, the stimulus being spatially uniform, the space integration

reduces to a constant, so that the detailed shape of the spatial RF plays a trivial role. The OPL response is then integrated into all pathways via a linear dynamical system:

$$\frac{dV_X}{dt} = -\frac{V_X}{\tau_X} + F_X(t), \quad X = E^{ON}, I^{ON}, I^{OFF}_{Gly}. \tag{3.3}$$

where V_X is the voltage of cell X (in Volt).

Next, all pathways provide input to the ganglion cell G:

$$\frac{dV_G}{dt} = -\frac{V_G}{\tau_G} + w_{E^{ON}} V_{E^{ON}} + n w_{I^{OFF}_{Gly}} p(V_{I^{OFF}_{Gly}}, \theta_{I^{OFF}_{Gly}}) + w_{I^{ON}} V_{I^{ON}}. \tag{3.4}$$

where $w_{E^{ON}}, w_{I^{OFF}_{Gly}}, w_{I^{ON}}$ are synaptic weights (in Hz). Voltages are rectified before integrated in the ganglion cell membrane potential via :

$$p(V, \theta) = \begin{cases} V - \theta & \text{if } V \geq \theta : \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

where θ is a threshold (in Volts).

The synaptic weight from I^{OFF}_{Gly} to G is modulated by a dimensionless variable n, used to simulate synaptic short-term plasticity. n, which interprets as a vesicle occupancy in the glycinergic amacrine synapse, obeys the kinetic equation (Hennig, 2013) :

$$\frac{dn}{dt} = (1 - n) k_{rec} - \beta k_{rel} p(V_{I^{OFF}_{Gly}}, \theta_{I^{OFF}_{gly}}) n. \tag{3.6}$$

k_{rec} and k_{rel} are rate constants (Hz) for vesicle release and replenishment and β (V^{-1}) is a scaling factor. Finally, the voltage response is passed through the piece-wise linear function p to obtain the firing rate

$$R(t) = s_G \, p(V_G(t), \theta_G), \tag{3.7}$$

where s_G is a scaling factor. The model was implemented with custom Python code.

Parameter Optimization

First, time constants and weights were fitted such that the linear response V_G of the model to an impulse stimulus best resembles the temporal STA of one example cell with OSR. this was done using an evolutionary optimization algorithm using the CMAES (Covariance Matrix Adaptation Evolutionary Strategy) Python toolbox (Hansen, Akimoto, and Baudis, 2019). Therefore, synaptic plasticity is removed from the model by setting $\beta = 0$. Scale factors were chosen such that the amplitude of voltage responses of intermediate units do not depend on the respective time constants setting $S_X = \frac{1}{\tau_X}$. The slope of the ganglion cell non-linearity S_G was chosen as to match experimentally observed firing rates. Parameters of the occupancy equation (3.6) were chosen as to yield a slope of 1, values were obtained by screening over a wide range of combinations of β and the ratio of recovery and release rate $\frac{k_{rec}}{k_{rel}}$. All parameter used in the simulations are listed in Table 3.1.

Parameter Values

Parameter	Value	Unit
$\tau_{E^{ON}}$	0.05	s
$\tau_{I^{ON}}$	0.08	s
$\tau_{I^{OFF}_{Gly}}$	0.08	s
τ_G	0.1	s
$w_{E^{ON}}$	50.0	Hz
$w_{I^{ON}}$	-95.0	Hz
$w_{I^{OFF}_{Gly}}$	-82.0	Hz
$S_{E^{ON}}$	1.0	Vs^{-1}
$S_{I^{OFF}_{Gly}}$	-0.625	Vs^{-1}
$S_{I^{ON}}$	0.625	Vs^{-1}
$\theta_{I^{OFF}_{Gly}}$	0.0	V
k_{rel}	4.5	Hz
k_{rec}	1.0	Hz
β	13.6	V^{-1}
θ_G	0.0	V
s_G	2200	HzV^{-1}

TABLE 3.1: Model parameter values used in simulations

Chapter 4

Temporal refinement of spatiotemporal pattern prediction via short-term plasticity

In this chapter we explore how dynamical synapses affect temporal predictions formed on the basis of more complex stimuli that also contain a spatial dimension. We therefore look at the phenomenon of predictive motion encoding in the retina, where retinal ganglion cells anticipate the future position of a moving bar (a spatiotemporal pattern). First, we show how a retinal network model with lateral amacrine connectivity can anticipate a moving object. Building up on the hypothesis that dynamical synapses in inhibitory connections implement temporal expectations of visual inputs, which we elaborated in Chapter 3, we then conduct a computational study on the effect of short-term depression on motion anticipation to various different stimulus speeds. Finally, we investigate how this model behaves in response to surprise in spatiotemporal patterns.

Contents

4.1 Introduction

In order to respond to a visual environment in real-time, the brain must compensate for delays in the neuronal responses. To this end, retinal cells have been shown to predict the trajectory of a moving object (Berry et al., 1999): When a bar is moving across the receptive field of a retinal ganglion cell, the peak-firing rate of the cell occurs earlier as if the bar is flashed above the receptive field center. This phenomenon has been coined "motion anticipation" and implies that already in the inner retina, cells form a prediction of the future position of a moving object (see Chapter 2.3.2). This predictive capacity is generally thought to be implemented via negative feedback suppression, such as a reduction in response gain.

In order to anticipate a moving object at the right position at the right time, the network should have a temporal expectation of the objects' speed. Berry et al., 1999 showed that between speeds of 0.1 - 1 mm/s, ganglion cells maintain a stable spatial anticipation with respect to the bar position but begin to show a lag in their responses at speeds faster than 1 mm/s. Several studies showed that inhibition via amacrine cells plays a role in motion anticipation (Johnston and Lagnado, 2015; Lee and Menz, 2020; Souihel and Cessac, 2021). However, it is not clear to which extent they can account for constant speed scaling for a relevant range.

In this study, we isolate the network effect via lateral amacrine connectivity on motion anticipation as first proposed by Souihel and Cessac, 2021 in combination with gain control. We show that this network alone can induce motion anticipation at bipolar cell outputs, supporting that lateral amacrine connectivity could be an underlying mechanism for gain control in bipolar cell responses. We then test how anticipation via network inhibition is affected by the bar speed and explore how synaptic plasticity in connectivity weights might aid the network to have a better speed representation. Additionally, we find that under certain conditions network inhibition can prime ganglion cells for early onset responses in our model, which has been observed experimentally but cannot be explained by gain control. Finally, we test how the lateral connectivity model responds to surprise in motion stimuli, such as onset and reversal motion.

4.2 Results

4.2.1 A retinal network model with reciprocal amacrine connectivity for motion anticipation

To study how lateral amacrine connectivity impacts motion anticipation, we designed a 1-D network model consisting of 3 consecutive steps to simulate the processing of an incoming stimulus by 3 retinal layers (Figure 4.1). Individual bipolar cells (BCs), amacrine cells (ACs) and retinal ganglion cells (RGCs) are simulated as point neurons and characterized by their voltage response $V(t)$, which is transformed into a synaptic response $R(t)$ after rectification (equation 4.7).

To simulate retinal processing, the spatiotemporal stimulus $s(x,t)$ is first convolved with a spatial and a temporal kernel to simulate how the OPL transforms a visual scene into a voltage response in bipolar cells. This response, $V_{drive}(t)$ (eq. 4.2), is purely evoked by the stimulus. The convolutional kernel (equation 4.3) describes the receptive field (RF) of the respective BC. The spatial profile is a Difference-of-Gaussian filter corresponding to the receptive field of bipolar cell i centered ad its position x_i, where the positive center comes from photoreceptor inputs while the

negative surround comes from horizontal cell inhibition. The temporal kernel sums up transmission delays in this first retinal processing step (see 2.4, Figure 2.8).

The second layer of the model simulates BC and AC interactions and is simulated as a dynamical system where the voltage response of each cell type has a characteristic time constant τ_B (fo BCs) or τ_A (for ACs) (see 4.6). The network consists of two sub-layers of N regularly spaced BCs $i = 1...N$ and ACs $j = 1...3N$, which share spatial position $x_i = x_j$. Each BC is reciprocally connected to neighboring ACs, such that the connectivity between the two layers is symmetric. This connectivity scheme is simplified compared to biological lateral connectivity in the retina but allows mathematical analysis.

The connectivity is simulated via a the connectivity matrices Γ_A^B and Γ_B^A, which define the connections from BCs to ACs and from ACs to BCs respectively. Each BC i projects onto ACs $j = i - 1$ and $j = i + 1$ and vice versa such that $\Gamma_{A_{i-1}}^{B_i} = \Gamma_{A_{i+1}}^{B_i} = 1$ and 0 otherwise. Given that the connectivity between BCs and ACs is symmetric, $\Gamma_A^B = \Gamma_B^A$. Connections from BCs to ACs are excitatory and have a synaptic weight $w^+ \geq 0$ while connections from ACs to BCs are inhibitory and have a synaptic weight $w^- \leq 0$.

The synaptic input from each pre-synaptic BC and AC into a postsynaptic cell, $R_{B/A}(t)$ (eq. 4.7), is obtained by rectifying the pre-synaptic voltage signal $V_{B/A}(t)$ with a nonlinear function $N_{B/A}(V)$ (eq. 4.8).

Finally, in the third layer of the model ganglion cells pool over bipolar cell outputs within their receptive field. The weight w_i for each pre-synaptic BC i is therefore given by a Gaussian distribution depending on the distance of the cell i from the RF-center of the RGC (see equation 4.11). This layer consists of N differential equations for RGCS $k = 1...N$, which integrate their input with a shared time constant τ_G (eq. 4.10). Their voltage response is passed through another nonlinear function N_G to simulate a firing rate $R_G(t)$. Given the reciprocal nature of lateral connection we refer to this model as "Reciprocal-Amacrine-Model" (RAM). The mathematical details of the model are given in the Methods section.

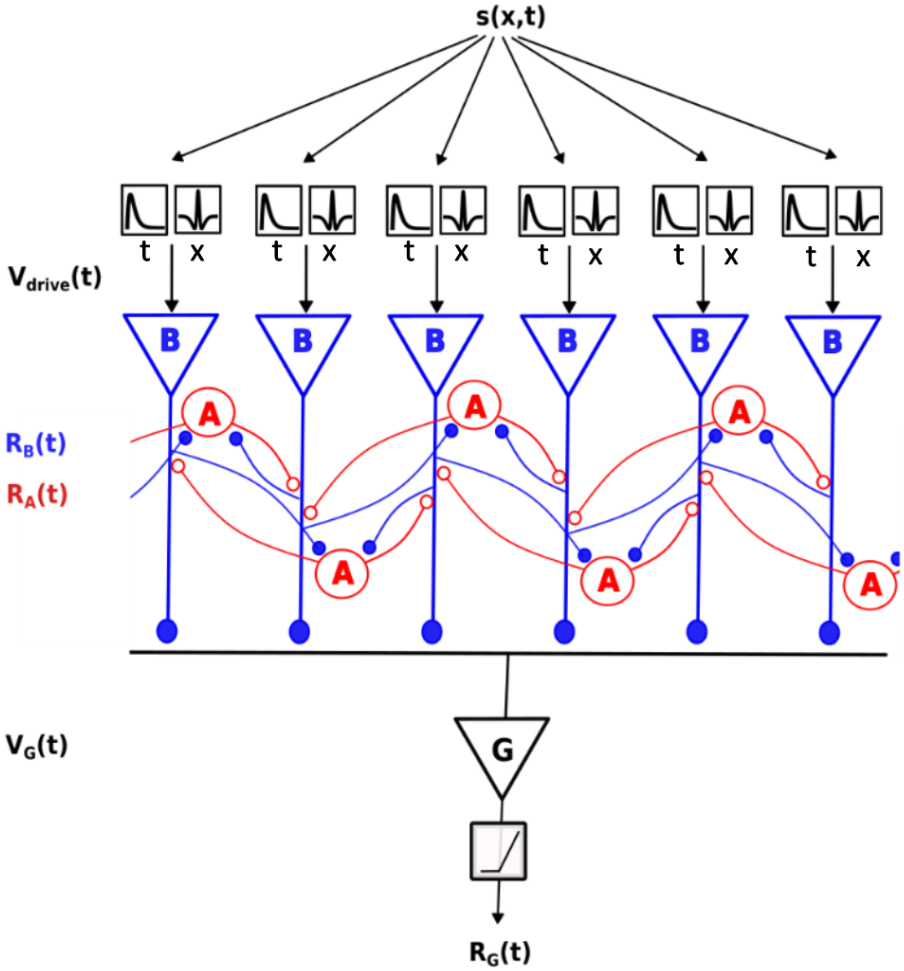

FIGURE 4.1: Schematic description of the Reciprocal Amacrine model (RAM)

The stimulus $s(x,t)$ is fed into a convolutional layer that simulates the transformation of the visual input into a neuronal voltage response, $V_{drive}(t)$, for each BC in the network. This convoluted signal is then fed into a network BCs and ACs, which are reciprocally connected and pass the synaptic signals $R_B(t)$ and $R_A(t)$ on to neighboring cells of the other type. A third layer of GCs pools over BCs within their receptive field and integrate their response $R_B(t)$ into their voltage $V_G(t)$, which is transformed into a Firing rate response $R_G(t)$ after rectification.

4.2.2 Lateral Connectivity can induce motion anticipation starting at the level of bipolar cell responses

In this section, we show how motion anticipation can arise via lateral amacrine connectivity in the RAM model. A phenomenological model that explains how motion anticipation can be achieved via gain control is the Adaptive-Cascade Model (ACM) proposed by Chen et al., 2013. This model also predicts amplified responses to motion onset and synchronized responses to motion reversal (see Chapter 2.3 and Chen et al., 2013; Chen et al., 2014), and yields an approximately constant amount of anticipation across a range of speeds between 0.1 and 1.0 mm/s (see Appendix A). Since we do not have experimental data for parameter optimization, we chose to set receptive field sizes, number of cells and spacing to the values used in the ACM as stated in Chen et al., 2013. We then optimized time constants and connectivity weights to match firing rate predictions of the ACM (see 4.4).

The mechanism with which amacrine connectivity causes motion anticipation becomes evident when we look at the internal responses of the model at each stage as shown in Figure 4.2: The moving bar evokes a voltage response in bipolar cells, simulated by the spatiotemporal convolution with its receptive field (Figure 4.2 **A**), which lags behind the stimulus. In a purely feed forward network, where $w^+ = 0$, this delay is passed on to BCs and slightly increases due to integration into membrane voltage (Figure 4.2 **B**, dotted blue). If $w^+ > 0$ however, BCs activate neighboring ACs as soon as they respond to the stimulus, which in turn inhibit BCs due to the reciprocal connections. This feedback loop truncates the bipolar response and yields an advancement of the response peak compared to its input (Figure 4.2 **B**, solid blue). The same mechanism takes place in all bipolar cells in the network, such that there is a cumulative effect of this feedback loop. The response of BC-AC interactions at site i propagates through the network and influences distant cells beyond nearest neighbor connectivity, impacting their response behavior. Ganglion cells pool over the bipolar cell layer with a Gaussian weighting (Equation 4.11) such that the response of each bipolar cell is scaled by a different synaptic weight onto the GC, $w_k R_{B_k}(t)$, (figure 4.2 **C**). Consecutively, ganglion cells that pool over the peak-advanced responses have an earlier response peak as well compared to as if no amacrine cells were present (Figure 4.2 **D**).

