



HAL
open science

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

Simone Ebert, Thomas Buffet, B.Semihcan Sermet, Olivier Marre, Bruno Cessac

► To cite this version:

Simone Ebert, Thomas Buffet, B.Semihcan Sermet, Olivier Marre, Bruno Cessac. Temporal pattern recognition in retinal ganglion cells is mediated by dynamical inhibitory synapses. *Nature Communications*, 2024, 15 (1), pp.6118. 10.1101/2023.01.12.523643 . hal-03939794v2

HAL Id: hal-03939794

<https://inria.hal.science/hal-03939794v2>

Submitted on 17 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

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

Simone Ebert ^{* †1,2,3}, Thomas Buffet ^{3,*}, B.Semihcan Sermet ^{‡3}, Olivier Marre ^{§3}, and Bruno Cessac ^{§,1,2}

¹Université Côte d'Azur, Inria Biovision team, 2004 Rte des Lucioles, 06902 Valbonne, France

²Institute for Modeling in Neuroscience and Cognition (NeuroMod), Université Côte d'Azur, 28 Av. Valrose, 06100 Nice, France

³Sorbonne Université, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, F-75012 Paris, France

Abstract

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

A fundamental task for the brain is to generate predictions of future sensory inputs, and signal errors in these predictions. Many neurons have been shown to signal omitted stimuli during periodic stimulation, even in the retina. However, the mechanisms of this error signaling are unclear. Here we show that depressing inhibitory synapses shape the timing of the response to an omitted stimulus in the retina. While ganglion cells, the retinal output, responded to an omitted flash with a constant latency over many frequencies of the flash sequence, we found that this was not the case once inhibition was blocked. We built 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 a sufficient amount of flashes in the stimulus, which we could confirm experimentally. Depressing inhibitory synapses could thus be a key component to generate the predictive responses observed in the retina, and potentially in many brain areas.

*These authors contributed equally to this work.

†corresponding author: simoneebert166@gmail.com

‡Present address: Netherlands Institute for Neuroscience Meibergdreef 47, 1105 BA Amsterdam

§These authors contributed equally to this work.

18 1 Introduction

19 A long-standing hypothesis is that visual neurons do not signal the visual scene
 20 per se, but rather surprising events, eg. mismatches between observation and
 21 expectation formed by previous inputs^{Barlow1961}. It has been observed in a
 22 number of sensory modalities that neurons strongly respond when a sequence of
 23 repetitive stimuli is unexpectedly interrupted^{Ulanovsky2003, Bullock1994, McAnany2009}.

24 In the retina, this phenomenon has been coined the Omitted Stimulus Re-
 25 sponse (OSR)^{Schwartz2007a}. When a periodic sequence of flashes suddenly ends,
 26 some ganglion cells emit a large response. Interestingly, the latency of this re-
 27 sponse shifts with the period of the flash sequence, so that the ganglion cell
 28 responds to the omitted flash with a constant latency. This suggests that the
 29 retina forms predictions of observed patterns, and responds to a violation of
 30 its internal expectation. Several studies have proposed potential mechanisms
 31 that may underlie this phenomenon, such as oscillatory activity in bipolar
 32 cells^{Gao2009}, summation of parallel pathways with different response polarities
 33 ^{Werner2008} or response kinetics that respond early and late^{Tanaka2019}. The
 34 fact that there is a rebound response at the end of the flash sequence can be
 35 simply explained by a model with a biphasic filter^{Werner2008, Tanaka2019}. Pre-
 36 vious work has shown that this rebound response depends on the ON bipolar
 37 cell pathway^{Schwartz2008}. However, how the latency changes with the period
 38 of the sequence, such that the response to the Omitted stimulus has a constant
 39 latency, remains unclear. None of the mechanisms proposed so far could be
 40 experimentally proven. Thus, the mechanisms by which the retina achieves this
 41 latency shift remain unclear and debated.

42 Here we investigated how inhibitory amacrine cells affect the OSR and
 43 showed that depression in inhibitory synapses can account for this character-
 44 istic latency shift. To this end, we performed electrophysiological recordings
 45 of retinal ganglion cells and found that blocking inhibitory transmission from
 46 glycinergic amacrine cells selectively abolished the predictive latency shift of the
 47 OSR. We then showed that the latency shift is specific to the periodicity of the
 48 stimulus and is not a consequence of stimulus duration or luminance. To better
 49 understand how glycinergic inhibition impacts the latency of the OSR, we de-
 50 veloped a circuit model equipped with a glycinergic amacrine cell. This model
 51 reproduced the latency shift of the OSR when the glycinergic synapse showed
 52 short-term depression, thereby adjusting its weight to the stimulus frequency.
 53 In addition, our model generated several predictions about the OSR, which we
 54 could confirm in experiments. The latency shift that is characteristic of the
 55 OSR is thus due to a depressing inhibitory synapse whose weight is changed
 56 by the stimulus frequency. For low-frequency sequences, the synaptic weight is
 57 large and this increases the latency of the response, while for high-frequency
 58 stimuli, the weight is low due to depression, and the latency is only shifted by
 59 a small amount. Our results suggest a generic circuit to generate responses to
 60 surprise that could be potentially implemented in several brain areas.

2 Results

2.1 ON 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^{Marre2012} (Figure 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, 1 B). Previous studies observed an OSR in both ON and OFF retinal ganglion cells. Given that previous pharmacological experiments revealed that ON inputs are necessary to drive an OSR^{Schwartz2007a}, 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 ($n = 143$ cells). A subset of cells with response after stimulus end shifted the response latency according to the stimulus frequency (27 %, $n = 40$, fig. 1 D, see Methods). They exhibited an "Omitted Stimulus Response" (OSR), as it was described in^{Schwartz2007a}: 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 (Figure 1C,D), such that the latency of the response to the omitted stimulus is constant (Figure 1E). This indicates that a subset of retinal ganglion cells have a temporal expectation of when the next flash should have occurred and shift their response latency accordingly. This is illustrated in Figure 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 (Figure 1G). In the following we will refer to this specific relation as "latency shift".