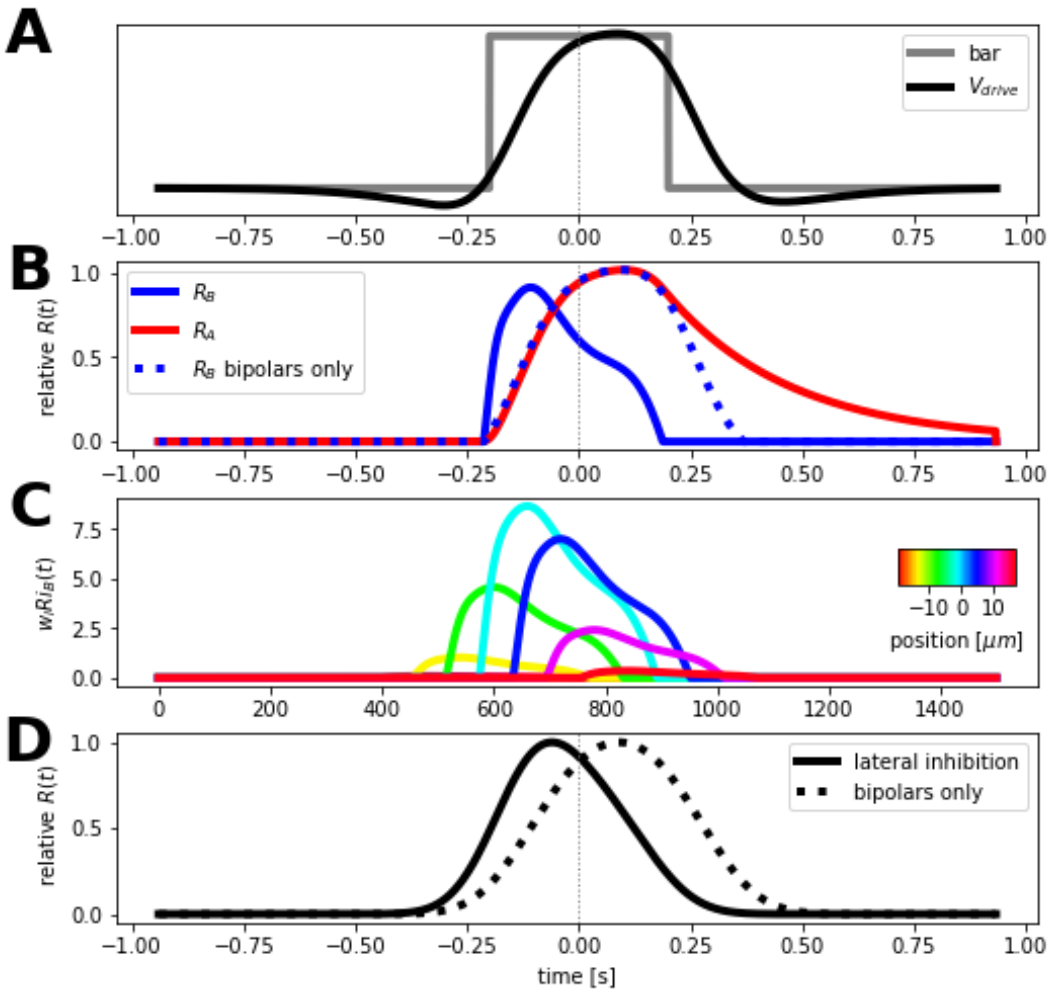

FIGURE 4.2: Lateral connectivity can evoke anticipation at the level of bipolar cells

Response of the model at each stage to a bar moving from left to right at 0.8 mm/s. **A.** The spatiotemporal convolution of bipolar cell receptive field and stimulus yields V_{drive}, which introduces a delay into the response peak. **B.** Bipolar cells respond (dotted blue, $w^+ = 0$) and transmit the signal to amacrine cells (red), which in turn inhibit bipolar cells and truncate the response (solid blue, $w^+ > 0$). **C.** Bipolar cell responses within a GCs' receptive field are weighted, summed and integrated into the GC response R_G. **D.** The peak of R_G is advanced and occurs before the bar middle reaches the receptive field center in the RAM model (solid black), while the response peaks delayed when no amacrine cells are present (dotted black).

4.2.3 Peak delay due to photo-transduction depends on bar speed

In the first stage of the model, the spatiotemporal convolution of the stimulus and a receptive field of bipolar cells yields V_{drive} (Eq. 4.2), which mimics the process of phototransduction and naturally introduces a delay between stimulus and response peak. In the case of a moving bar stimulus, this delay is quantified as $\delta t_{drive} = t_{drive} - t_{bar}$, where t_{bar} is the time-point at which the bar is at the receptive field center of a given cell and t_{drive} the time-point when the response of this cell peaks. If δt_{drive} is negative, the response peak lags behind the bar center, whereas if δt_{drive} is positive, the response peak is ahead of the bar and the response anticipates the bar position.

For a fixed bar width, δt_{drive} already depends on the speed of the moving bar

and the shape of the temporal filter. Figure 4.3 **A** shows V_{drive} in response to moving bars with 3 different speeds where either a monophasic or biphasic temporal filter was used (Figure 4.3 **B**). For a fixed network length D, a moving bar stimulus with 0.1 mm/s takes 10x longer to cross the network than one with 1.0 mm/s. Thus, for better comparison between traces across the speed range tested, we plot V_{drive} at time t against the distance of the spatial position of the bar center from the RF center at time t. Similarly, we transform the delay between response peak and the time-point at which the bar is at the receptive field center, δt_{drive}, into a spatial measure $\delta X_{drive} = \delta t_{drive} v$. This corresponds to the spatial position of the response peak with respect to the center position of the moving bar.

For a monophasic filter with a time constant τ_{RF} and a receptive field center size σ_C, we can calculate the time it takes for the bar to cross half of the receptive field as $\tau_{cross} = \frac{\sigma_C}{v}$ and define the parameter $\rho = \frac{\tau_{RF}}{\tau cross}$ as a measure of the distance covered by the bar within the integration time of the cell (Figure 4.3 **C**). For low speeds, that is to say $\rho < 1$, the bar takes longer to cross (half of) the RF than the characteristic time τ_{RF}. V_{drive} then reaches its maximum close to when the bar is centered above the RF (Figure 4.3 **A**, upper panel). As the bar moves faster, ρ increases. When $\rho > 1$, the bar already crossed half of the RF before the cell can fully integrate the stimulus. The peak response of V_{drive} significantly starts to lag behind the bar (Figure 4.3 **A**, lower panel).

Many bipolar cells are however thought to have a biphasic temporal response profile (Baccus and Meister, 2002) and are commonly modelled with a biphasic temporal kernel. If the temporal filter of the convolution is biphasic, the response peak is already ahead of the bar center for low speeds (Figure 4.3 **A**, upper panel). V_{drive} peaks ahead of the bar for low speeds because of the biphasic shape of the temporal filter, which acts as a change detector, responding to the appearance of the bar, then ceasing to respond because the rebound phase of the filter cancels out the response and finally hyper-polarizing when the bar disappears. For faster speeds, the cancellation via the filter rebound starts after the bar has already left the receptive field and the response peak is delayed (Figure 4.3 **A**, lower panel).

In both scenarios, that is to say with mono- and biphasic temporal filter, albeit absolute values differ, the scaling between peak delay and speed is similar: With increasing speed, the δX_{drive} increases (Figure 4.3 **D**).

However, the scaling between response amplitude and speed is opposite for mono- and biphasic temporal profiles. In the monophasic scenario, the amplitude of the response decreases the faster the bar gets, because the response does not reach its maximum before the bar leaves the RF. On the contrary, in the biphasic case, the amplitude increases with speed, because the positive part of the filter is faster than the negative one and over-weights (Figure 4.3 **E**).

Altogether, we have seen that the filter shape already has a strong effect on the peak response of V_{drive} and can already cause anticipation, that is to say a positive δX_{drive}, if the filter is biphasic. To ensure that anticipatory effects in our simulations come exclusively from lateral connectivity, we from now on consider a network where bipolar cells have a monophasic filter.

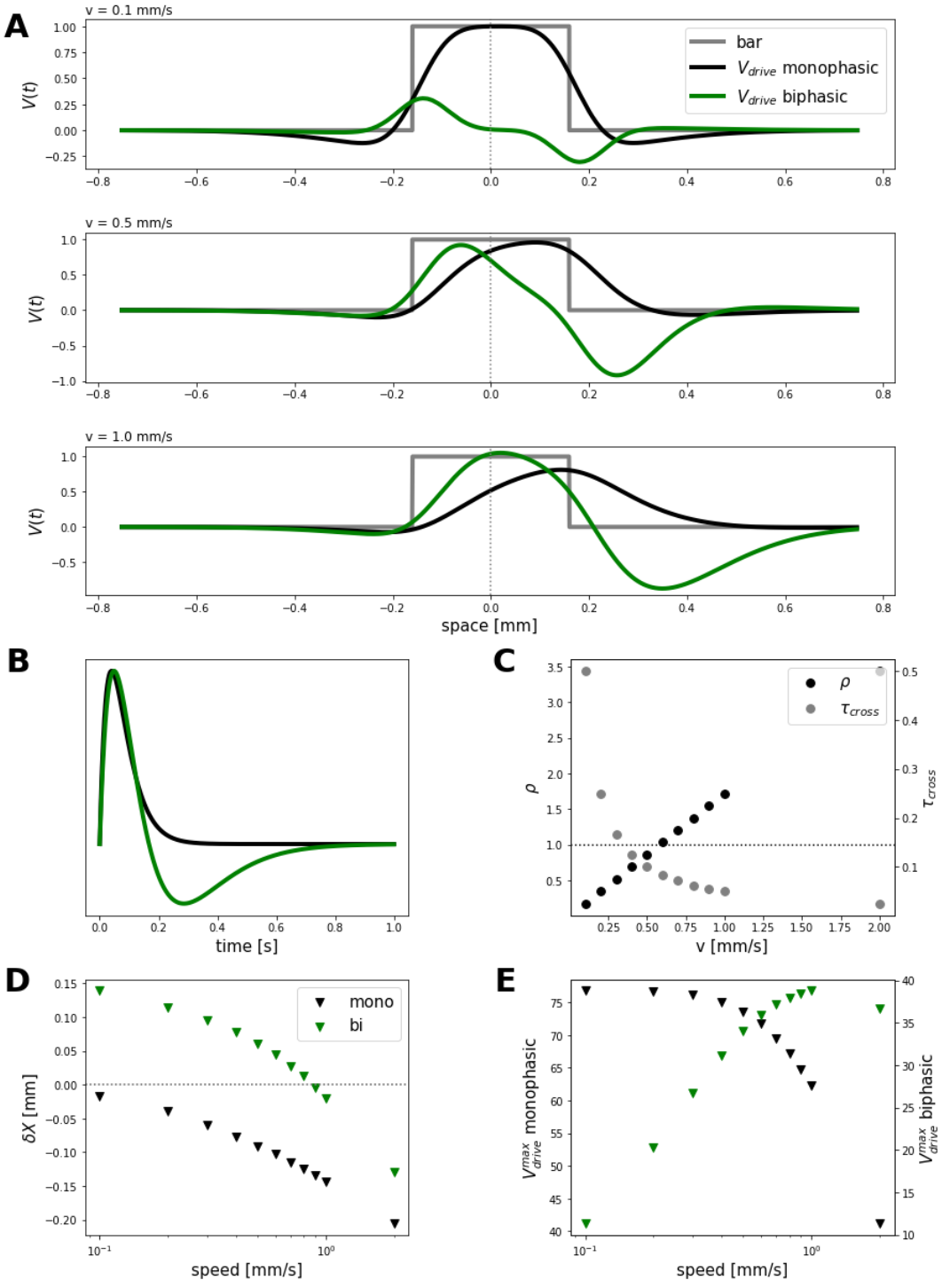

FIGURE 4.3: The response peak of V_{drive} depends on bar speed and temporal filter shape

A. V_{drive} response to moving bars at 0.1, 0.5 and 1.0 mm/s with monophasic or biphasic temporal profile. $V_{drive}(t)$ is plotted against the distance of spatial position of the bar center from the RF centre at time t. Response amplitudes are normalized to the maximum response across all speeds for monophasic and biphasic simulations separately. **B.** Temporal kernels used for monophasic and biphasic simulations. **C.** ρ and τ_{cross} for different speeds of the moving bar and a monophasic filter. **D.** Spatial position of the response peak with respect to the center position of the moving bar, $\delta A_{drive} = \delta t_{drive} v$ for different speeds with monophasic or biphasic filter. **E.** Peak amplitudes in V_{drive} across speeds for monophasic and biphasic simulations.

4.2.4 Lateral amacrine network effect counteracts photo-transduction delay stronger for faster speeds

It is observed experimentally that the spatial anticipation of the response peak with respect to the bar position is approximately constant across a range of speeds between 0.1 and 1.0 mm/s. However, the results of photo-transduction simulation imply that the peak delay due to photo-transduction increases with increasing speed (see Figure 4.3). It is thus necessary that an anticipation mechanism not only shifts the peak ahead of the bar position but also has a stronger impact on faster speeds. We observed that the lateral amacrine network indeed counteracts the photo-transduction delay specifically for faster speeds (Figure 4.4). To understand how, it is useful to look at the shift of the peak response at the different stages in the model, across different speeds.

Amacrine cell network can yield anticipation but not a constant scaling for a speed range of 0.1-1.0 mm/s

Figure 4.4 shows how the different stages of the model, V_{drive}, the synaptic inputs R_B, R_A and the RGC firing rate $R_G(t)$ respond to bars moving at 0.1, 0.5, and 1.0 mm/s. In a purely feed-forward architecture without lateral connections, the scaling described above for $V_{drive}(t)$ propagates to the bipolar cell and ganglion cell response, while the integration into voltage response adds another small delay to all speeds at all stages. The delay between BC and RGC response peak at t_B and t_G, and the bar center,t_{bar} is defined as $\delta t_{R/G} = t_{R/G} - t_{bar}$. This temporal measure can be transformed into a spatial measure as well, yielding $\delta X_{R/G} = \delta t_{R/G} v$.

Lateral amacrine connectivity corrects for this delay via a negative feed-back mechanism, which decreases and truncates the bipolar response causing bipolar cell responses to peak earlier than in a purely excitatory network (4.2 **B**, compare dotted to solid blue line).

ACs introduce a peak advancement in bipolar and ganglion responses R_B and R_G compared to the network input V_{drive}, which we quantify as $\Delta X = \delta X_{drive} - \delta X_{B/G}$ for BCs and RGCs (4.4 **D**). If the ΔX is larger than the delay introduced by photo-transduction δX_{drive}, the peak response is shifted before the moving bar hits the RF center, yielding anticipation. The amount of this peak advancement depends on the bar speed in the following manner.

If the bar moves slowly across the network (Figure 4.4 **A**), bipolar cells are depolarized for an extended amount of time which lasts quite longer than the time constant of integration into AC voltage (in red). ACs thus closely follow the slow time-course of the stimulus and BC responses. They effectively decrease the overall response amplitude in BCs while only slightly affecting the peak timing.

For increasing bar speeds (Figure 4.4 **B**), the bipolar cell response starts to lag behind the stimulus and decays more rapidly. The delay in amacrine cell integration thus introduces a discrepancy in amacrine and bipolar cell response time-courses. Due to this discrepancy, the negative feed-back via ACs, BC responses first increase without inhibitory feedback and then start to decrease when AC inhibition starts, leading to a truncated response that peaks earlier. Lateral amacrine connectivity thus starts to counteract the increase in response delay with increasing speed.

If speed increases even more (Figure 4.4 **C**), the discrepancy between bipolar and amacrine response time course increases as well, until the amacrine feedback with eventually arrive after the bipolar cell already passed it's peak, thus again affecting the peak timing only little.

In a bounded range of speeds, lateral amacrine connectivity can thus counteract the increase in response delay with increasing speed. For a network with a given parameter set, there is a speed for which the anticipatory effect via lateral amacrine connectivity is maximal, which is around 0.6 mm/s for the choice of parameters used in this study. Up-until this speed, the network exhibits an almost equal, slightly increasing anticipation time, whereas the curve steeply decreases afterwards. (4.4 **E**).

Thus, our RAM model does not reach an equally advanced anticipation time across the full speed range but shows a parabolic scaling with a maximum anticipation at 0.6 mm/s. This does not comply with experimental recordings where is has been found that anticipation remains constant in a speed range between 0.1-1.0 mm/s. Notably, the shape of the anticipation-speed curve depends on the parameters of the model, especially the time constants of the second network layer, τ_B and τ_A, and synaptic connectivity weights w^+ and w^-. In the following section we explore how anticipation depends on these parameters.

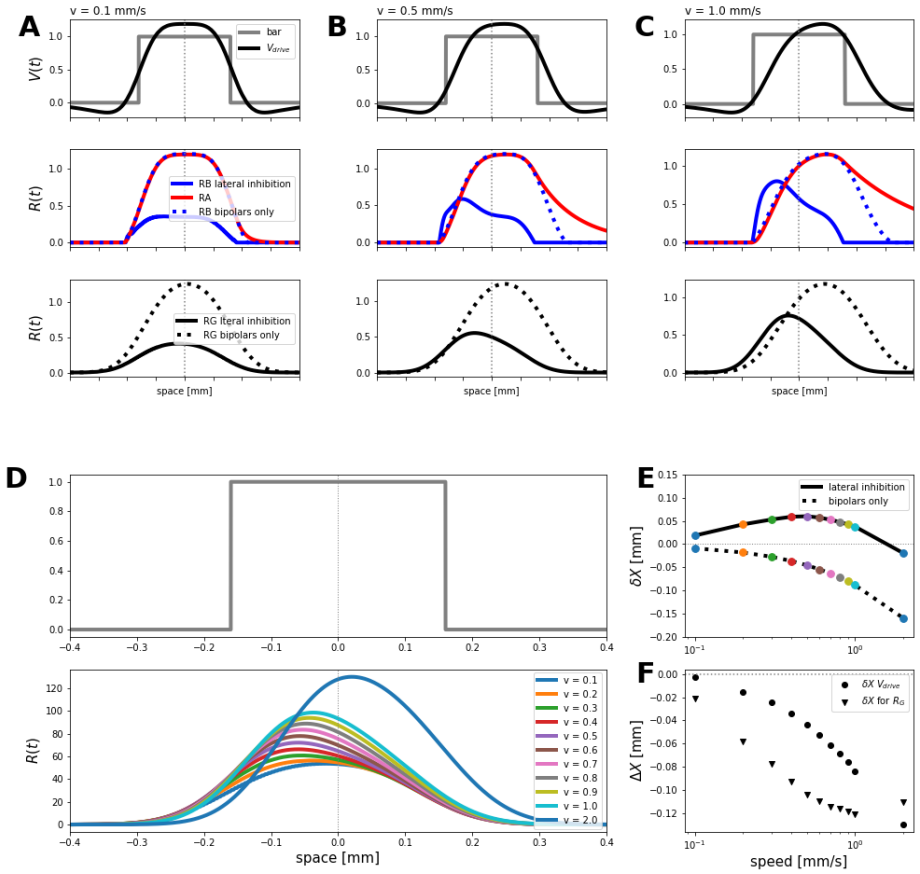

FIGURE 4.4: Amacrine network cancels photo-transduction delays for a certain speed range

A -C. Response traces at 3 stages of the model, plotted against the distance of spatial position of the bar center from the RF centre at time t, for 0.1 (**A**), 0.5 (**B**) and 1.0 mm/s (**C**), motion from left to right. Upper: Bar stimulus (grey) and V_{drive} (black) for a cell positioned at $x = 1.5$ mm. Middle: Bipolar response R_B (solid blue) with lateral amacrine connectivity, or without amacrine input (dotted blue). Amacrine response R_A plotted in red. Lower: Ganglion cell response rate R_G. **D.** R_G in responses (lower panel) to a moving bar (upper panel) to different speeds. **E.** Spatial position of the response peak with respect to the center position of the moving bar, $\delta X_G = \delta t v$ of R_G int with (solid) and without (dotted) amacrine inhibition for different speeds. **F.** Spatial peak advancement δX_G in R_G compared to V_{drive} due to amacrine network (black triangles) against speed. Delay of V_{drive} response peak (black circles) is plotted for reference.

Peak advancement depends on parameters for lateral connectivity

It has been shown previously that a biphasic temporal response profile in bipolar and ganglion cells may come from amacrine inhibition in the retinal network (Evgenia et al., 2023). In a computational model similar to the one used here, Kartsaki and colleagues have shown that the shape of the temporal response profile can switch from monophasic to biphasic depending on the parameter $\mu = w\tau^2$, where $w = -w^- w^+$ and $\tau = \frac{1}{\tau_B} - \frac{1}{\tau_A}$ ((Souihel and Cessac, 2021), see section 2.4.1). Increasing w corresponds to a increased strength in the feed-back loop between ACs and BCs. The parameter τ defines the difference between the time constants τ_A and τ_B. Finally, the parameter μ controls the eigenvalues of the dynamical system which simulates BC and AC interactions (Eq. 4.6). The switch from monophasic to

biphasic corresponds to the emergence of complex eigenvalues in a specific region of the parameter space. The biphasicness of the network response thus depends on a combination of lateral connectivity weights w as well as the combination of time constants τ. We thus explored how anticipation depends on the contributions of weights and time constants, and how this dependence is affected by the bar speed. Figure 4.5 shows that anticipation indeed depends on the combination of weights and time constants of the model, and that a region can be distinguished in which anticipation occurs. Notably, the shape of this region and amplitudes of anticipation change with the speed of the bar.

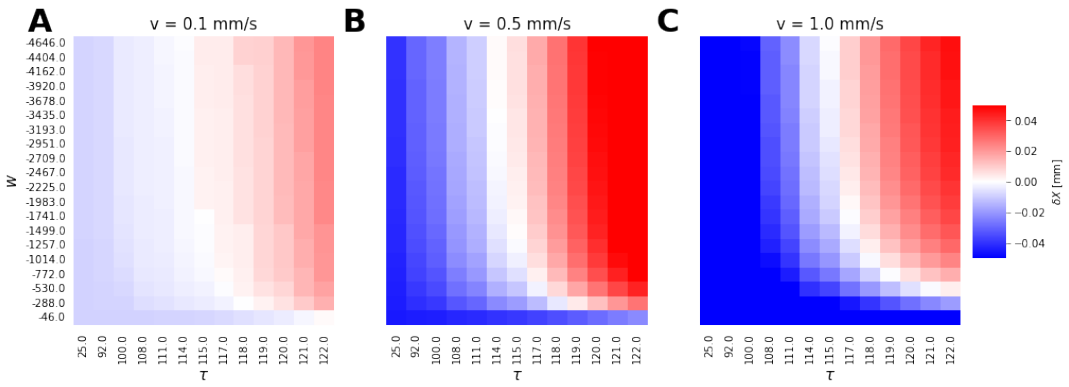

FIGURE 4.5: Peak advancement depends on parameters for lateral connectivity.

A-C. Heatmap showing the spatial anticipation A in R_G across w and τ for bars moving at 0.1 (**A**), 0.5 (**B**) and 1.0 mm/s (**C**). Red regions anticipate the bar while the peak response in blue region lags behind the bar center.

This suggests that, by dynamically adapting either weights, time constants or both to the speed of the bar, the representative part of the dynamical system can be moved into the red area. In this way a constant anticipation time across speeds may be obtained. It is well established that synaptic outputs adapt to a current stimulus via short-term plasticity, which dynamically tunes synaptic weights. This process of short-term plasticity can also affect the time constant of a postsynaptic cell since the change in postsynaptic conductance changes the overall membrane resistance (Abbott et al., 1997). In the following section we will thus show how STP might adjust connectivity weights to the speed of the bar to maintain a constant anticipation time between 0.1 and 1.0 mm/s. For simplicity we will restrain to the effect of STP on synaptic weights and keep time constants fixed.

4.2.5 Short-term plasticity could maintain constant anticipation across speeds

In the following section, we explore how plasticity affects the scaling between anticipation and speed. To this end, we look at the effect of plasticity on anticipation to different speeds in two main cases. First, we explore short-term depression via vesicle depletion at the output of the Reciprocal-Bipolar-Amacrine network, in the synapse from BCs to GCs. Second, we look at STP within the lateral network, at the reciprocal synapse from ACs to BCs. 2.4.3

Plasticity in Bipolar to Ganglion synapses increases anticipation

We found that short-term depression at the output synapse of bipolar cells to ganglion cells overall amplified the peak advancement in the GC response (Figure 4.6). It does so via a similar negative feedback mechanism as gain control in bipolar cells (see Chapter 2.3.2): As soon as BCs start responding, the vesicle occupancy n^B decreases which leads to a reduced vesicle release onto GCs. The output of bipolar cells R_B is thus truncated by synaptic depression and peaks early (Figure 4.6 **A**). The depletion of vesicles varies with the speed of the bar (Figure 4.6 **B**).

Even though the equilibrium occupancy is the same for all speeds, because all bars have the same contrast, different speeds yield different levels of occupancy via the time the bar spends in the cells RF. At low speed, the cell is exposed to the moving bar for the longest duration, the occupancy is thus lowest for the slowest speed tested. The faster the bar, the less time it spends above the cells RF, vesicle occupancy thus increases with increasing speed. We quantified the additional peak advancement via plasticity as the difference between anticipation in the plastic network ($\delta X_{plastic}$) compared to one fixed synaptic weights (δX_{fixed}) as $\delta X_{diff} = \delta X_{fixed} - \delta X_{plastic}$ (Figure 4.6 **C**), which is biggest for slow speeds because the negative feedback via vesicle depletion is strongest here. Additionally, the preceding AC inputs left the peak largely unaffected. For fast speeds, there is almost no effect due to plasticity, since occupancy decrease is comparatively small and the effect on the amacrine network over-weights. Overall, plasticity from bipolar to ganglion cells shifts the anticipation-speed curve upwards but does not affect the shape of the anticipation-speed curve (Figure 4.6 **D**)

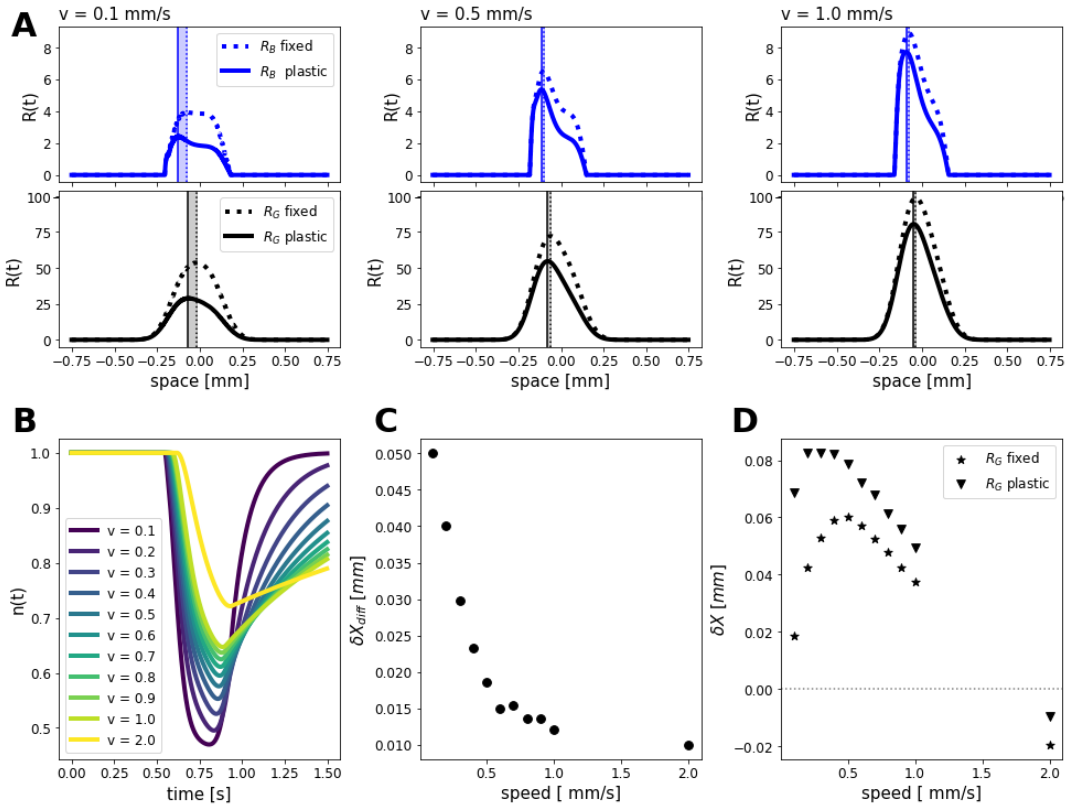

FIGURE 4.6: Depression in bipolar cell increases anticipation.

A. Response traces plotted against the distance of spatial position of the bar center from the RF centre at time t, for 0.1 (left), 0.5 (middle) and 1.0 mm/s (right). Upper: Bipolar response R_B after undergoing STP (solid blue) or with fixed output weight (dashed blue). Lower: Ganglion cell response rate R_G with or without STP in the bipolar input (solid vs dashed green). Shaded area highlights δX_{diff}, the peak shift due to plasticity. **B.** Occupancy traces $n(t)$ plotted against the distance of spatial position of the bar center from the RF centre at time t for speeds between 0.1-1.0 mm/s. **C.** δX_{diff} across speeds. Positive values indicate an increased anticipation compared to fixed simulations. **D.** Spatial Anticipation δX against speeds in the network with fixed synaptic weights (stars) and with dynamic adaptation (triangles).

Plasticity in Amacrine to Bipolar Synapses Decreases Anticipation

Within the lateral connectivity between bipolar and amacrine cells, the effect of plasticity is opposite. It overall decreases the peak advancement caused by lateral connectivity. Implemented in the synapse from amacrine to ganglion cells, the mechanism is the following: depression in amacrine synapses leads to a smaller suppression of bipolar cells via the amacrine network, which in turn will decrease the peak advancement via the network effect (see 4.5). The slower the bar, the lower the vesicle pool occupancy and thus, the weaker the peak advancement effect. For fast speeds on the other hand, plasticity only slightly decreases the amacrine effect. Thus depression in amacrine cells decreases peak advancement especially for slow bars. Plasticity from bipolar to amacrine cell has the same qualitative effect and is not shown here.

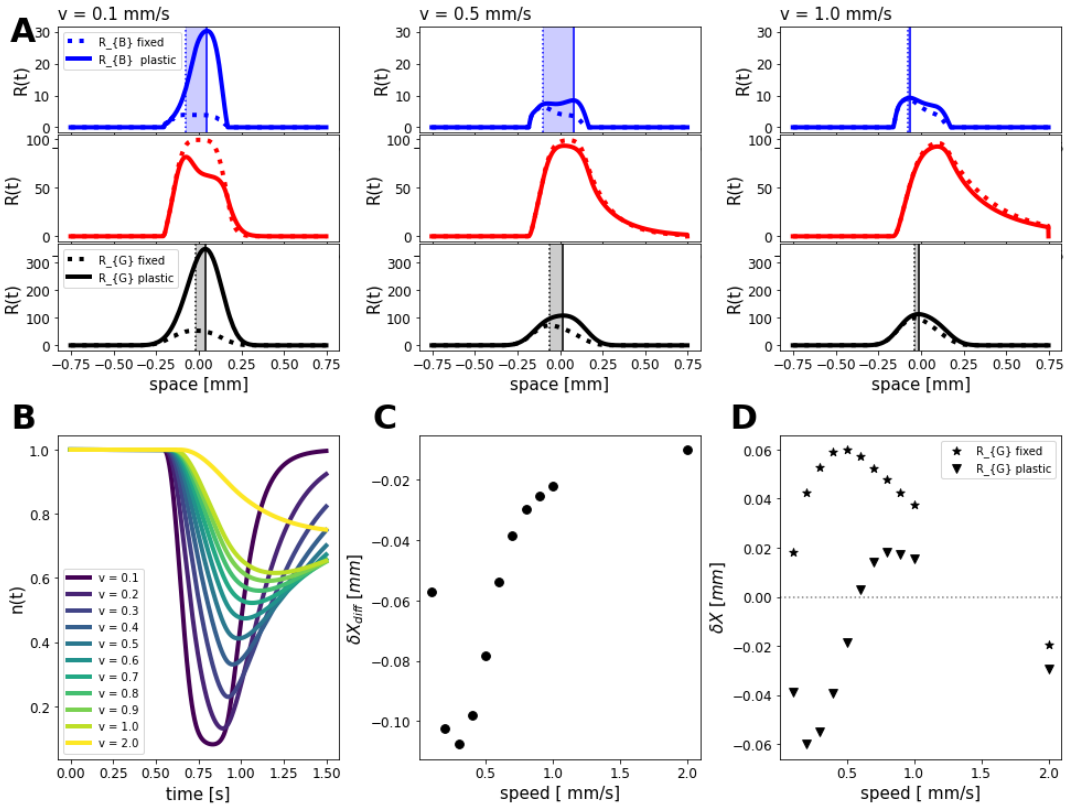

FIGURE 4.7: Depression in Amacrine cells decreases anticipation.