To see whether there are ON cells showing this latency shift can be related to a known RGC type, we clustered them according to their response to a full-field 'chirp' stimulation, which reveals response polarity, kinetics and preference to temporal frequencies and contrasts^{Baden2016}. We identified roughly 4 different response types which we classified as "ON-transient", "ON-transient-sustained", "ON-sustained" and "ON-OFF" (see Figure S1). This suggests that the OSR is not a response behavior specific to a given cell type but rather exhibited by a broader range of cells. While some cells showed a response to the onset of the flash sequence and/or showed small responses to flashes, the majority of cells did not respond at all to the flashed stimuli (see Figure S1C). Additionally, the total latency between the last flash and the OSR varied substantially between cells, ranging between 142 and 483 ms for the fastest frequency (see Figure S2, right). Even though the omitted stimulus responses occur late for some cells, the latency difference of 120 ms between the fastest and the lowest frequency is very close to the difference in period (110 ms) of these stimuli. For further analysis, we chose to focus on the omitted stimulus response peak characterized by this latency shift, pooling together ON cells of different types.

105 **2.2 Amacrine cells are required for the latency shift in the** 106 **Omitted Stimulus Response**

107 It remains unclear how the retinal circuit generates an OSR. Previous studies
108 have shown that the ON bipolar cell pathway is necessary to have a response
109 per se^{Schwartz2007a}, but the components of the retinal circuit needed for the la-
110 tency shift are yet to be determined. We hypothesized that the inhibitory cells
111 are responsible for the shift in latency, as inhibition has been shown to shift la-
112 tency in various neuronal circuits^{Tsodyks1998, Wehr2003, Wehr2005, Fontaine2014}.
113 Amacrine cells are the main class of inhibitory interneurons in the mouse retina.
114 To investigate this further, we blocked glycinergic transmission using strychnine
115 ($2\mu M$) and recorded the spike responses of retinal ganglion cells to flash trains of
116 varying frequencies (see Methods). As glycinergic transmission is only employed
117 by certain classes of inhibitory amacrine cells in the mouse retina^{Waesle2009}
118 this blocks only a subset of amacrine cells. In the following, we analyze the
119 effect of strychnine on those ON retinal ganglion cells which exhibited a OSR
120 with latency shift in control conditions.

121 While the response after the sequence end remained after strychnine applica-
122 tion, we observed that the slope between response latency and stimulus period
123 decreased from an average of 1.13 ± 0.08 in the control condition to 0.07 ± 0.18
124 after strychnine was added (Figure 1 F and G, mean \pm SEM, $n = 12$, see
125 Methods). While the OSR occurred at roughly the same time in the high-
126 est frequency tested, the peak was significantly advanced after low-frequency
127 flashes compared to the control condition (Figure 1 D). As a consequence, the
128 OSR peak did not have a constant latency relative to the omitted stimulus after
129 strychnine was added (Figure 1 E). These results demonstrate that glycinergic
130 amacrine cells are a key contributor to the OSR latency shift. Although they
131 do not generate the response alone, they are crucial for the peak latency shift,
132 which is a hallmark of the OSR. We also looked at the onset time of the OSR
133 response but found that the relation between onset time and stimulus period
134 is highly variable with an average slope of 0.39 (see Figure S3). The effect of
135 glycine application on the onset of the OSR is also highly variable, resulting
136 in a non-significant difference in the slope in control and strychnine conditions.
137 Overall, these results suggest that spiking onset does not contain predictive in-
138 formation about the arrival of the next flash as its timing does not shift in a
139 1-to-1 manner with the period of the flash train.

140 **2.3 Stimulus Luminance has no effect on response latency** 141 **to opposite polarity stimuli**

142 Previous works have shown that overall changes in luminance or contrast can in-
143 duce temporal adaptation that could affect ganglion cell responses^{Oesch2011, Oesch2019}.
144 To further understand the mechanisms of the OSR, and constrain a model that
145 would reproduce the experimental results, we investigated whether the latency
146 shift of the OSR is really triggered only by the periodicity of the stimulus or if it
147 is affected by other stimulus features. Periodic flash trains of different frequen-

148 cies are not only characterized by periodicity but also vary in terms of average
 149 luminance and duration. As flashes have the same duration of 40 ms across all
 150 frequencies but occur in different windows of time, the average illumination and
 151 stimulus duration is lower for high than for low frequencies.

152 To investigate whether the latency shift of the OSR is simply a consequence
 153 of duration or different levels of average luminance, we presented two sets of
 154 stimuli: non-flashing dark step-stimuli of different lengths corresponding to the
 155 duration of flash trains presented in the control experiments, and non-flashing
 156 dark step-stimuli with different luminance levels corresponding to the average
 157 luminance during flash trains (Figure 2A,B upper panel). Both stimuli evoked
 158 an offset response after stimulus in cells that show an OSR to periodic flashes
 159 (Figure 2A,B lower panel), but the latency of this response was nearly equal
 160 across all luminance levels and durations tested (Figure 2C). This suggests that
 161 the latency shift in the OSR is specific to periodic stimuli and that luminance
 162 or duration variations are not sufficient.

163 We then tested two further modifications of our periodic flash trains in which
 164 mean luminance was kept constant across stimulus frequencies (Figure 2D).
 165 The first modification was to keep the luminance steady by increasing light
 166 intensity in between the dark 40ms flashes, which yielded an OSR with latency
 167 shift (Figure 2E, F, G, in blue). This suggests that overall contrast change is
 168 not necessary for the latency shift. In the second modification, we kept the
 169 average luminance constant by adjusting the duration of the flashes to half of
 170 the stimulus period. Surprisingly, we found that changing the duration of the
 171 flashes maintains the OSR peak but erases the latency scaling (Figure 2E, F, G,
 172 in green). These experimental results suggest that the retina detects periodic
 173 patterns but can only form temporal predictions for specific input patterns (see
 174 also [Schwartz2008](#)).

175 Surprisingly, the responses to the stimulus with increased light intensity be-
 176 tween the flashes were almost identical to the responses in the control stimulus.
 177 To both the control stimulus and the modification, only 4 out of 20 ON cells
 178 with OSR showed a response during the stimulus, mostly for low frequencies.
 179 We did not identify any quantifiable differences in the OSR response in these
 180 two cells compared to the remaining 16 cells that did not respond during the
 181 stimulus. Our hypothesis is that the transient increases are either too fast (for
 182 high frequencies) or too small (for low frequencies), to evoke a response above
 183 threshold.

184 **2.4 A circuit model with depressing synapses in inhibitory** 185 **glycinergic amacrine cells explains the latency shift**

186 Our experimental results provided compelling evidence that glycinergic amacrine
 187 cells provide inputs which are required to achieve the latency shift of the OSR
 188 and that the latency shift is specific to some stimulus patterns. However, it
 189 remained unclear how the retinal circuit achieves such selectivity in a glycin-
 190 ergic dependent way. We developed a mechanistic model in which we explicitly

191 simulated evoked inputs from glycinergic amacrine cells to understand their role
 192 in the latency shift of the OSR.

193 We focused on ON ganglion cells, and equipped our model with two ON
 194 inputs, one being excitatory E^{ON} and mimicking ON bipolar cell input, and
 195 one being inhibitory I^{ON} , conveying broad delayed inhibition. This generates a
 196 biphasic response profile needed for a rebound response to dark flashes **Werner2008**.
 197 This delayed inhibitory input summarizes the influence of various inhibitory
 198 pathways (horizontal cells, GABAergic amacrine cells, ON glycinergic amacrine
 199 cells) that generate the biphasic response profile (see Discussion). In addition,
 200 we explicitly included a glycinergic amacrine cell with an OFF polarity, I_{Gly}^{OFF} ,
 201 in order to provide inhibition to dark stimuli. All three units receive inputs
 202 from the outer plexiform layer (OPL), which is passed through a nonlinearity
 203 to simulate nonlinear processing in the first synaptic retinal layer. All 3 units
 204 synapse onto a ganglion cell G, where synaptic inputs to G are again passed
 205 through a nonlinearity (Figure 3 A, see Methods for details).

206 A characteristic feature of the Omitted Stimulus Response is that the latency
 207 shifts by the same amount as the stimulus period. In other contexts, it has been
 208 shown that the relative strengths of excitation and inhibition can determine the
 209 latency of the response **Tsodyks1998, Wehr2003, Wehr2005, Fontaine2014**. We rea-
 210 soned that the effective strength of the inhibitory input should thus depend
 211 on the frequency of the flash sequence. This can be achieved with a dynamic
 212 synapse, i.e. a synapse whose strength varies with the frequency of the stimu-
 213 lation.

214 Several previous reports have shown that inhibitory synapses can be depress-
 215 ing, i.e. have a decreasing weight depending on previous inputs **Li2007, Vickers2012, Nikolaev2013, Kastner2019, Hua**
 216 It has been hypothesized that the variation in synaptic strength results from
 217 varying availability of vesicles in the readily releasable vesicle pool, which gets
 218 gradually depleted upon persistent inputs **Singer2006, Burrone2000, Oesch2011**. In
 219 the retina, this has been modeled via dynamical systems of vesicle pools, where
 220 the value of one of the variables directly serves as input to the postsynaptic
 221 cell **Ozuysal2012, Schroeder2020**. To keep our model as simple as possible, we
 222 modeled the glycinergic synapse as a depressing synapse using one kinetic equa-
 223 tion to simulate synaptic vesicle occupancy, similar to cortical models of short-
 224 term plasticity **TsodyksMarkram1997, Zhou2012, Henning2013**, but replaced spiking
 225 inputs by continuous presynaptic voltage. The vesicle occupancy then scales
 226 the output of the cell (see Methods).

227 This mechanistic model allowed us to reproduce the main properties of the
 228 recorded ganglion cells. Thanks to excitatory and inhibitory ON inputs, it has
 229 an ON biphasic impulse response. It also responds with a peak at the end of
 230 dark periodic flash stimuli, due to delayed disinhibition, because inhibition has
 231 a slower temporal filter than excitation (Figure 3 B).

232 Our model also successfully simulated an OSR whose latency with respect
 233 to the last flash increased with the stimulus period with a slope around 1, and
 234 thus with a constant latency with respect to the omitted stimulus (Figure 3 B
 235 and C, black line).

236 We next simulated the experimental effect of strychnine with the model by

237 removing the glycinergic amacrine cell input. Since the ON inhibitory cell of
 238 our model also includes the effect of ON glycinergic amacrine cells, we simulta-
 239 neously decreased the weight of the ON inhibitory input (note however that our
 240 results did not depend on that additional modification, see Discussion). Fig-
 241 ure 3 B and C (red) show that the model replicates the experimental effect of
 242 strychnine. The slope of the latency shift decreases from 1.16 to 0.34 when I_{Gly}^{OFF}
 243 is removed from the circuit.

244 To test whether the model achieves the latency shift of the response peak
 245 thanks to the dynamical synapse, we simulated the response of the model while
 246 keeping the occupancy of the glycinergic amacrine cell synapse constant at 1 (
 247 Figure 4 A). The latency increased in all frequencies simulated, but the slope of
 248 the latency shift decreased to 0.32 (see Figure 4 B-D). Without the dynamical
 249 synapse, the linear filters in the distinct pathways of our model can be summed
 250 together to one ON biphasic filter profile. Followed by the non-linearity placed
 251 in the ganglion cell our model is equivalent to a LN-model and very similar to the
 252 one proposed by [Werner2008](#). However, these models, which are somewhat sim-
 253 plification of ours (they have no dynamical synapse) are not able to predict the
 254 latency shift with a slope of 1. The dynamical synapse is thus essential to achieve
 255 the latency shift of the response with a slope of 1 observed experimentally. Note
 256 that there was also an overall increase in latency which can be explained by the
 257 fact that the fixed synapse is stronger than the dynamic counterpart, since it is
 258 never at full occupancy when stimulated due to depression.

259 Additionally, we observed that the OSR amplitude decreased when stimulus
 260 frequency increased (Figure S4B). There are no significant changes in this rela-
 261 tion after strychnine application, albeit the curve seems slightly flattened. Our
 262 model captures this relation accurately (Figure S4A).

263 Finally, we simulated the model's response to modified flash trains with con-
 264 stant average luminance across periods (Figure 2D). As observed in experiments,
 265 the model maintains an OSR with latency shift to periodic flash trains with in-
 266 creased light intensity in interflash-intervals but loses this scaling in response to
 267 flash trains where the duration of the flashes is adjusted (see Figure S5).