A. Response traces plotted against the distance of spatial position of the bar center from the RF center at time t, for 0.1 (left), 0.5 (middle) and 1.0 mm/s (right). Upper: Bipolar response R_B receiving dynamic (solid blue) or with fixed input from amacrine cells (dashed blue). Middle: Amacrine response R_A after undergoing STP (solid red) or with fixed output weight (dashed red). Lower: Ganglion cell response rate R_G with or without STP in the bipolar input (solid vs dashed green). Shaded area highlights δX_{diff}, the peak shift due to plasticity. **B.** Occupancy traces $n(t)$ plotted against the distance of spatial position of the bar center from the RF centre at time t for speeds between 0.1-1.0 mm/s. **C.** δX_{diff} across speeds. Negative values indicate an decreased anticipation compared to fixed simulations without STP. **D.** Spatial Anticipation δX against speeds in the network with fixed synaptic weights (stars) and with dynamic adaptation (triangles).

Vesicle Kinetics impact Anticipation Tuning

Notably, the effect of plasticity strongly depends on the recovery and release rates of the vesicle dynamics. The parameters determining vesicle kinetics are the recovery and release rates k_{rec} and k_{rel}, as well as the scale factor β. The effect of these parameters on the evolution of vesicle occupancy can be summarized via a vesicle time constant τ_n (see Chapter 2.4.3, equation 2.22). If the vesicle dynamics are too fast, such that a steady state is reached for all speeds tested, the effect of plasticity will be the same for all speeds within the time τ_n (Figure 4.8 **A,B**, uniform blue rows). On the other extreme, if vesicle dynamics are too slow, the change in occupancy in the time the bar spends in the cells' RF will be too small to have an effect on anticipation (Figure 4.8 **A,B**, uniform red rows). The dynamics of STP thus need to be tuned to react in the range of τ_{cross} for a desired range of speeds.

Figure 4.8 shows how anticipation to the speeds range tested is impacted by depression in bipolar and amacrine cells, respectively. In bipolar cells (Figure 4.8 A),

anticipation remains largely unaffected for slow vesicle dynamics, while if τ_n^B becomes very small, the anticipation speed curve is shifted upwards without affecting the anticipation scaling with speed (Figure 4.8 C). Only in the dynamic range, depression in bipolar cells affects the slow speeds more than fast ones as described above.

In amacrine cells (Figure 4.8 B), anticipation remains unaffected for slow vesicle dynamics as well, whereas very fast vesicle dynamics result in a response very similar to a network without lateral connectivity, because the weight of amacrine output quickly drops to 0. In the dynamic range, depression in ACs decreases anticipation more strongly for faster speeds due to lower occupancy. Interestingly, as vesicle dynamics speed up, anticipation first decreases, until it reaches a minimum and then increases again. This is because as vesicle dynamics speed up, τ_n^A becomes smaller and allows for faster speeds to affect vesicle dynamics within τ_{cross} (Figure 4.8 D).

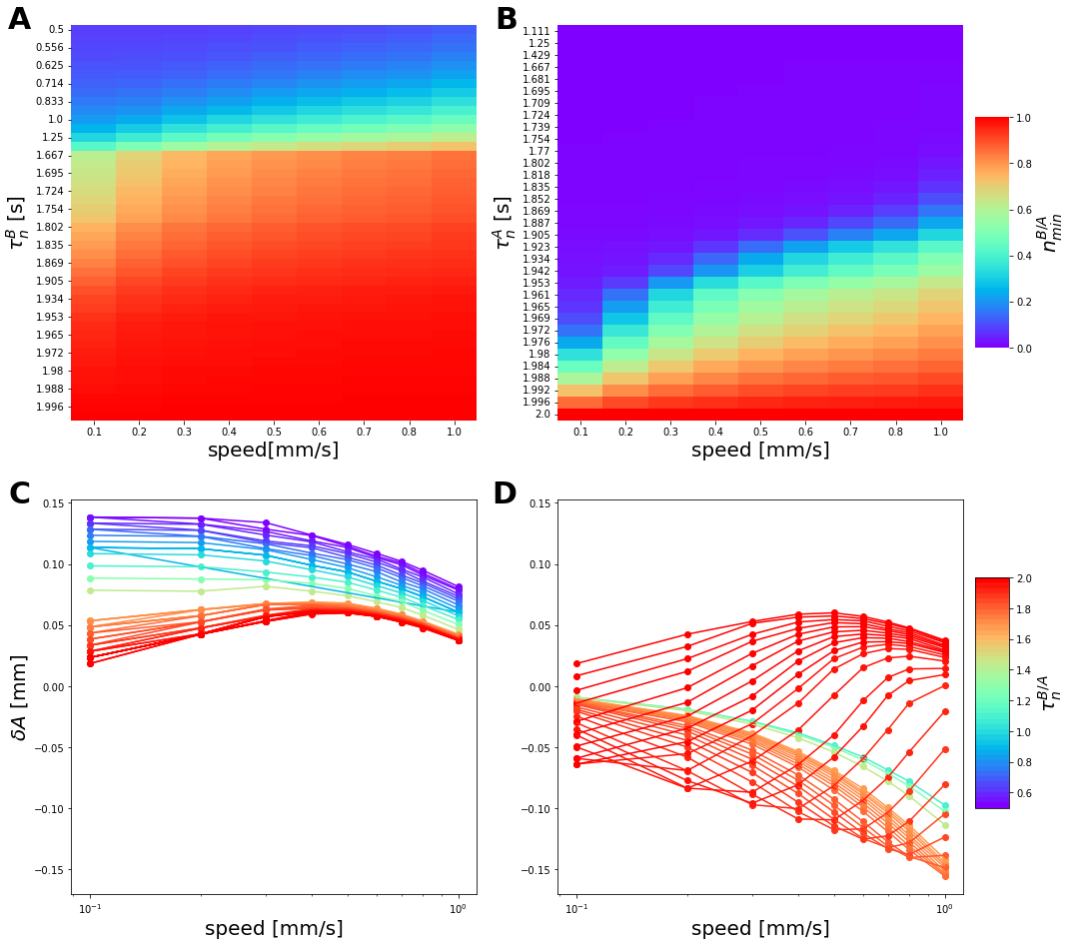

FIGURE 4.8: The effect of depression in anticipation depends of vesicle kinetics.

A-B. Heatmap showing the minimum vesicle occupancy n_{min} observed at different bar speeds for depression bipolar to ganglion (**A**) or amacrine-to-bipolar (**B**) synaptic terminals with different kinetic time constants τ_n. The time constant of vesicle dynamics usually depends on pre-synaptic activity. To illustrate the time constant based on parameters and independent of pre-synaptic activity, we plot $\tau_n = \frac{1}{k_{rec}+k_{rel}\beta}$. **C-D** Spatial Anticipation δX of the moving bar against speed for different kinetic time constants τ_n during depression bipolar to ganglion (**A**) or amacrine-to-bipolar (**B**) synaptic terminals.

Combined Depression in Bipolar and Amacrine cells may support constant anticipation across speeds

Alone, neither plasticity at the output or within the lateral network can thus yield a constant anticipation time. Since the two mechanisms have opposite effects on anticipation, we next explored if a combination of both with different kinetics may stabilize anticipation time across the full range of speeds between 0.1-1.0 mm/s.

We thus explored the effect of combined plasticity in bipolar and amacrine cells across a large range of time constants for ACs and BCs. Figure 4.9 shows the anticipation map for τ_n^A and τ_n^B for a slow, intermediate and fast speeds. For each speed, there is a clearly distinguishable (red) region in which the network anticipates from the region where the response lags behind the bar (blue). This anticipatory region shifts towards slower dynamics for BCs and faster dynamics for ACs with increasing speeds.

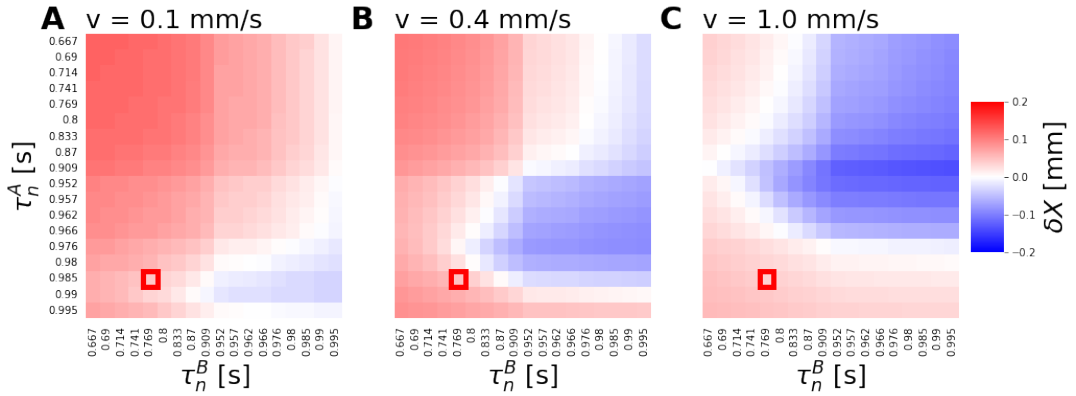

FIGURE 4.9: Combined depression in B and A yields minimal difference in anticipation by speed.

A-C. Heatmap showing the spatial anticipation δA in R_G across τ_n^B and τ_n^A for bars moving at 0.1 (**A**), 0.5 (**B**) and 1.0 mm/s (**C**). Red regions anticipate the bar while the peak response in blue region lags behind the bar center. The red square indicates the combination of τ_n^B and τ_n^A with minimal difference in δA across speeds between 0.1-1.0 mm/s

We then searched for the combination of time constants for BCs and ACs that yields the minimal difference between speeds in the range of 0.1 - 1.0 mm/s (red square in figure 4.9). The ganglion cell responses of this network to moving bars are shown in Figure 4.10 **A**. Anticipation time can be held approximately constant (4.10 B) if both excitatory and inhibitory plasticity are kept in the dynamic range (4.10 C). For low speeds around 0.1 mm/s the effect of excitatory plasticity over-weights and increases anticipation, whereas for the remaining speeds until 1.0 mm/s, anticipation is slightly decreased. Overall, combined depression in excitatory and inhibitory synapses can have a variety of effects on anticipation in GC responses across speeds and can be tuned to a certain anticipation values by tuning vesicle kinetics.

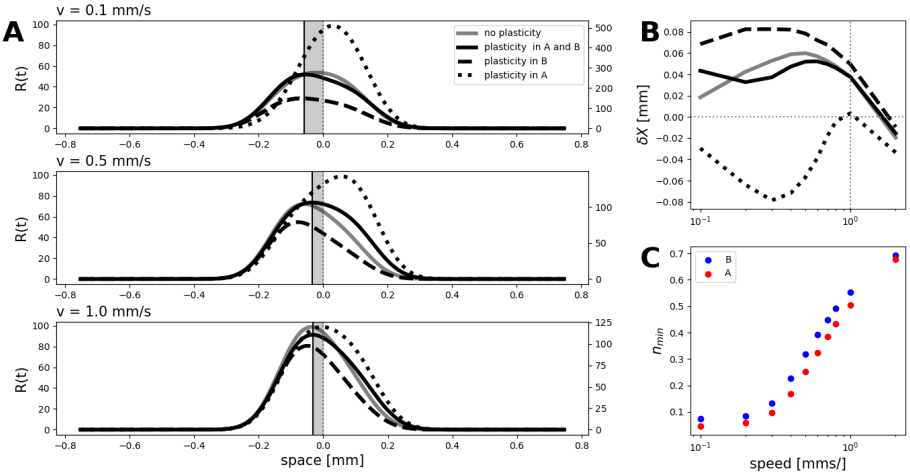

FIGURE 4.10: Combined Depression in B and A yields approximately stable anticipation across speeds.

A Ganglion cell response traces plotted against the distance of spatial position of the bar center from the RF centre at time t, for 0.1 (upper), 0.5 (middle) and 1.0 mm/s (lower). Shown are response traces in a network with fixed synaptic weights (grey), plasticity in the BC-to-GC or AC-to-BC synapse only (dashed or dotted) and with depression in both synapses (solid). Shaded area highlights δA, the spatial anticipation for the networ with combined plasticity. **B** Spatial Anticipation δA against speeds in the network with fixed synaptic weights (grey), plasticity in the BC-to-GC or AC-to-BC synapse only (dashed or dotted) and with depression in both synapses (solid). **C** Minimum occupancy n^A_{min} (red) and n^A_{min} (blue) for different speeds.

4.2.6 Early response onset via lateral amacrine connectivity

Next, we investigated how lateral amacrine connectivity affects the start of the response to a moving bar. It has been observed that motion anticipation can occur by advancing the onset of the response in addition to the peak (Lee and Menz, 2020). This phenomenon cannot be accounted for by gain control, as a reduction in response gain truncates the latter part of the response but leaves the initial rise unaffected, since gain reduction starts with a delay only after the cell already started to respond.

In our RAM model, bipolar cell responses start earlier in the presence of amacrine cells compared to a purely feed-forward network without lateral inhibition. Note that this holds only if lateral connections are kept in the linear synaptic range: The reduction in BC voltage due to feedback inhibition decreases amacrine cell activity, which then will cause bipolar activity to raise their potential again, creating a feedback loop that imposes oscillatory activity on the network. In addition, due to the nearest-neighbor connectivity, a decrease in amacrine cell n potential will activate bipolar cell $n+2$, causing an increase in membrane potential ahead of the bar, which leads cells to respond earlier. Since all connections are reciprocal, all of these loops will propagate back to the initially activated cell and contribute to their response as well. These interactions will add additional layers of complexity to the response and can trigger early onset responses.

Figure 4.11 shows that, starting at the level of BCs, the response start is earlier when linear lateral amacrine connections are present as compared to a network with bipolar sub-units only.

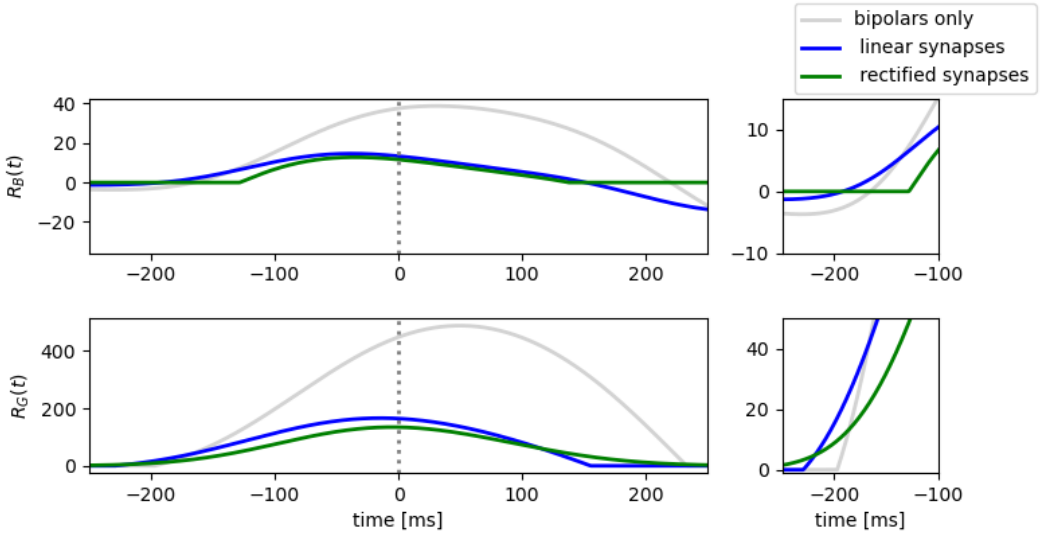

FIGURE 4.11: Linear lateral connectivity can produce anticipation in response onset

A-B. BC Response (**A**) and GC Response (**B**), without amacrine inhibition (grey), with linear synapses (blue) and with rectified synapses (green). Zoom-in to show response onset on the right.

4.2.7 Lateral connectivity does currently not account for motion onset and reversal responses

Next, we tested if our RAM model could also detect surprise in spatio-temporal motion patterns. We therefore probed the model with two types of stimuli that have previously been shown to evoke surprise responses in the retina: motion onset and motion reversal.

If a static bar centered above a cells' receptive field suddenly starts moving, ganglion cells strongly respond to motion onset - even more strongly than to smooth motion across the entire receptive field (Chen et al., 2013). Similarly, when a moving bar suddenly reverses its motion near the ganglion cells' receptive field, it emits a large response which occurs with a constant latency regardless of the exact location of the reversal, resulting in a large synchronised response within nearby ganglion cells, signalling the unexpected change in trajectory (Chen et al., 2014).