268 The model thus captures experimental findings accurately.

269 **2.5 The depressing inhibitory synapse induces a latency** 270 **shift**

271 To understand how the model can account for the latency shift in the OSR,
 272 it is helpful to look at the temporal evolution of its internal variables. The
 273 ON excitatory input evokes a hyperpolarization in response to dark flashes,
 274 and cancels the depolarization evoked by the slightly slower ON inhibitory cells
 275 during the flash sequence. At the end of the flash sequence, due to differing time
 276 constants, there is a time window where depolarization due to the inhibitory
 277 delayed cell exceeds the hyperpolarization due to the bipolar cell, and triggers a
 278 spiking response of the retinal ganglion cell (Figure 5 A). This is similar to many
 279 classical rebound responses recorded experimentally, and it can be predicted
 280 with a biphasic filter. But by itself, this biphasic filter would not predict the

281 latency shift as observed experimentally (see above). As we will describe in the
 282 following paragraph, this latency shift can be explained thanks to the specific
 283 effect of the glycinergic amacrine cell equipped with a depressing synapse.

284 Since this glycinergic amacrine cell is an OFF cell, it inhibits the ganglion cell
 285 in response to dark flashes. This has the effect of delaying the spiking response
 286 at the end of the flash sequence and to increase its latency (Figure 5 B, black
 287 compared to grey). The latency of the response is then shifted for different
 288 stimulus frequencies because the depressing synapse changes the strength of the
 289 glycinergic inhibition and the time scale in the response.

290 In this synapse, the vesicle occupancy represents the amount of synaptic
 291 depression and decreases when stimulation starts (Figure 6 A, 3rd row). It then
 292 reduces the current input from I_{Gly}^{OFF} to the ganglion cell (Figure 6 A, 4th row).
 293 This reduction of inhibition shifts the OSR towards an earlier response, reducing
 294 its latency (Figure 6 A, 5th row).

295 Fast-frequency stimuli cause stronger depression, reducing the I_{Gly}^{OFF} current
 296 input by about 30 %. This has a strong impact on the latency, which is more
 297 than 100 ms shorter when the synapse is depressed. In contrast, slow-frequency
 298 stimuli cause only weak depression, reducing I_{Gly}^{OFF} by about only 10 %. The
 299 latency was thus only slightly reduced in that case (compare Figure 6 A and
 300 B).

301 In summary, the steady-state vesicle occupancy of the synapse is determined
 302 by the stimulus frequency (Figure 6 C and D). The vesicle occupancy can then
 303 reduce the inhibitory current, yielding a large reduction for fast inputs and a
 304 small reduction for slow inputs (Figure 6 E). As a result, vesicle occupancy acts
 305 as a scaling factor to tune the inhibitory current input and thereby shifts the
 306 response latency based on stimulus frequency. This explains how the latency
 307 shift observed experimentally is achieved via glycinergic inputs.

308 2.6 The depressing inhibitory synapse predicts other fea- 309 tures of the Omitted Stimulus Response

310 Can our model give other predictions about the OSR ? Since the glycinergic
 311 inhibitory synapse is more depressed at high frequency, the OSR is less inhibited,
 312 and its amplitude is thus stronger compared to low frequencies (compare Figure
 313 6 A and B, 5th row). This trend was also observed in our experiments. Both
 314 simulated and experimental amplitudes showed a negative correlation between
 315 the response amplitude and the stimulus period (-0.87 ± 0.02 , mean \pm SEM
 316 $n = 14$) in data and as well in simulations (-0.93).

317 An important consequence of our depressing synapse model is that it needs
 318 several flashes to reach the steady state in the vesicle occupancy. If we shorten
 319 the flash sequence, the vesicle occupancy will barely reach that steady state, and
 320 this should have predictable consequences on response amplitude and latency.
 321 We simulated the response to long flash trains consisting of 12 flashes (as in the
 322 experiments and simulations above) and shorter sequences of only 5 flashes.

323 Our simulations predicted that the amplitude of the OSR decreases when the
 324 stimulus contains only 5 flashes. We tested that in experiments and found the

325 same tendency: the OSR amplitude was smaller in all but the lowest frequency
 326 tested (see Figure 7 A)

327 Another model prediction was that the slope of the relation between OSR la-
 328 tency and stimulus period should decrease for shorter flash trains, reaching only
 329 a value of 0.67 for 5 flashes, compared to 1.16 when 12 flashes were presented
 330 (see Figure 7 B, left). In our model, this is a consequence of the dynamics of
 331 the depressing synapse.

332 In the 5-flashes scenario, our model predicts that the stimulus is too short
 333 for the synapse to reach a reliable steady-state occupancy that is not perturbed
 334 by the oscillations due to the flashes. (Figure 7 C). This perturbs the response
 335 latency scaling and changes the slope of the relation between latency shift and
 336 stimulus period (Figure 7 D) for short flash trains.

337 In our experiments, while there was no difference for low-frequency stimuli,
 338 the absolute latency of the OSR was much larger after 5 flashes than after 12
 339 when the stimulus frequency was high (see Figure 7 B). This change in latency
 340 led to a reduction of the mean \pm SEM slope value from 0.84 ± 0.02 to 0.69 ± 0.04
 341 in experiments, consistent with the model prediction. These agreements provide
 342 further evidence for the validity of our model, and for the key role of a depressing
 343 inhibitory synapse.

344 3 Discussion

345 The Omitted Stimulus Response is an example of sophisticated feature detection
 346 that takes place already in the retina. This phenomenon implies that retinal
 347 ganglion cells can carry a dynamic prediction of their future visual input with
 348 high temporal precision, and selectively respond when this prediction is not
 349 matched (Figure 8A). Although high-contrast full-field periodic flashes are arti-
 350 ficial stimuli that are unlikely to occur in natural scenes, they isolate temporal
 351 aspects of visual input patterns. The underlying mechanisms of temporal pro-
 352 cessing in the rather artificial OSR may be embedded in more complex networks
 353 detecting spatiotemporal patterns in more realistic scenes.