Both motion onset and reversal responses have been accounted for by the Adaptive-Cascade Model (ACM) (Chen et al., 2013; Chen et al., 2014), see A.

Using the parameter-set from the previous sections, we found that our model produces responses to both motion onset and motion reversal, but it does not match important properties of the experimentally observed responses (Figure 4.12). When the RAM is tested with the motion onset stimulus (Figure 4.12 A), it emits a transient response to the appearance of the bar at the beginning of the stimulus and when the bar starts moving, which is in agreement with experimental observations. However, while the bar is steadily displayed above the RF center, a GC in the RAM does not return to baseline as experimentally observed, instead it keeps responding at a steady rate. In addition, the response to motion onset is not bigger than the response to a smoothly moving bar.

When the RAM is tested with the motion reversal stimulus, GCs in the network exhibit a response peak to the bar moving in the initial direction, followed by a second peak evoked by the bar moving in the opposite direction after reversal.

In contrast to the synchronized reversal responses observed experimentally, GC responses to the reversing bar shift backwards the further the reversal occurs from the receptive field center in the RAM. For cells very close to the reversal location, the excitation triggered by motion in the new direction merges with excitation triggered by motion before reversal, such that there is no peak signalling the reversal. The further the RF is located from the reversal position, the longer the bar takes to return and the later the response occurs. The reversal responses thus occur consecutively and not synchronized among ganglion cells.

Although the RAM cannot account for motion onset or reversal responses at this stage, it should be noted that the parameter used in this simulations were tuned only to match smooth motion responses of the Adaptive Cascade Model and not to experimental data of smooth motion, motion onset or motion reversal responses. It might be possible that the RAM gives much better approximations of experimental findings if parameters are optimized to match experimental data (see 4.3).

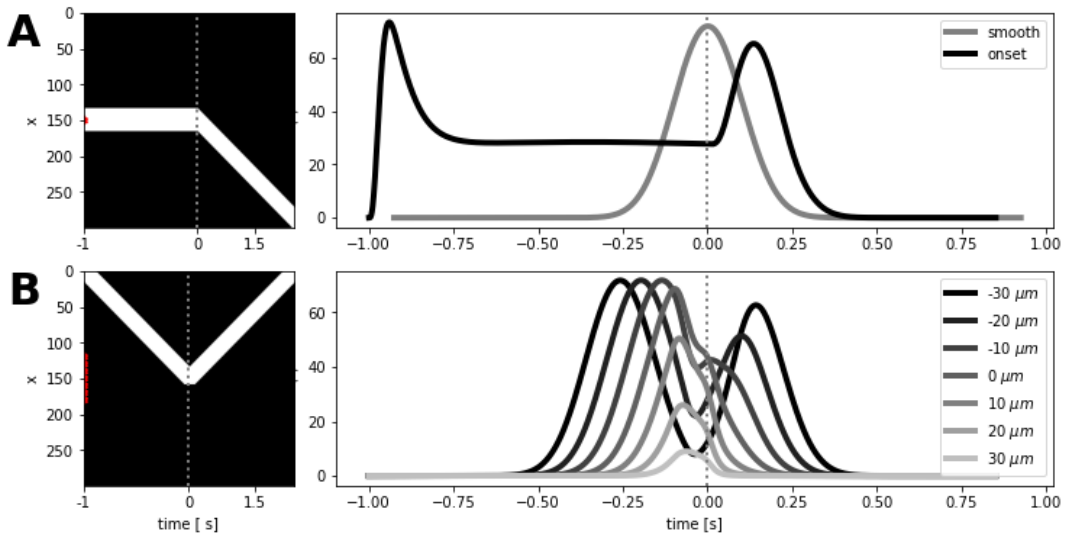

FIGURE 4.12: The plastic RAM cannot accurately explain motion onset and reversal responses for the initial parameter set

A. Left panel shows the stimulus and right panel shows the RAM R_G response to smooth motion (grey) and motion onset (black). Traces are aligned such that that motion starts when the smoothly moving bar middle is centered over the cells' RF (time 0), such that both stimuli are identical after motion onset. **B.** RAM R_G response to motion reversal. Shown are responses of 8 different cells with different distances to the reversal location. Traces are aligned such that the bar reverses at time 0.

4.3 Discussion

The retina is known to anticipate moving objects in order to compensate for transmission delays. This anticipation of moving stimuli requires a precise temporal representation of speed in order to expect the object at the right location at the right time. In this study, we explored how dynamical synapses in a laterally connected retinal network could serve to establish such a temporal accuracy in spatio-temporal pattern predictions. We show that reciprocal connectivity between amacrine and bipolar cells can cause motion anticipation, potentially being an underlying mechanism of gain control at the bipolar cell level. By adjusting synaptic weights dependent on the recent stimulus history, the network may adjust its steady-state as to contain

a temporal representation of when a stimulus is expected to arrive. This leads a constant anticipatory amount to a given range of speeds.

Gain control models predict retinal responses to motion very well, however, they provide a rather phenomenological explanation and yield little mechanistic insight. Several previous studies proposed mechanistic implementations based on amacrine cell inputs, which have been shown to play a crucial role in motion anticipation in the goldfish retina (Johnston and Lagnado, 2015) and to contribute to shifting responses to earlier times via dis-inhibitory inputs in the mouse (Lee and Menz, 2020). The aforementioned studies mainly considered cells as isolated units with nonlinear processing, but ignored higher order inputs due to complex interactions in the network. Souihel and Cessac, 2021 proposed that amacrine cells can act on motion anticipation in the following manner: In a mathematical study of a retinal network with lateral inhibitory connectivity between bipolar cells, they show that complex indirect effects can arise from the network connectivity, inducing a wave of activity that amplifies the bipolar cell response and enhances the effect of gain control. However, it was not explained how anticipation via a lateral network connectivity scales with different stimulus speeds. Dynamical synapses have previously been proposed to enable neuronal circuits in the retina to form expectations of future inputs (Kastner and Baccus, 2014) on a spatio-temporal scale, yet the role of short term-plasticity in motion anticipation has not been explored. With this study, we thus provide a first model in which short-term plasticity tunes motion anticipation to the speed of the moving object, an idea which to our knowledge has not been explored so far.

Bipolar cells are generally simulated with a biphasic response profile, which can arise from intracellular processes such s vesicle dynamics, hyper-polarizing currents producing delayed inhibition or inhibitory input from amacrine cells (Suh and Baccus, 2014). Here we showed that the biphasic shape of the receptive field indeed already has a big impact on motion anticipation. This implies that underlying mechanisms of biphasic response profiles might be important for motion anticipation as well, rendering amacrine inputs plausible candidates.

4.3.1 Limitations

Our simple model of plasticity requires that synapses are rectified, while synapses in the biological networks could operate at a more linear range at various connections. This could add complex interactions to the network response that can lead to anticipation in response onset. It would be interesting to explore network responses where synapses are both linear and dynamic.

While a retinal network with dynamic reciprocal connections between neighboring bipolar and amacrine cells can anticipate the motion of a moving bar with a stable anticipation distance for a certain range of speeds, we were not able account for temporal synchronization of motion reversal responses of an amplified response to motion onset as compared to smooth motion with the same parameter set.

To account for motion reversal, responses to the bar after reversal should occur synchronized and not sequential. This behavior could potentially arise in the RAM with strong, long-lasting amacrine inhibition which could act similar to gain control by fully suppressing bipolar outputs for the duration of their time constants. For motion onset, GC responses should show two modifications: First, they should return to baseline after the appearance of the bar instead of maintaining a steady response rate. In order to achieve this, lateral connectivity weights and/or rectification thresholds would need to be tuned to yield zero response to the steady bar. Second, the response peak to motion onset should be bigger than the one to smooth motion. This

might be possible to achieve by tuning plasticity kinetics to fully suppress amacrine output, such that bipolar cells responding to the bar after motion onset are not suppressed by lateral inputs. These two modifications are contradictory and likely not possible to achieve with the same parameter set. Adding a second type of inhibitory cells to separate the response cancellation to a steady input from amplitude tuning could be a potential solution to this problem.

4.3.2 Perspectives

In our model architecture, we considered nearest neighbors connectivity where each cell connects to its two neighboring cells, also within a ganglion cells receptive field. We include only one single type of amacrine and bipolar cells, and do not account for surround inhibition via amacrine cells. In our previous work (see Chapter 3), we found evidence that depressing glycinergic synapses specifically fine-tune temporal expectations of regular temporal patterns, while excitation and inhibitory inputs that closely follow each other function as change detectors. In the future, it would be very interesting to see if a separate inhibitory population providing inputs specifically targeting temporal representations may yield better approximations of real neuronal responses to spatio-temporally surprising scenes rather than implementing plasticity within the prediction-input interactions. If these temporal tuning inputs would be of opposite stimulus polarity as the change detector, they could as well contribute to response onset anticipation.

Altogether, with this work we provide a first attempt to hypothesize how dynamical synapses could implement temporal precision in spatiotemporal pattern predictions. However, the current stage of the model developed in this chapter fails to reproduce important experimental features in spatiotemporal surprise detection. It would be beneficial for this work to fit its parameter to experimental data in order to more thoroughly explore its potential and limitations.

4.4 Methods

4.4.1 1-D Network model with lateral connectivity

In order to simulate the integration of lateral inputs we implement the retinal network as a dynamical system rather than a cascade of convolutions. The model can be described as follows.

We consider a 1-D moving bar stimulus $s(t)$ with a speed v (in mm/s) and width $2b$ (in mm). The stimulus is described as follows:

$$s(x,t) = \begin{cases} 1, & \text{if } -b + vt \leq x \leq b + vt\ ; \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

We simulate the voltage input from the OPL to a bipolar cell i with a receptive field center located at x_i via a spatiotemporal convolution:

$$V_i^{drive}(t) = \int_{-\infty}^{t} \int_{0}^{D} K(x - x_i, t - u) s(x,u) dx du \tag{4.2}$$

where the kernel :

$$K(x,t) = k_t(t) k_x(x) \tag{4.3}$$

consists of a temporal and a spatial profile. The temporal profile is given as a difference of α-kernels :

$$k_t(t) = \frac{t}{{\tau_{RF}^+}^2} e^{-\frac{t}{\tau_{RF}^+}} - wt_s \frac{t}{{\tau_{RF}^-}^2} e^{-\frac{t}{\tau_{RF}^-}} \tag{4.4}$$

where τ_{RF}^+ and τ_{RF}^- are the characteristic integration times of positive and negative inputs and wt_s is the relative weight of negative inputs. The spatial kernel has the Gaussian shape:

$$k_x(x) = e^{-\frac{(x-x_i)^2}{2\sigma_s^2}} - wx_s e^{-\frac{(x-x_i)^2}{2\sigma_c^2}} \tag{4.5}$$

where σ_c^B and σ_s^B are (half of) the size of the receptive field center and surround (in mm) of cell i with position x_i.

We then consider a retinal network spanning a 1-D plane with N = 300 bipolar cells positioned at x_i and N = 300 amacrine cells with the same spatial location. Cells cover a distance $D = 3\ mm$ and are spaced by $\delta = 0.005\ mm$. Each cell is characterized by its membrane potential V_{Bi} and V_{Aj} respectively. The dynamics of the cells are described by the dynamical system:

$$\begin{cases} \frac{dV_{Bi}}{dt} = -\frac{1}{\tau_B} V_{Bi} + \sum_{j=1}^{N} w^- \Gamma_{Bi}^{Aj} N_A(V_{Aj}) + F_i(t), \\ \frac{dV_{Aj}}{dt} = -\frac{1}{\tau_A} V_{Aj} + \sum_{j=1}^{N} w^+ \Gamma_{Aj}^{Bi} N_B(V_{Bi}), \end{cases} \tag{4.6}$$

The connectivity is simulated via the connectivity matrices Γ_A^B and Γ_B^A, which define the connections from BCs to ACs and from ACs to BCs, respectively. Each BC i projects onto ACs $j = i-1$ and $j = i+1$ and vice versa such that $\Gamma_{A_{i-1}}^{B_i} = \Gamma_{A_{i+1}}^{B_i} = 1$. All other entries are set to 0. We assume null boundary conditions. Given that the connectivity between BCs and ACs is symmetric, $\Gamma_A^B = \Gamma_B^A$. Connections from BCs to ACs are excitatory and have a synaptic weight $w^+ \geq 0$ while connections from ACs to BCs are inhibitory and have a synaptic weight $w^- \leq 0$. Here we chose the coupling between amacrine and bipolar cells to be symmetric to avoid instabilities in the system, as eigenvalues always have a negative real part with this condition (Kartsaki, 2022).

Synaptic inputs $R_B(t)$ and $R_A(t)$are obtained by rectifying pre-synaptic voltages such that:

$$R_X = N(V_X, \theta_X). \tag{4.7}$$

The piece-wise linear function N rectifies the synaptic connections and is given as:

$$N_X(V) = \begin{cases} V - \theta_X, & \text{if } V \geq \theta_X\ ; \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

The term:

$$F_i(t) = \frac{V_i^{drive}}{\tau_B} + \frac{dV_i^{drive}}{dt}, \tag{4.9}$$

is the stimulus driven input into BCs. It takes this form to ensure that $V_{Bi} = V_i^{drive}$ in the absence of AC coupling.

Finally, a layer of N = 300 Ganglion cells (GCs) follow the dynamical system

$$\frac{dV_{Gk}}{dt} = -\frac{V_{Gk}}{\tau_G} + \sum_{i=1}^{N} w_i N(V_{Bi}((t)), \tag{4.10}$$

where each GC k pools over the bipolar cell layer with Gaussian weights w_i, centered at the GC's position x_k (same as BC and AC position), and a width σ_c^G in mm. The scale factor w_{GB} determines the overall strength of synapses from bipolar to ganglion cells.

$$w_i(x) = w_{GB} e^{-\frac{(x-x_i)^2}{2\sigma_c^{G\,2}}}. \tag{4.11}$$

In the last step, the GC voltage V_{Gk} is transformed into a firing rate R_{Gk} via the nonlinear function N_X (4.8):

$$R_{Gk} = N(V_{Gk}, \theta_G). \tag{4.12}$$

4.4.2 Parameter Calibration

We chose to set receptive field sizes, number of cells and spacing to the values used in the ACM as stated in Chen et al., 2013 and optimized time constants and connectivity weights such that predicted firing rates of our model match those of the ACM for smooth motion to a moving bar at 0.8 mm/s. We used an evolutionary optimization algorithm using the CMAES (Covariance Matrix Adaptation Evolutionary Strategy) Python toolbox (Hansen, Akimoto, and Baudis, 2019).

4.4.3 Parameter Values

Parameter	Value	Unit
σ_c^B	0.05	mm
σ_s^B	0.20	mm
σ_c^G	0.065	mm
δ	0.005	mm
τ_{RF}^+	0.086 / 0.04	s
τ_{RF}^-	0.087	s
wt_s	1.0 / 0.0	V
τ_B	0.08	s
τ_A	0.28	s
τ_G	0.01	s
w^-	46.0 / 0.0	Hz
w^+	46.0	Hz
w_{GB}	0.04	Hz
s_B	1	HzV^{-1}
θ_B	0.0	V
s_A	1	HzV^{-1}
θ_A	0.0	V
s_G	1110	HzV^{-1}
θ_G	0.0	V
k_{rel}^B	0.5	Hz
k_{rec}^B	1.0	Hz
β^B	0.6 / 0.0	V^{-1}
k_{rel}^A	0.5	Hz
k_{rec}^A	1.0	Hz
β^A	0.04 / 0.0	V^{-1}

TABLE 4.1: RAM parameter values used in simulations

Chapter 5

Conclusion

The goal of this thesis was to investigate how the retina dynamically encodes unexpected inputs. Based in the hypothesis that the retina mainly encodes "surprise", i.e., deviations from the expectation made on the basis of the visual context, we sought to answer the two following questions:

1. What are the potential mechanisms allowing the retinal network to form expectations and compute surprise ?
2. What is the role of dynamical synapses in surprise computation?