354 With this work, we provide evidence that the latency shift of the OSR,
 355 which allows a constant latency relative to the omitted stimulus, is generated
 356 by inhibition from glycinergic amacrine cells and triggered by the periodicity of
 357 the stimulus. Using computational modeling, we show how inhibition can enable
 358 retinal ganglion cells to respond to the missing flash at the end of a sequence.
 359 Short-term depression in inhibitory synapses allows shifting the latency of this
 360 response (Figure 8B,C).

361 Several theoretical models have been proposed to elucidate the mechanisms
 362 behind the Omitted Stimulus Response. Werner and Passaglia^{Werner2008} pro-
 363 posed a dual LN-model with biphasic ON-OFF pathway interactions, which
 364 accurately captures the response peak after the stimulus ends via the rebound
 365 phase of the pathway selective to the opposite polarity than the stimulus. How-
 366 ever, it fails to shift the peak latency as a function of the stimulus period with a
 367 slope of 1, which is a defining feature of the OSR. This slope value is necessary

368 to have a response of constant latency with respect to the omitted stimulus.
 369 When removing the depressing synapse, our model is amenable to a biphasic
 370 LN model, since the responses of our intermediate units are then linear and
 371 could be represented by a single linear filter as well. Our model then simu-
 372 lates the OSR in the same manner as this previous study, but the depressing
 373 inhibitory synapse was necessary to obtain the slope of 1, which is the signature
 374 of a predictive latency shift.

375 Gao and Berry^{Gao2009} proposed intrinsic oscillatory activity in ON bipolar
 376 cells that evoked a latency shift via resonance tuned to the stimulus frequency.
 377 However, such oscillatory activity was not found in bipolar cells^{Deshmukh2019}.
 378 Further, our experiments show that glycinergic amacrine inhibition is necessary
 379 for the latency scaling of the OSR, which cannot be explained by this model.

380 More recently, Tanaka et al.^{Tanaka2019} proposed that the OSR with its la-
 381 tency shift can arise in a deep neural network model via summation of multiple
 382 excitatory inputs with different time constants. It is difficult to evaluate whether
 383 the model accurately captures the latency shift as observed in experiments, with
 384 the correct slope value. The explanation behind this model is that the OSR la-
 385 tency is determined by the sum of 2 ON bipolar cells which are activated only
 386 by certain stimulus frequencies due to different temporal filtering. This purely
 387 excitatory mechanism of latency scaling is not in line with our experimental
 388 findings, suggesting that amacrine cells likely contribute to temporal filtering as
 389 well. Our hypothesis is thus that the components they isolated correspond to a
 390 mix of bipolar and amacrine cell properties.

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

396 Previous experimental studies^{Schwartz2007a, Schwartz2008, Werner2008} reported
 397 that the OSR is found in a higher proportion of retinal ganglion cells than what
 398 we experimentally observed in this study. This difference could come from the
 399 fact that we define OSR as a response with a latency shift having a slope of at
 400 least 0.7, while it is not clear whether previous studies took multiple frequencies
 401 into account when classifying the OSR. Schwartz et al.^{Schwartz2008} also showed
 402 that blocking inhibition from amacrine cells had no effect on the OSR. But
 403 again, this study only investigated the presence or absence of the OSR under
 404 amacrine blockade, and did not investigate if the latency of the OSR shifted with
 405 the stimulus frequency. In addition, some previous studies were mostly carried
 406 out in salamander^{Schwartz2007a, Schwartz2008}, where the underlying mechanisms
 407 may be different from the mouse.

408 In order to realistically simulate the model’s response with glycinergic amacrine
 409 cells blocked, we had to decrease the weight of the inhibitory ON input I_{ON} in
 410 our simulations. Leaving the weight of this input untouched while setting $w_{I_{Gly}}^{OFF}$
 411 to 0, we still obtain a decrease in latency shift but this configuration gener-
 412 ated a response to each flash of the sequence, something we did not observe

413 experimentally. We therefore deemed this configuration as less realistic than
 414 decreasing also the weight of the ON inhibitory cell, since strychnine is likely
 415 to affect glycinergic ON inhibition as well. Additionally, the components of our
 416 circuit might represent several cell types pooled together, and more detailed
 417 circuit models might give similar predictions. For example, we chose to sim-
 418 ulate synaptic depression via a modified version of cortical STP-models with
 419 only 2 parameters rather than the more complex systems used in the retina
 420 previously [Schroeder2020](#). Overall, we chose to only include the minimal compo-
 421 nents necessary to specifically explain the peak latency shift in the OSR and its
 422 abolishment via strychnine and arrived at a model with (still) 19 parameters.

423 In addition, bath application of strychnine acts on the whole retinal network
 424 and likely has side effects such as changes (increases) in the baseline activity
 425 across bipolar, other amacrine and ganglion cells due to a reduction of main-
 426 tained inhibition. However, these maintained effects, although they would likely
 427 decrease absolute latencies, should not depend on the stimulus and are thus un-
 428 likely to explain changes in the latency shift relative to stimulus period. For
 429 simplicity, we thus do not take these side effects into account in our model-
 430 ing approach and only simulate the minimal amount of direct inputs we found
 431 necessary for an OSR with latency shift.

432 Our present model accounts only for an OSR to dark-type flashes in ON cells.
 433 It has however been reported as well that OFF cells emit a similar response after
 434 cessation of OFF-type stimulation which our model as it is does not explain.
 435 This could be obtained by adding another independent OFF pathway that would
 436 deliver an input to the same ganglion cell. This pathway would trigger a response
 437 during the flashes, while the circuit described in our model would generate an
 438 OSR that scales with latency.

439 We experimentally identified an important role for glycinergic inhibition in
 440 the latency shift of the OSR in experiments. The role of GABAergic inhibi-
 441 tion in generating the OSR is less clear. In principle, depressing GABAergic
 442 synapses could play a similar role as glycinergic synapses or be the source of
 443 disinhibition needed for the rebound response. We tried to block GABAergic
 444 inputs with Gabazine to test this hypothesis experimentally. We indeed found
 445 a small number of cells that also lost their latency scaling and others that lost
 446 the entire OSR after application of gabazine (data not shown), supporting both
 447 hypotheses. However, we observed a large heterogeneity in the effect of gabazine
 448 on the OSR across experiments, so these results remain inconclusive (see Figure
 449 S7).

450 In this study, we focused on the OSR defined as the response to the abrupt
 451 cessation of a periodic stimulus, following several previous studies [Schwartz2007a](#), [Werner2008](#), [Schwartz2008](#), [Gao2008](#).
 452 It has however been shown that retinal ganglion cells in both mouse and sala-
 453 mander emit an OSR to more complex pattern violations, for example in re-
 454 sponse to an omission in the middle of a periodic stimulus sequence [Schwartz2007a](#),
 455 which we did not test here. We would expect to observe such a response if the
 456 overall latency of the OSR for a given frequency was shorter than twice the
 457 stimulus period. In a stimulus where one flash is omitted in the middle of
 458 a sequence, this would allow for the OSR to signal this omission before flash

459 sequence continues. According to absolute latencies of cells recorded in this
 460 study, several cells should signal omissions in the middle of a flash stimulus up
 461 to frequencies of 12 Hz. (see Figure S7).

462 Dynamical synapses have previously been proposed to enable neuronal cir-
 463 cuits in the retina to form expectations of future inputs^{Hosoya2005Johnston2019}
 464 and are thus a plausible candidate to play an important role in the OSR. Previ-
 465 ous works have shown that inhibitory synapses can be depressing^{Kastner2011, Kastner2013b, Nikolaev2013}.
 466 In particular, glycinergic synapses that input to bipolar cells can be depress-
 467 ing^{Huang2022}. However, there is no method to experimentally remove the de-
 468 pressing nature of the synapse without affecting the inhibitory weight, so we
 469 could not show experimentally that the depressing nature of the synapse is
 470 necessary for the OSR.

471 Adaptation in excitatory ribbon synapses is a well-studied phenomenon to
 472 reliably encode luminance and contrast and could be the source of the OSR
 473 latency shift as well. While a more simple model with depression in bipolar
 474 cell terminals can provide an OSR with latency shift as well (see Figure S6),
 475 such a model is not congruent with our findings that stimulus luminance and
 476 contrast are neither necessary nor sufficient for the OSR latency shift and can-
 477 not readily explain our experimental observations under strychnine. While it
 478 has been shown that GABAergic inhibition modulates gain changes in ribbon
 479 synapses^{Oesch2019}, glycinergic effects at this stage have not been observed so
 480 far. Therefore we conclude that, even though plasticity in bipolar cells certainly
 481 plays an important role in retinal computations, the idea of inhibitory plasticity
 482 provides a simpler explanation of this specific phenomenon.

483 In addition, our model predicts that the depressing inhibitory synapse should
 484 have several functional consequences, that we verified in the data. In particular,
 485 a key prediction of the depressing synapse is that the OSR requires a long enough
 486 flash sequence to accurately shift the latency and we confirmed this prediction
 487 experimentally.

488 Ultimately, our results might be of relevance to understanding neuronal
 489 mechanisms of predictive coding beyond the retina. Very similar surprise re-
 490 sponses exist in other sensory domains, such as the mismatch negativity re-
 491 sponse in the auditory cortex^{Naatanen1993, Garrido2009, Ulanovsky2003, Li2017},
 492 where neural activity is enhanced following a "deviant" tone in a sequence of
 493 "standard" tones. A recent study suggested that synaptic adaptation could
 494 be a key contributor to this phenomenon^{Amsalem2020}. Following the predictive
 495 coding theory, one possible explanation is that this response emerges from an in-
 496 teraction between feed-forward and feedback connectivity^{Rao1999, Millidge2021}.

497 Here we show that a purely feed-forward micro-circuit can generate this re-
 498 sponse to a violation of prediction via an interplay of excitation and inhibition,
 499 where synaptic depression takes place in inhibitory connections. All the com-
 500 ponents used in this micro-circuit are generic and can be found in other sensory
 501 areas^{Deneve2016, TsodyksMarkram1997, Abbott1997}, and it is thus likely that a
 502 similar circuit could be at work at the cortical level, for more complex pattern
 503 recognition than full-field flashes.

504 4 Methods

505 4.1 Experimental Setup

506 4.1.1 Recordings

507 Recordings were performed on isolated retinas from 8 C57BL6/J adult mice
508 aged between 2 and 7 months. Strychnine Experiments were performed on
509 two males (age 84 and 52 days) and one female (85 days). Luminance exper-
510 iments were performed on two retinas from one female (161 days). Gabazine
511 experiments were performed on two females (58 and 163 days) and one male
512 (185 days). Experiment with different flashnumbers was performed one a fe-
513 male mouse (30 days). The animals were housed in enriched cages with ad
514 libitum food, and watering. The ambient temperature was between 22 and 25
515 C, the humidity was between 50 and 70% and the light cycle was 12–14h of
516 light, 10–12h of darkness. Animals were killed according to institutional animal
517 care standards of Sorbonne Université. The retina was isolated from the eye
518 under dim illumination and transferred as quickly as possible into oxygenated
519 Ames’ medium (Merck, A1420). The retina was extracted from the eye cup
520 and lowered with the ganglion cell side against a multi-electrode array whose
521 electrodes were spaced by $30\ \mu\text{m}$ ^{Marre2012}. During the recordings, the Ames’
522 medium temperature was maintained at 37°C. Raw voltage traces were digitized
523 and stored for off-line analysis using a 252-channel preamplifier (MultiChannel
524 Systems, Germany) at a sampling frequency of 20kHz. The activity of single
525 neurons was obtained using Spyking Circus V 1.1.0, a custom spike sorting
526 software developed specifically for these arrays^{Yger2018}.

527 4.1.2 Visual stimulation

528 Visual stimuli were presented using a white LED and a Digital Mirror Device
529 (DMD). Flash sequences contained 5 or 12 flashes of 5 different frequencies (6Hz,
530 8Hz, 10Hz, 12Hz, 16Hz). Polarities were either switched from grey to black
531 (dark flashes) or from grey to white (bright flashes). 60 trials were conducted
532 for each stimulus, with 2-4 s between each trial. The order of magnitude of the
533 background illumination was $10^6\ \text{R}^*$.