In order to answer these questions we started to investigate the circuit mechanisms underlying the Omitted Stimulus Response, which is a simple yet not fully explained example of temporal pattern recognition and surprise coding in the retina. Here, the retina sends a massive response when a stimulus in a periodic sequence of flashes is omitted - only after the stimulus did not arrive at the expected time. Although the stimulation with full-field flashes at fixed frequency is very artificial and unlikely present in natural visual scenes, it allowed us to formulate two hypothesises in response to our initial questions :

1. Temporally displaced excitation and inhibition can act as temporal change detector tuned to an opposite feature and can successfully generate a 'surprise response' to an unexpected stimulus.
2. Dynamical synapses tune the timing of the surprise signal, thereby impinging temporal expectations on a neuronal network.

The idea of excitation and closely following inhibition serving as neuronal implementation of predictive coding has been around for more than 50 years (Srinivasan, Laughlin, and Dubs, 1982; Huang and Rao, 2011), where excitatory signals convey sensory inputs while inhibitory signals serve as predictions. The idea of dynamical synapses tuning for temporal expectation however is largely unexplored in the retinal literature. Thus we tested if the principle of temporal expectation tuning via dynamical synapses could be applied to other examples of predictive coding in the retina. We chose the phenomenon of motion anticipation, where retinal ganglion cells respond to a moving object before it reaches their receptive field center, thereby creating a representation of the object position that corrects for transmission delays in the network. We proposed that the retina might adapt its synaptic connections via dynamical synapses in order to adjust its anticipation to the speed of the moving object in order to anticipate a moving object at the right position at the right time. Thereby, the retina creates temporal expectations that are stored in the synaptic strengths within the network.

The results of this work led to first insights into the role of STP on a network level, but also raise a number of questions towards a better understanding of temporal pattern recognition, the effect of STP at the retinal network level, and circuit mechanisms for predictive coding. In the final pages of this thesis, I will outline future directions along these three axes.

5.1 Towards a better understanding of temporal pattern recognition in the retina and beyond

In the first part of this study, we explained temporal pattern recognition in a computational model with 2 important components: A biphasic response profile tuned to the opposite polarity of the flashes, here generated by ON excitation and slightly slower ON inhibition, and OFF selective inhibition with dynamic weight.

While we experimentally identified an important role for glycinergic inhibition in the latency shift of the OSR in experiments, the remaining cellular units of our model are not explicitly backed up by experiments. Especially ON inhibitory inputs are crucial in our model to generate an OSR, which could arise from GABAergic inhibition in the retina. We currently do not know if GABAergic transmission plays a role in the OSR, it would thus be important to test its involvement experimentally.

All units of our model directly form feed-forward inputs onto the RGC, which is the computationally simplest architecture but might not be the most realistic one in terms of biological connectivity. To get a better idea of circuit and cellular motifs relevant for the OSR it would further be interesting to see if the cells showing an OSR relate to known functional RGC types (Baden et al., 2013). A very prominent retinal circuit which involves glycinergic inhibition in the retina is the AII-circuit, where the AII amacrine cell receives glutamatergic inputs from ON Rod Bipolar Cells (RBCs) and provide glycinergic inhibition to OFF ganglion cells (Demb and Singer, 2012). In addition, they from gap junctions with ON cone bipolar cells. Probing the retina with flashed stimuli at very low vs. very high light levels could test if this pathway is involved here.

While the simple model from this study successfully replicates the latency shift in the OSR peak and allows us to identify the mechanistic contribution of each component, it fails to give a realistic representation of for example the response onset. The computational model we sketched could be extended by feed-back connectivity, thresholding or further adaptation mechanisms to provide a more detailed and realistic representation of retinal circuits in the mouse.

Further, if the retina forms temporal expectations of future inputs, it seems interesting to ask how precise these expectations are. A first approach here could be to test the robustnesss of the OSR to jittered flash sequences with different levels of noise in the frequency of the flash trains.

Finally, another important question that remains is how an error signal like the OSR is treated by upstream circuits in the visual system. Deviant stimuli evoke an amplified response if presented in between a row of similar stimuli in the auditory and visual domains - the mismatch negativity (see Chapter 2). Is there a link between the OSR and these cortical phenomena ?

5.2 Underlying mechanisms of expectation and surprise encoding

In the second part of this thesis, we explored how temporal fine tuning of anticipation to the speed of a moving object could be achieved via synaptic plasticity in a retinal network model with reciprocal nearest neighbor connectivity between bipolar and amacrine cells. While it was possible to adjust the timing of anticipation to the bar speed by introducing STD in this network, we could not show a realistic reproduction of surprise responses to motion stimuli thus far.

The composition of this network is inspired by the work of Souihel and Cessac, 2021, who showed that this architecture can successfully introduce motion anticipation, and we started by adding short-term plasticity to this network while maintaining the global architecture. Thus, there exists only one species of amacrine cells which jointly performs the task of prediction and temporal tuning whereas we previously developed the necessity for two separate inhibitory inputs. It would thus be interesting to see how different the principle of temporally displaced excitation and inhibition tuned to an opposite feature would impact motion anticipation, and responses to surprise in motion trajectories. It has been shown that certain ganglion cells have extraclassical receptive fields, that is to say regions beyond the suppressive surround that evokes change in their response (Ikeda and Wright, 1972), supposedly via dis-inhibition. Identifying extended spatiotemporal patterns that excite a given RGCs could help to shed light on the underlying circuitry innervating a given cell, which will ultimately help to establish a better and more realistic mechanistic understanding of how elements of the retinal network contribute to information processing.

Given the at least 40 types of different amacrine cells in the retina (Masland, 2012b), inhibition can come in many different flavors and can shape retinal responses in various ways. For example, simultaneous excitation and inhibition may modulate the gain of the cell locally by increasing membrane conductance (Chance, Abbott, and Reyes, 2002). Additionally, factors such as shunting inhibition (Abbott et al., 1997) and presynaptic inhibition (Zhang, Wu, and Rasch, 2015; Ayaz and Chance, 2009) influence a cells gain. All these mechanisms mediate divisive modulation of the cellular response to incoming inputs, reducing the postsynaptic membrane potential by division rather than subtraction (Abbott et al., 1997). It has been proposed that divisive inhibition serves as a better mechanism to compare an input to an internal prediction than subtractive inhibition (Chalk et al., 2017), and also to contribute to dynamically reshaping receptive fields (Cui et al., 2016). It would be interesting to implement divisive inhibition in combination with STP in the models we developed as to investigate how it affects network predictions compared to subtractive inhibition.

Another important type of synaptic connections which are abundantly found in the retina are gap junctions, which have not been addressed in this study so far. Gap junctions have been suggested to contribute to motion anticipation responses in the retina (**Souhiel2020**; Liu et al., 2021), it thus seems interesting to ask how they could affect surprise encoding.

5.3 The effect of dynamical synapses on the retinal code

The main hypothesis elaborated in this thesis is that dynamical synapses shape temporal expectations in retinal computations.

Yet we only have indirect evidence for its involvement, derived from predictions of our plasticity model on the effect of the number of flashes in the OSR, which have been confirmed in experiments. STP has been studied in great detail at the level of single synapses. Thanks to these studies we have a detailed understanding of synaptic dynamics and their function in retinal adaptation to contrast and luminance. However, little is known about how STP affects the circuit computations in the network. It would be very valuable to establish a protocol to identify the effect of STP in the preceding retinal circuit at the level of RGC output. It has recently been shown that STP can be measured in the spiking output of cortical cells via a combination of STP and LN- models (Ghanbari et al., 2017). It could be interesting to adapt this method to the retina.

Further, the results of this thesis imply that short-term plasticity can shape temporal (and potentially spatial) preferences of RGCs. Since it is known that STP is massively affected by luminance and contrast in the stimulus, it could be interesting to measure spatial and temporal frequency tuning curves at different levels of contrast and luminance.

The ability to detect and represent temporal patterns is critical in order to interact with a dynamically changing environment. With this work, we emphasize that temporal representations are already present in the retina and propose a new purpose for short-term adaptation in the retina that is to store information about the temporal evolution of a visual scene.

Neurons that selectively respond to temporal features have been observed across species and sensory modalities (see Chapter 2.3.1). Implying that temporal representations are of fundamental importance. It is thus possible that short-term plasticity may generally serve to establish temporal expectations, possibly at the cortical level, for more complex temporal pattern recognition.

It therefore seems important to further explore and characterize how dynamical synapses shape the retinal code.

Appendix A

Adaptive Cascade Model

The Adaptive-Cascade Model first proposed in Chen et al., 2013 phenomenologically explains motion anticipation, amplified responses to motion onset and synchronized responses to motion reversal. We implemented the model as described in the literature to confirm the scaling between anticipation and bar speed. We took all parameter values provided in (Chen et al., 2013) and approximated receptive field parameter form the given figures. We then fitted the parameter of gain control to match the speed simulated responses to moving bars of different speeds shown in the publication. All parameters that we use are within 20% of the values stated in Chen et al., 2013. In this setting, we can reproduce qualitatively similar responses to motion onset and motion reversal as stated in the literature. The model predicts an almost constant latency of around 100 μm ahead of the actual position of the bar, which slightly decreases with increasing speed.

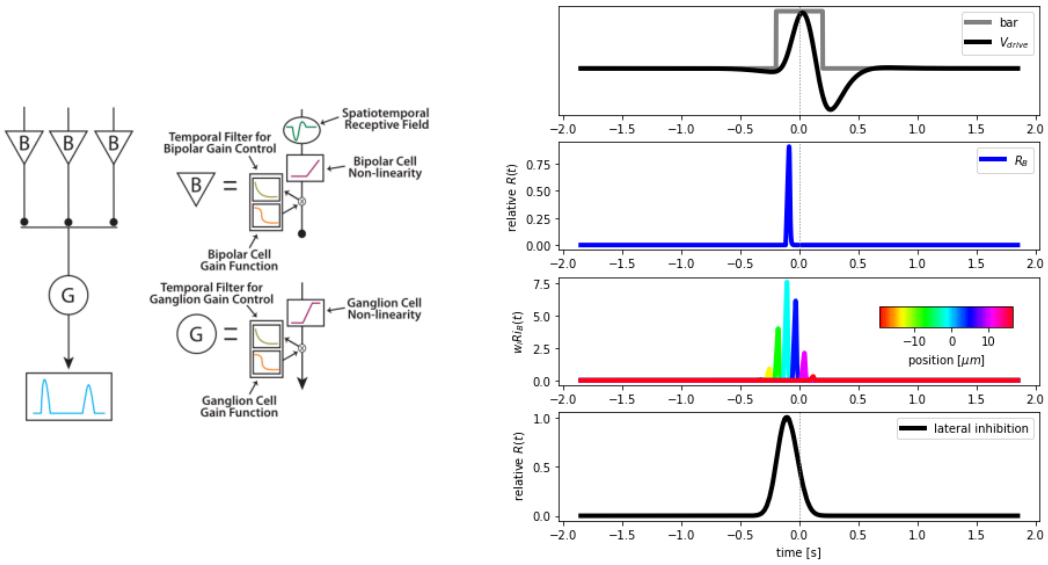

FIGURE A.1: The Adaptive Cascade Model and motion anticipation

A Schematic description of the model. From Chen et al., 2013 . **B** Responses at each stage in the model.

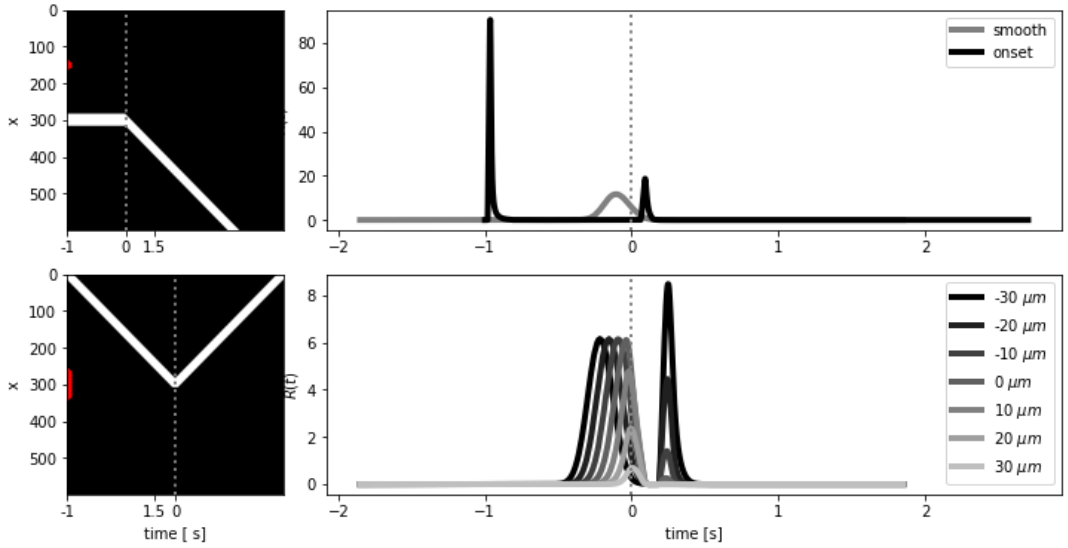

FIGURE A.2: Motion onset and motion reversal in the ACM

A Motion onset response is bigger than response to smooth motion **B** Motion reversal response is synchronized amoing ganglion cells near the reversal location.

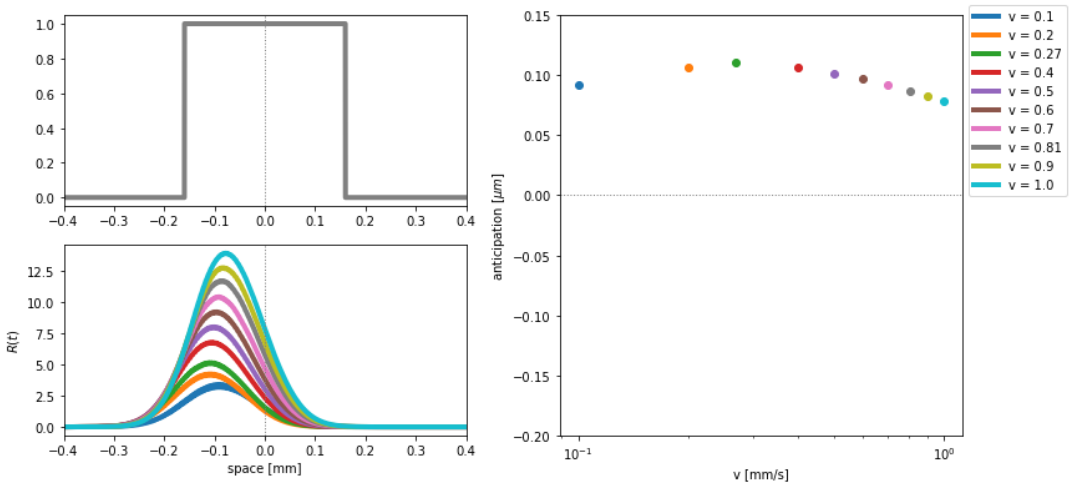

FIGURE A.3: Scaling of motion anticipation for different speeds

A Responses to moving bars of different speeds between 0.1-1.0 mm/s. **B** Scaling of anticipation to bar speed is nearly constant.

Parameter	Value	Unit
σ_B^{center}	0.05	mm
$\sigma_B^{surround}$	0.20	mm
σ_G^{center}	0.065	mm
δ	0.005	mm
τ_{OPL}	0.086 / 0.04	s
τ_{OPL2}	0.087	s
S	1.0 / 0.0	$-$
s_B	1	HzV^{-1}
θ_B	7.35	V
s_G	1110	HzV^{-1}
θ_G	0.0	V
H_G	0.000459	1
τ_G^{Act}	0.1995	s
H_B	0.1	1
τ_B^{Act}	0.00611	s

TABLE A.1: ACM parameter values used in simulations

Bibliography

Abbott, L. F. et al. (Jan. 1997). "Synaptic Depression and Cortical Gain Control". In: *Science* 275 (5297), pp. 221–224. ISSN: 0036-8075. DOI: 10.1126/science.275.5297.221.

Alexander, Kenneth R. (2017). "Information Processing: Retinal Adaptation". In: Elsevier. DOI: 10.1016/B978-0-12-809324-5.01403-6.

Amsalem, Oren et al. (2020). "Dense Computer Replica of Cortical Microcircuits Unravels Cellular Underpinnings of Auditory Surprise Response". In: *bioRxiv*. DOI: 10.1101/2020.05.31.126466. eprint: https://www.biorxiv.org/content/early/2020/06/01/2020.05.31.126466.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/01/2020.05.31.126466.