534 4.1.3 Spike Triggered Average

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

541 4.1.4 Pharmacology

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

545 4.1.5 Latency Analysis

546 To determine slope of latency shift, we measured the latency between the peak
 547 firing rate and the end of the last flash in the stimulus for all frequencies tested.
 548 We plotted these latencies against the respective period of the stimulus and
 549 fitted a straight line to determine the slope of the latency shift using SciPy's
 550 optimization package^{SciPy2020}. Cells were classified as having an OSR in the
 551 control condition when the slope was at least 0.7 or higher. All cells where the
 552 peak time point could not be unambiguously determined in any condition were
 553 excluded from the analysis.

554 4.1.6 Statistical Analysis

555 Statistical testing was performed using Scipy's stats package^{SciPy2020}. Welch's
 556 t-test for independent samples with unequal variance was used for slope compar-
 557 isons between conditions. For comparison of amplitudes and latencies between
 558 different flash numbers, paired two-sided t-tests with Bonferroni- Holm correc-
 559 tion were used such that significance levels α were adjusted to $\frac{\alpha}{m+1-k}$ for the k^{th}
 560 p-value of m total comparisons. Pearson correlation coefficients were calculated
 561 using Python's numpy build-in functions.

562 4.2 Modeling

563 4.2.1 Model Implementation

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

$$\alpha(t) = \frac{t}{\tau_{\text{OPL}}^2} \exp\left(-\frac{t}{\tau_{\text{OPL}}}\right) H(t), \quad (1)$$

567 where τ_{OPL} is the characteristic time of integration (in s) in photoreceptors and
 568 $H(t)$ the Heaviside function.

569 Thus, the output of the OPL reads:

$$F(t) = [\alpha * s](t), \quad (2)$$

570 where $*$ is the space-time convolution. Note that, the stimulus being spatially
 571 uniform, the space integration reduces to a constant, so that the detailed shape
 572 of the spatial RF plays a trivial role here.

573 The OPL response is then passed through a sigmoidal nonlinearity to account
574 for nonlinear processing steps in early bipolar cell processing, given by:

$$\sigma_{ab}(F) = \frac{S_X}{1 + e^{-a(F-b)}} \quad (3)$$

575 where a and b are parameter defining the shape of the nonlinearity (they depend
576 on X but we didn't add this dependence to alleviate notations). $|S_X| = \frac{S_{mV}}{\tau_X}$ is
577 a scale factor which scales the input into each unit to $\frac{mV}{s}$ via $S_{mV} = 20mV$ and
578 the time constant τ_X of each unit, such that the amplitude of the response is
579 normalized to the time constant of the respective unit. For OFF cells, the OPL
580 inputs are multiplied by -1 before rectification to reverse polarity. The rectified
581 signal is then integrated into each pathways via a linear dynamical system:

$$\frac{dV_X}{dt} = -\frac{V_X}{\tau_X} + \sigma_{ab}(F(t)), \quad X = E^{ON}, I^{ON}, I_{Gly}^{OFF}. \quad (4)$$

582 where V_X is the voltage of cell X (in Volt) and τ_X its characteristic time constant
583 (in s). The Current I_X of unit X is computed as

$$I_X = \frac{w_X p_\theta(V_X)}{c_m} \quad (5)$$

584 with a membrane capacitance set to $c_m = 0.1nF$.

585 Note that this dynamical system is non autonomous as $F(t)$, the input,
586 explicitly depends on time.

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

$$\frac{dV_G}{dt} = -\frac{V_G}{\tau_G} + w_{EON} p_\theta(V_{EON}) + n w_{I_{Gly}^{OFF}} p_\theta(V_{I_{Gly}^{OFF}}) + w_{ION} p_\theta(V_{ION}). \quad (6)$$

588 where $w_{EON}, w_{I_{Gly}^{OFF}}, w_{ION}$ are synaptic weights (in Hz). Voltages are rectified
589 before integrated in the ganglion cell membrane potential via :

$$p_\theta(V) = \begin{cases} V - \theta & \text{if } V \geq \theta : \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

590 where θ is a threshold (in V) that differs between pathways.

591
592 The synaptic weight from I_{Gly}^{OFF} to G is modulated by a dimensionless vari-
593 able n , used to simulate synaptic short-term plasticity. n , which interprets
594 as a vesicle occupancy in the glycinergic amacrine synapse, obeys the kinetic
595 equation **Henning2013** :

$$\frac{dn}{dt} = (1 - n)k_{rec} - \beta k_{rel} p(V_{I_{Gly}^{OFF}}, \theta_{I_{gly}^{OFF}}) n. \quad (8)$$

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

$$R(t) = s_G p(V_G(t), \theta_G), \quad (9)$$

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

601 4.2.2 Parameter Optimization

602 First, time constants and weights were fitted such that the linear response V_G
603 of the model (without nonlinearities) to an impulse stimulus best resembles the
604 temporal STA of one example cell with OSR. This was done using an evolu-
605 tionary optimization algorithm using the CMAES (Covariance Matrix Adapta-
606 tion Evolutionary Strategy) Python toolbox^{Hansen2019}. Synaptic plasticity is
607 removed from the model by setting $\beta = 0$. The slope of the ganglion cell non-
608 linearity S_G was then chosen as to match experimentally observed firing rates.
609 Parameters of the occupancy equation (8) were manually chosen as to yield a
610 slope of 1. Thresholds for synaptic rectifications were calculated such that the
611 rest potential of G remained at 0 without input. Slopes and thresholds for non-
612 linear processing within units were manually chosen. Values were obtained by
613 screening over a wide range of combinations of β and the ratio of recovery and
614 release rate $\frac{k_{rec}}{k_{rel}}$. All parameters used in the simulations are listed in Table 1.

615 5 Data Availability

616 The data generated in this study have been deposited in the Zenodo database un-
617 der access code [10.5281/zenodo.11396185](https://zenodo.org/record/11396185). Due to large file sizes, raw datasets
618 are available from the corresponding author upon request. Source data for all
619 Figures are provided with this paper.