Antinucci, Paride and Robert Hindges (Feb. 2018). "Orientation-Selective Retinal Circuits in Vertebrates". In: *Frontiers in Neural Circuits* 12. ISSN: 1662-5110. DOI: 10.3389/fncir.2018.00011.

Anwar, Haroon et al. (Apr. 2017). "Functional roles of short-term synaptic plasticity with an emphasis on inhibition". In: *Current Opinion in Neurobiology* 43, pp. 71–78. ISSN: 09594388. DOI: 10.1016/j.conb.2017.01.002.

Astikainen, Piia, Elina Lillstrang, and Timo Ruusuvirta (Dec. 2008). "Visual mismatch negativity for changes in orientation - a sensory memory-dependent response". In: *European Journal of Neuroscience* 28 (11), pp. 2319–2324. ISSN: 0953816X. DOI: 10.1111/j.1460-9568.2008.06510.x.

Astikainen, Piia et al. (2013). "Event-related potentials to unattended changes in facial expressions: detection of regularity violations or encoding of emotions?" In: *Frontiers in Human Neuroscience* 7. ISSN: 1662-5161. DOI: 10.3389/fnhum.2013.00557.

Awatramani, Gautam B. and Malcolm M. Slaughter (Sept. 2000). "Origin of Transient and Sustained Responses in Ganglion Cells of the Retina". In: *The Journal of Neuroscience* 20 (18), pp. 7087–7095. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.20-18-07087.2000.

Ayaz, Asli and Frances S. Chance (Feb. 2009). "Gain modulation of neuronal responses by subtractive and divisive mechanisms of inhibition". In: *Journal of Neurophysiology* 101 (2), pp. 958–968. ISSN: 00223077. DOI: 10.1152/jn.90547.2008.

Baccus, S. A. et al. (July 2008). "A Retinal Circuit That Computes Object Motion". In: *Journal of Neuroscience* 28 (27), pp. 6807–6817. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4206-07.2008.

Baccus, Stephen A. and Markus Meister (Dec. 2002). "Fast and Slow Contrast Adaptation in Retinal Circuitry". In: *Neuron* 36 (5), pp. 909–919. ISSN: 08966273. DOI: 10.1016/S0896-6273(02)01050-4.

Baden, Tom, Thomas Euler, and Philipp Berens (Jan. 2020). "Understanding the retinal basis of vision across species". In: *Nature Reviews Neuroscience* 21 (1), pp. 5–20. ISSN: 1471-003X. DOI: 10.1038/s41583-019-0242-1.

Baden, Tom et al. (Aug. 2013). "Spikes and ribbon synapses in early vision". In: *Trends in Neurosciences* 36 (8), pp. 480–488. ISSN: 01662236. DOI: 10.1016/j.tins.2013.04.006.

Baden, Tom et al. (Jan. 2016). "The functional diversity of retinal ganglion cells in the mouse". In: *Nature* 529 (7586), pp. 345–350. ISSN: 0028-0836. DOI: 10.1038/nature16468.

Baden, Tom et al. (Mar. 2018). "The Functional Organization of Vertebrate Retinal Circuits for Vision". In: Oxford University Press. DOI: 10.1093/acrefore/9780190264086.013.68.

Barlow, H. (Jan. 2001). "Redundancy reduction revisited". In: *Network: Computation in Neural Systems* 12 (3), pp. 241–253. ISSN: 0954-898X. DOI: 10.1080/net.12.3.241.253.

Barlow, Horace B et al. (1961). "Possible principles underlying the transformation of sensory messages". In: *Sensory communication* 1.01.

Barne, Louise Catheryne et al. (June 2022). "Temporal prediction elicits rhythmic preactivation of relevant sensory cortices". In: *European Journal of Neuroscience* 55 (11-12), pp. 3324–3339. ISSN: 0953-816X. DOI: 10.1111/ejn.15405.

Berry, Michael J. et al. (Mar. 1999). "Anticipation of moving stimuli by the retina". In: *Nature* 398 (6725), pp. 334–338. ISSN: 0028-0836. DOI: 10.1038/18678.

Bullock, T. H. et al. (Sept. 1990). "Event-related potentials in the retina and optic tectum of fish". In: *Journal of Neurophysiology* 64 (3), pp. 903–914. ISSN: 0022-3077. DOI: 10.1152/jn.1990.64.3.903.

Bullock, Theodore H. et al. (July 1994). "Dynamic properties of human visual evoked and omitted stimulus potentials". In: *Electroencephalography and Clinical Neurophysiology* 91 (1), pp. 42–53. ISSN: 00134694. DOI: 10.1016/0013-4694(94)90017-5.

Burrone, Juan and Leon Lagnado (2000). "Synaptic Depression and the Kinetics of Exocytosis in Retinal Bipolar Cells". In: *Journal of Neuroscience* 20.2, pp. 568–578. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.20-02-00568.2000. eprint: https://www.jneurosci.org/content/20/2/568.full.pdf. URL: https://www.jneurosci.org/content/20/2/568.

Cammann, Rainer (June 1990). "Is there a mismatch negativity (MMN) in visual modality?" In: *Behavioral and Brain Sciences* 13 (2), pp. 234–235. ISSN: 0140-525X. DOI: 10.1017/S0140525X00078420.

Cessac, Bruno (Jan. 2022). "Retinal Processing: Insights from Mathematical Modelling". In: *Journal of Imaging* 8 (1), p. 14. ISSN: 2313-433X. DOI: 10.3390/jimaging8010014.

Chalk, Matthew et al. (June 2017). "Sensory noise predicts divisive reshaping of receptive fields". In: *PLoS Computational Biology* 13 (6). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005582.

Chance, Frances S, L.F Abbott, and Alex D Reyes (Aug. 2002). "Gain Modulation from Background Synaptic Input". In: *Neuron* 35 (4), pp. 773–782. ISSN: 08966273. DOI: 10.1016/S0896-6273(02)00820-6.

Chander, Divya and E. J. Chichilnisky (Dec. 2001). "Adaptation to Temporal Contrast in Primate and Salamander Retina". In: *The Journal of Neuroscience* 21 (24), pp. 9904–9916. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.21-24-09904.2001.

Chen, Eric Y. et al. (Jan. 2013). "Alert response to motion onset in the retina". In: *Journal of Neuroscience* 33 (1), pp. 120–132. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.3749-12.2013.

Chen, Eric Y. et al. (Nov. 2014). "The neural circuit mechanisms underlying the retinal response to motion reversal". In: *Journal of Neuroscience* 34 (47), pp. 15557–15575. ISSN: 15292401. DOI: 10.1523/JNEUROSCI.1460-13.2014.

Chichilnisky, E. J. (Feb. 2001). "A simple white noise analysis of neuronal light responses". In: *Network: Computation in Neural Systems* 12 (2), pp. 199–213. ISSN: 0954-898X. DOI: 10.1080/713663221.

Chichilnisky, E. J. and Rachel S. Kalmar (Apr. 2002). "Functional Asymmetries in ON and OFF Ganglion Cells of Primate Retina". In: *The Journal of Neuroscience* 22 (7), pp. 2737–2747. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.22-07-02737.2002.

Cui, Yuwei et al. (Nov. 2016). "Divisive suppression explains highprecision firing and contrast adaptation in retinal ganglion cells". In: *eLife* 5 (NOVEMBER2016). ISSN: 2050084X. DOI: 10.7554/eLife.19460.001.

Dacheux, Ramon F. and Robert F. Miller (Mar. 1976). "Photoreceptor-Bipolar Cell Transmission in the Perfused Retina Eyecup of the Mudpuppy". In: *Science* 191 (4230), pp. 963–964. ISSN: 0036-8075. DOI: 10.1126/science.175443.

Dayan, P and L Abbott (Jan. 2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Vol. 15.

Demb, Jonathan B. and Joshua H. Singer (Jan. 2012). "Intrinsic properties and functional circuitry of the AII amacrine cell". In: *Visual Neuroscience* 29 (1), pp. 51–60. ISSN: 0952-5238. DOI: 10.1017/S0952523811000368.

— (Nov. 2015). "Functional Circuitry of the Retina". In: *Annual Review of Vision Science* 1 (1), pp. 263–289. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-082114-035334.

Denève, Sophie and Christian K Machens (Mar. 2016). "Efficient codes and balanced networks". In: *Nature Neuroscience* 19 (3), pp. 375–382. ISSN: 1097-6256. DOI: 10.1038/nn.4243.

Deshmukh, Nikhil R. and Michael J. Berry (2019). "Nonlinear transfer and temporal gain control in ON bipolar cells". In: *bioRxiv*. DOI: 10.1101/514364. eprint: https://www.biorxiv.org/content/early/2019/01/08/514364.full.pdf. URL: https://www.biorxiv.org/content/early/2019/01/08/514364.

Despotović, Danica (2022). "Encoding surprise by retinal ganglion cells". In: *PhD Thesis*. URL: https://theses.hal.science/tel-03646660.

Diamond, Jeffrey S (Sept. 2011). "Grilled RIBEYE stakes a claim for synaptic ribbons". In: *Nature Neuroscience* 14 (9), pp. 1097–1098. ISSN: 1097-6256. DOI: 10.1038/nn.2914.

Diamond, Jeffrey S. (Sept. 2017). "Inhibitory Interneurons in the Retina: Types, Circuitry, and Function". In: *Annual Review of Vision Science* 3 (1), pp. 1–24. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-102016-061345.

Ding, Jennifer et al. (June 2021). "Spatially displaced excitation contributes to the encoding of interrupted motion by a retinal direction-selective circuit". In: *eLife* 10. ISSN: 2050084X. DOI: 10.7554/eLife.68181.

Dong, Cun-Jian and Frank S. Werblin (Apr. 1998). "Temporal Contrast Enhancement via $GABA_C$ Feedback at Bipolar Terminals in the Tiger Salamander Retina". In: *Journal of Neurophysiology* 79 (4), pp. 2171–2180. ISSN: 0022-3077. DOI: 10.1152/jn.1998.79.4.2171.

Dunn, Felice A., Martin J. Lankheet, and Fred Rieke (Oct. 2007). "Light adaptation in cone vision involves switching between receptor and post-receptor sites". In: *Nature* 449 (7162), pp. 603–606. ISSN: 0028-0836. DOI: 10.1038/nature06150.

Dunn, Felice A. and Fred Rieke (Mar. 2008). "Single-Photon Absorptions Evoke Synaptic Depression in the Retina to Extend the Operational Range of Rod Vision". In: *Neuron* 57 (6), pp. 894–904. ISSN: 08966273. DOI: 10.1016/j.neuron.2008.01.031.

Ermentrout, G. Bard and David H. Terman (2010). *Mathematical Foundations of Neuroscience*. Vol. 35. Springer New York. ISBN: 978-0-387-87707-5. DOI: 10.1007/978-0-387-87708-2.

Euler, Thomas et al. (Aug. 2014). "Retinal bipolar cells: elementary building blocks of vision". In: *Nature Reviews Neuroscience* 15 (8), pp. 507–519. ISSN: 1471-003X. DOI: 10.1038/nrn3783.

Evgenia, Kartsaki et al. (2023). *How does the inner retinal network shape the ganglion cells receptive field : a computational study*. DOI: https://hal.science/hal-04161982.

Fontaine, Bertrand, José Luis Peña, and Romain Brette (Apr. 2014). "Spike-Threshold Adaptation Predicted by Membrane Potential Dynamics In Vivo". In: *PLoS Computational Biology* 10 (4), e1003560. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003560.

Franke, Katrin et al. (Feb. 2017). "Inhibition decorrelates visual feature representations in the inner retina". In: *Nature* 542 (7642), pp. 439–444. ISSN: 14764687. DOI: 10.1038/nature21394.

Gao, Juan et al. (2009). "An oscillatory circuit underlying the detection of disruptions in temporally-periodic patterns". In: *Network: Computation in Neural Systems* 20 (2), pp. 106–135. ISSN: 0954898X. DOI: 10.1080/09548980902991705.

Garrido, Marta I. et al. (Mar. 2009). "The mismatch negativity: A review of underlying mechanisms". In: *Clinical Neurophysiology* 120 (3), pp. 453–463. ISSN: 13882457. DOI: 10.1016/j.clinph.2008.11.029.

Gersdorff, Henrique von and Gary Matthews (Mar. 1997). "Depletion and Replenishment of Vesicle Pools at a Ribbon-Type Synaptic Terminal". In: *The Journal of Neuroscience* 17 (6), pp. 1919–1927. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.17-06-01919.1997.

Ghanbari, Abed et al. (Sept. 2017). "Estimating short-term synaptic plasticity from pre- and postsynaptic spiking". In: *PLOS Computational Biology* 13 (9), e1005738. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005738.

Gollisch, Tim (Nov. 2013). "Features and functions of nonlinear spatial integration by retinal ganglion cells". In: *Journal of Physiology-Paris* 107 (5), pp. 338–348. ISSN: 09284257. DOI: 10.1016/j.jphysparis.2012.12.001.

Gollisch, Tim and Markus Meister (Jan. 2010). "Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina". In: *Neuron* 65 (2), pp. 150–164. ISSN: 08966273. DOI: 10.1016/j.neuron.2009.12.009.

Guo, Tianruo et al. (2014). "Understanding the Retina: A Review of Computational Models of the Retina from the Single Cell to the Network Level". In: *Critical Reviews in Biomedical Engineering* 42 (5), pp. 419–436. ISSN: 0278-940X. DOI: 10.1615/CritRevBiomedEng.2014011732.

Hansen, Nikolaus, Youhei Akimoto, and Petr Baudis (Feb. 2019). *CMA-ES/pycma on Github*. Zenodo, DOI:10.5281/zenodo.2559634. DOI: 10.5281/zenodo.2559634. URL: https://doi.org/10.5281/zenodo.2559634.

Hennig, Matthias H. (Apr. 2013). "Theoretical models of synaptic short term plasticity". In: *Frontiers in Computational Neuroscience* (APR 2013). ISSN: 16625188. DOI: 10.3389/fncom.2013.00045.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6, pp. 65–70.

Hosoya, Toshihiko, Stephen A. Baccus, and Markus Meister (July 2005). "Dynamic predictive coding by the retina". In: *Nature* 436 (7047), pp. 71–77. ISSN: 00280836. DOI: 10.1038/nature03689.

Huang, Xiaolin et al. (June 2022). "Visual Stimulation Induces Distinct Forms of Sensitization of On-Off Direction-Selective Ganglion Cell Responses in the Dorsal and Ventral Retina". In: *The Journal of Neuroscience* 42 (22), pp. 4449–4469. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.1391-21.2022.

Huang, Yanping and Rajesh P. N. Rao (Sept. 2011). "Predictive coding". In: *WIREs Cognitive Science* 2 (5), pp. 580–593. ISSN: 1939-5078. DOI: 10.1002/wcs.142.

Hubel, D. H. and T. N. Wiesel (Jan. 1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of Physiology* 160 (1), pp. 106–154. ISSN: 00223751. DOI: 10.1113/jphysiol.1962.sp006837.

Ikeda, Hisako and M. J. Wright (Oct. 1972). "The outer disinhibitory surround of the retinal ganglion cell receptive field". In: *The Journal of Physiology* 226 (2), pp. 511–544. ISSN: 00223751. DOI: 10.1113/jphysiol.1972.sp009996.

Jarsky, T. et al. (July 2011). "A Synaptic Mechanism for Retinal Adaptation to Luminance and Contrast". In: *Journal of Neuroscience* 31 (30), pp. 11003–11015. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2631-11.2011.

Johnston, Jamie and Leon Lagnado (2015). "General features of the retinal connectome determine the computation of motion anticipation". In: DOI: 10.7554/eLife.06250.001.

Johnston, Jamie et al. (June 2019). "A Retinal Circuit Generating a Dynamic Predictive Code for Oriented Features". In: *Neuron* 102 (6), 1211–1222.e3. ISSN: 08966273. DOI: 10.1016/j.neuron.2019.04.002.

Kartsaki, Evgenia (2022). "How specific classes of retinal cells contribute to vision : a computational model". In: *PhD Thesis*. URL: https://theses.hal.science/tel-03869570v3.

Kastner, David B. and Stephen A. Baccus (Oct. 2011). "Coordinated dynamic encoding in the retina using opposing forms of plasticity". In: *Nature Neuroscience* 14 (10), pp. 1317–1322. ISSN: 10976256. DOI: 10.1038/nn.2906.

— (Nov. 2013a). "Insights from the retina into the diverse and general computations of adaptation, detection, and prediction". In: *Current Opinion in Neurobiology* 25, pp. 63–69. ISSN: 09594388. DOI: 10.1016/j.conb.2013.11.012.

Kastner, David B and Stephen A Baccus (Apr. 2014). "Insights from the retina into the diverse and general computations of adaptation, detection, and prediction". In: *Current Opinion in Neurobiology* 25, pp. 63–69. ISSN: 09594388. DOI: 10.1016/j.conb.2013.11.012.

Kastner, David B. et al. (Aug. 2019). "Adaptation of Inhibition Mediates Retinal Sensitization". In: *Current Biology* 29 (16), 2640–2651.e4. ISSN: 09609822. DOI: 10.1016/j.cub.2019.06.081.

Kastner, David B. and Stephen A. Baccus (Aug. 2013b). "Spatial Segregation of Adaptation and Predictive Sensitization in Retinal Ganglion Cells". In: *Neuron* 79 (3), pp. 541–554. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.06.011.

Kremláček, Jan et al. (July 2016). "Visual mismatch negativity (vMMN): A review and meta-analysis of studies in psychiatric and neurological disorders". In: *Cortex* 80, pp. 76–112. ISSN: 00109452. DOI: 10.1016/j.cortex.2016.03.017.

Kuffler, Stephen W. (Jan. 1953). "Discharge Patterns and Functional Organization of Mammalian Retina". In: *Journal of Neurophysiology* 16 (1), pp. 37–68. ISSN: 0022-3077. DOI: 10.1152/jn.1953.16.1.37.

Lagnado, Leon and Frank Schmitz (Nov. 2015). "Ribbon Synapses and Visual Processing in the Retina". In: *Annual Review of Vision Science* 1 (1), pp. 235–262. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-082114-035709.

Lee, Dongsoo and Stephen A. Baccus Michael D. Menz (2020). "Representations of the amacrine cell population underlying retinal motion anticipation". In: *bioRxiv*.

Leonardo, Anthony and Markus Meister (2013). "Nonlinear dynamics support a linear population code in a retinal target-tracking circuit". In: *Journal of Neuroscience* 33 (43), pp. 16971–16982. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.2257-13.2013.

Li, G.-L., J. Vigh, and H. von Gersdorff (July 2007). "Short-Term Depression at the Reciprocal Synapses between a Retinal Bipolar Cell Terminal and Amacrine Cells".

In: *Journal of Neuroscience* 27 (28), pp. 7377–7385. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0410-07.2007.

Li, Jingcheng et al. (June 2017). "Primary Auditory Cortex is Required for Anticipatory Motor Response". In: *Cerebral Cortex* 27 (6), pp. 3254–3271. ISSN: 1047-3211. DOI: 10.1093/cercor/bhx079.

Lindner, Benjamin et al. (Feb. 2009). "Broadband Coding with Dynamic Synapses". In: *The Journal of Neuroscience* 29 (7), pp. 2076–2087. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3702-08.2009.

Liu, Belle et al. (Sept. 2021). "Predictive encoding of motion begins in the primate retina". In: *Nature Neuroscience* 24 (9), pp. 1280–1291. ISSN: 15461726. DOI: 10.1038/s41593-021-00899-1.

Manookin, Michael B. and Jonathan B. Demb (May 2006). "Presynaptic Mechanism for Slow Contrast Adaptation in Mammalian Retinal Ganglion Cells". In: *Neuron* 50 (3), pp. 453–464. ISSN: 08966273. DOI: 10.1016/j.neuron.2006.03.039.

Marre, Olivier et al. (2012). "Mapping a complete neural population in the retina". In: *Journal of Neuroscience* 32.43, pp. 14859–14873.

Masland, Richard H. (Oct. 2012a). "The Neuronal Organization of the Retina". In: *Neuron* 76 (2), pp. 266–280. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.10.002.

— (Jan. 2012b). "The tasks of amacrine cells". In: *Visual Neuroscience* 29 (1), pp. 3–9. ISSN: 0952-5238. DOI: 10.1017/S0952523811000344.

McAnany, J. Jason and Kenneth R. Alexander (Mar. 2009). "Is there an omitted stimulus response in the human cone flicker electroretinogram?" In: *Visual Neuroscience* 26 (2), pp. 189–194. ISSN: 09525238. DOI: 10.1017/S0952523808080991.

Millidge, Beren, Anil Seth, and Christopher L Buckley (July 2021). "Predictive Coding: a Theoretical and Experimental Review". In: *ArXiv*. URL: http://arxiv.org/abs/2107.12979.

Moraitis, Timoleon, Abu Sebastian, and Evangelos Eleftheriou (Sept. 2018). "The Role of Short-Term Plasticity in Neuromorphic Learning: Learning from the Timing of Rate-Varying Events with Fatiguing Spike-Timing-Dependent Plasticity". In: *IEEE Nanotechnology Magazine* 12 (3), pp. 45–53. ISSN: 1932-4510. DOI: 10.1109/MNANO.2018.2845479.

Mostofi, Naghmeh et al. (Oct. 2020). "Spatiotemporal Content of Saccade Transients". In: *Current Biology* 30 (20), 3999–4008.e2. ISSN: 09609822. DOI: 10.1016/j.cub.2020.07.085.

Münch, Thomas A et al. (Oct. 2009). "Approach sensitivity in the retina processed by a multifunctional neural circuit". In: *Nature Neuroscience* 12 (10), pp. 1308–1316. ISSN: 1097-6256. DOI: 10.1038/nn.2389.

Nikolaev, Anton et al. (July 2013). "Synaptic mechanisms of adaptation and sensitization in the retina". In: *Nature Neuroscience* 16 (7), pp. 934–941. ISSN: 10976256. DOI: 10.1038/nn.3408.

Näätänen, R., A.W.K. Gaillard, and S. Mäntysalo (July 1978). "Early selective-attention effect on evoked potential reinterpreted". In: *Acta Psychologica* 42 (4), pp. 313–329. ISSN: 00016918. DOI: 10.1016/0001-6918(78)90006-9.

Näätänen, R. et al. (Sept. 1993). "Attention and mismatch negativity". In: *Psychophysiology* 30 (5), pp. 436–450. ISSN: 0048-5772. DOI: 10.1111/j.1469-8986.1993.tb02067.x.

Odermatt, B. and L. Lagnado (2009). "Ribbon Synapses". In: Elsevier, pp. 373–381. DOI: 10.1016/B978-008045046-9.00923-2.

Oesch, Nicholas W. and Jeffrey S. Diamond (Dec. 2011). "Ribbon synapses compute temporal contrast and encode luminance in retinal rod bipolar cells". In: *Nature Neuroscience* 14 (12), pp. 1555–1561. ISSN: 10976256. DOI: 10.1038/nn.2945.

— (May 2019). "Synaptic inhibition tunes contrast computation in the retina". In: *Visual Neuroscience* 36, E006. ISSN: 0952-5238. DOI: 10.1017/S095252381900004X.

Ooyen, Arjen van (June 2011). "Using theoretical models to analyse neural development". In: *Nature Reviews Neuroscience* 12 (6), pp. 311–326. ISSN: 1471-003X. DOI: 10.1038/nrn3031.

Ozuysal, Yusuf and Stephen A. Baccus (Mar. 2012). "Linking the Computational Structure of Variance Adaptation to Biophysical Mechanisms". In: *Neuron* 73 (5), pp. 1002–1015. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.12.029.

Palmer, Stephanie E. et al. (June 2015). "Predictive information in a sensory population". In: *Proceedings of the National Academy of Sciences* 112 (22), pp. 6908–6913. ISSN: 0027-8424. DOI: 10.1073/pnas.1506855112.

Pillow, Jonathan W. et al. (Aug. 2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population". In: *Nature* 454 (7207), pp. 995–999. ISSN: 0028-0836. DOI: 10.1038/nature07140.

Rao, Rajesh P. N. and Dana H. Ballard (Jan. 1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature Neuroscience* 2 (1), pp. 79–87. ISSN: 1097-6256. DOI: 10.1038/4580.

Roberts, Paul A. et al. (July 2016). "Mathematical and computational models of the retina in health, development and disease". In: *Progress in Retinal and Eye Research* 53, pp. 48–69. ISSN: 13509462. DOI: 10.1016/j.preteyeres.2016.04.001.

Saigusa, Tetsu et al. (Jan. 2008). "Amoebae Anticipate Periodic Events". In: *Physical Review Letters* 100 (1), p. 018101. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.100.018101.

Schröder, Cornelius et al. (2020). "System identification with biophysical constraints: A circuit model of the inner retina". In: *bioRxiv*. ISSN: 2692-8205. DOI: 10.1101/2020.06.16.154203.

Schwartz, Greg and Michael J. Berry (Apr. 2008). "Sophisticated temporal pattern recognition in retinal ganglion cells". In: *Journal of Neurophysiology* 99 (4), pp. 1787–1798. ISSN: 00223077. DOI: 10.1152/jn.01025.2007.

Schwartz, Greg et al. (May 2007a). "Detection and prediction of periodic patterns by the retina". In: *Nature Neuroscience* 10 (5), pp. 552–554. ISSN: 10976256. DOI: 10.1038/nn1887.

Schwartz, Greg et al. (Sept. 2007b). "Synchronized Firing among Retinal Ganglion Cells Signals Motion Reversal". In: *Neuron* 55 (6), pp. 958–969. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.07.042.

— (Sept. 2007c). "Synchronized Firing among Retinal Ganglion Cells Signals Motion Reversal". In: *Neuron* 55 (6), pp. 958–969. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.07.042.

Schwartz, Gregory W et al. (Nov. 2012). "The spatial structure of a nonlinear receptive field". In: *Nature Neuroscience* 15 (11), pp. 1572–1580. ISSN: 1097-6256. DOI: 10.1038/nn.3225.

Shapley, Robert and Christina Enroth-Cugell (Jan. 1984). "Chapter 9 Visual adaptation and retinal gain controls". In: *Progress in Retinal Research* 3, pp. 263–346. ISSN: 02784327. DOI: 10.1016/0278-4327(84)90011-7.

Sidney P. Kuo, Gregory W. Schwartz and Fred Rieke (Apr. 2016). "Nonlinear Spatiotemporal Integration by Electrical and Chemical Synapses in the Retina". In: *Neuron* 90 (2), pp. 320–332. ISSN: 08966273. DOI: 10.1016/j.neuron.2016.03.012.

Silveira, Rava Azeredo da and Botond Roska (Oct. 2011). "Cell Types, Circuits, Computation". In: *Current Opinion in Neurobiology* 21 (5), pp. 664–671. ISSN: 09594388. DOI: 10.1016/j.conb.2011.05.007.

Singer, Joshua H. and Jeffrey S. Diamond (May 2006). "Vesicle depletion and synaptic depression at a mammalian ribbon synapse". In: *Journal of Neurophysiology* 95 (5), pp. 3191–3198. ISSN: 00223077. DOI: 10.1152/jn.01309.2005.

Souihel, Selma and Bruno Cessac (Dec. 2021). "On the potential role of lateral connectivity in retinal anticipation". In: *Journal of Mathematical Neuroscience* 11 (1). ISSN: 21908567. DOI: 10.1186/s13408-020-00101-z.

Srinivasan, Mandyam Veerambudi, Simon Barry Laughlin, and A. Dubs (Nov. 1982). "Predictive coding: a fresh view of inhibition in the retina". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 216 (1205), pp. 427–459. ISSN: 0080-4649. DOI: 10.1098/rspb.1982.0085.

Sterratt, David et al. (June 2011). *Principles of Computational Modelling in Neuroscience*. Cambridge University Press. ISBN: 9780521877954. DOI: 10.1017/CBO9780511975899.

Suh, Bongsoo and Stephen A. Baccus (Oct. 2014). "Building Blocks of Temporal Filters in Retinal Synapses". In: *PLoS Biology* 12 (10), e1001973. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001973.

Tanaka, Hidenori et al. (Dec. 2019). "From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction". In: *ArXiv*. URL: http://arxiv.org/abs/1912.06207.

Tsodyks, Misha and Henry Markram (1997). "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability." In: *Proceedings of the National Academy of Sciences of the United States of America* 94 2, pp. 719–23.

Tsodyks, Misha, Klaus Pawelzik, and Henry Markram (May 1998). "Neural Networks with Dynamic Synapses". In: *Neural Computation* 10 (4), pp. 821–835. ISSN: 0899-7667. DOI: 10.1162/089976698300017502.

Ulanovsky, Nachum, Liora Las, and Israel Nelken (Apr. 2003). "Processing of low-probability sounds by cortical neurons". In: *Nature Neuroscience* 6 (4), pp. 391–398. ISSN: 10976256. DOI: 10.1038/nn1032.

Vaney, David I. (Jan. 1990). "Chapter 2 The mosaic of amacrine cells in the mammalian retina". In: *Progress in Retinal Research* 9, pp. 49–100. ISSN: 02784327. DOI: 10.1016/0278-4327(90)90004-2.

Vickers, Evan et al. (Aug. 2012). "Paired-pulse plasticity in the strength and latency of light-evoked lateral inhibition to retinal bipolar cell terminals". In: *Journal of Neuroscience* 32 (34), pp. 11688–11699. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0547-12.2012.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wehr, Michael and Anthony M. Zador (Nov. 2003). "Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex". In: *Nature* 426 (6965), pp. 442–446. ISSN: 0028-0836. DOI: 10.1038/nature02116.

— (Aug. 2005). "Synaptic mechanisms of forward suppression in rat auditory cortex". In: *Neuron* 47 (3), pp. 437–445. ISSN: 08966273. DOI: 10.1016/j.neuron.2005.06.009.

Wei, Wei (2018). "Neural Mechanisms of Motion Processing in the Mammalian Retina". In: 22 (1). DOI: 10.1146/annurev-vision-091517. URL: https://doi.org/10.1146/annurev-vision-091517-.

Weidmann, M. D (2009). "Exploring the Cellular Mechanism of the Omitted Stimulus Response." In: *Thesis*.

Werner, Birgit, Paul B. Cook, and Christopher L. Passaglia (Aug. 2008). "Complex temporal response patterns with a simple retinal circuit". In: *Journal of Neurophysiology* 100 (2), pp. 1087–1097. ISSN: 00223077. DOI: 10.1152/jn.90527.2008.

Wohrer, Adrien and Pierre Kornprobst (Apr. 2009). "Virtual Retina: A biological retina model and simulator, with contrast gain control". In: *Journal of Computational Neuroscience* 26 (2), pp. 219–249. ISSN: 0929-5313. DOI: 10.1007/s10827-008-0108-4.

Wässle, Heinz et al. (2009). "Glycinergic transmission in the mammalian retina". In: *Frontiers in Molecular Neuroscience* 2. ISSN: 1662-5099. DOI: 10.3389/neuro.02.006.2009. URL: https://www.frontiersin.org/articles/10.3389/neuro.02.006.2009.

Yger, Pierre et al. (2018). "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo". In: *Elife* 7, e34518.

Zaghloul, Kareem A. et al. (June 2007). "Functional Circuitry for Peripheral Suppression in Mammalian Y-Type Retinal Ganglion Cells". In: *Journal of Neurophysiology* 97 (6), pp. 4327–4340. ISSN: 0022-3077. DOI: 10.1152/jn.01091.2006.

Zhang, Danke, Si Wu, and Malte J. Rasch (Feb. 2015). "Circuit motifs for contrast-adaptive differentiation in early sensory systems: The role of presynaptic inhibition and short-term plasticity". In: *PLoS ONE* 10 (2). ISSN: 19326203. DOI: 10.1371/journal.pone.0118125.

Zhou, Y. et al. (July 2012). "Generation of Spike Latency Tuning by Thalamocortical Circuits in Auditory Cortex". In: *Journal of Neuroscience* 32 (29), pp. 9969–9980. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.1384-12.2012.

Ölveczky, Bence P., Stephen A. Baccus, and Markus Meister (May 2003). "Segregation of object and background motion in the retina". In: *Nature* 423 (6938), pp. 401–408. ISSN: 0028-0836. DOI: 10.1038/nature01652.