620 6 Code Availability

621 The code used for the computational model is accessible on gitlab.

622 7 Acknowledgments

623 We thank Romain Brette, Matthias Henning and Romain Veltz for helpful dis-
624 cussions about the model, and Matias Goldin and Brice Bathellier for critical
625 reading of the manuscript. We thank Giulia Spampinato for allowing us to use
626 her excellent artwork. This work has been funded by a PhD fellowship from the
627 interdisciplinary Institute for Modeling in Neuroscience and Cognition (Neu-
628 roMod) of Université Côte d’Azur to S.E., funded by the National Research
629 Agency (ANR-15- IDEX-01) and, by an ERC grant (No 101045253, DEEP-
630 RETINA) to O.M., ANR grants (DECORE, ANR-18-CE37-0011, and PerBaCo,
631 ANR-22-CE37-0016-02) to O.M., a grant from Retina France to O.M., and ANR
632 ShootingStar (ANR-20-CE37-0018-04) to O.M. and B.C. T.B. was funded by a
633 PhD fellowship from ENS and supported by the Fondation pour la Recherche
634 Médicale, grant number FDT202304016465.

635 **8 Author contributions**

636 S.E., T.B., O.M. and B.C. designed the study. T.B. and B.S.S. performed the
 637 experiments. S.E. and T.B. analyzed the data with help from O.M. and B.C.
 638 S.E. did the model with help from T.B., O.M. and B.C. S.E., T.B., O.M. and
 639 B.C. wrote the paper.

640 **9 Competing interests**

641 The authors declare no competing interests.

Parameter	Value	Unit
τ_{OPL}	0.003	s
τ_{EON}	0.08	s
τ_{ION}	0.085	s
$\tau_{\text{I}_{\text{Gly}}^{\text{OFF}}}$	0.12	s
τ_{G}	0.11	s
a_{ON}	14.0	1
a_{OFF}	12.0	1
b_{ON}	-0.5	1
b_{OFF}	0.5	1
w_{EON}	50.0	Hz
w_{ION}	-65.0 /36.0	Hz
$w_{\text{I}_{\text{Gly}}^{\text{OFF}}}$	-53.0 /0.0	Hz
S_{EON}	250.0	mVs ⁻¹
$S_{\text{I}_{\text{Gly}}^{\text{OFF}}}$	-235.3	mVs ⁻¹
S_{ION}	166.7	mVs ⁻¹
θ_{EON}	26.0	mV
θ_{ION}	-20.0	mV
$\theta_{\text{I}_{\text{Gly}}^{\text{OFF}}}$	0.0	mV
k_{rel}	5.0	Hz
k_{rec}	10.0	Hz
β	0.0826	mV ⁻¹
θ_{G}	0.0	V
s_{G}	2200	HzmV ⁻¹

Table 1: Model parameter values used in simulations.

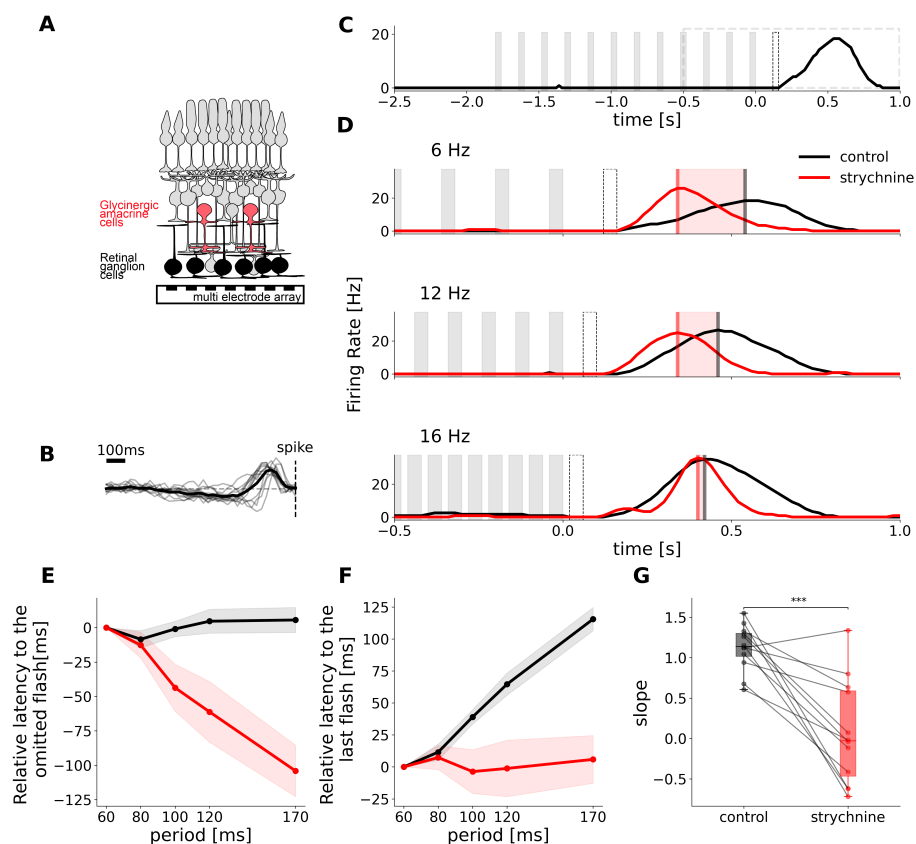


Figure 1: Glycinergic Amacrine cells are necessary for predictive timing of the OSR. **A.** Schematic representation of the retina, with the activity of retinal ganglion cells being recorded with a multi electrode array. **B.** Temporal traces of the receptive fields of the cells that loose the latency shift with strychnine, $n = 12$. Bold trace shows mean of all cells. **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 black 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 grey patches, vertical lines indicate the maximum of the response peak, red shaded areas indicate the temporal discrepancy between control and strychnine conditions. **E.** Mean \pm SEM 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 control condition. **F.** Mean \pm SEM latency between OSR and last flash in the stimulus plotted against stimulus period for a population of $n = 12$ cells. Control latencies shift with the period of the stimulus with 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.** Slope of relative latency vs stimulus period for control and strychnine conditions. *** indicates statistical significance.

G. Slope of the latency shift illustrated as a boxplot (band indicates the median, box indicates the first and third quartiles and whiskers indicate ± 1.5 interquartile range). Circles show individual datapoints, lines connect values of the same cell. The slope decreases significantly when strychnine is added, $p = 1.09e^{-4}$ (two-sided Welch t-test). Source data are provided as a Source Data file. Figure 1A is used with permission from Giulia Spampinato.

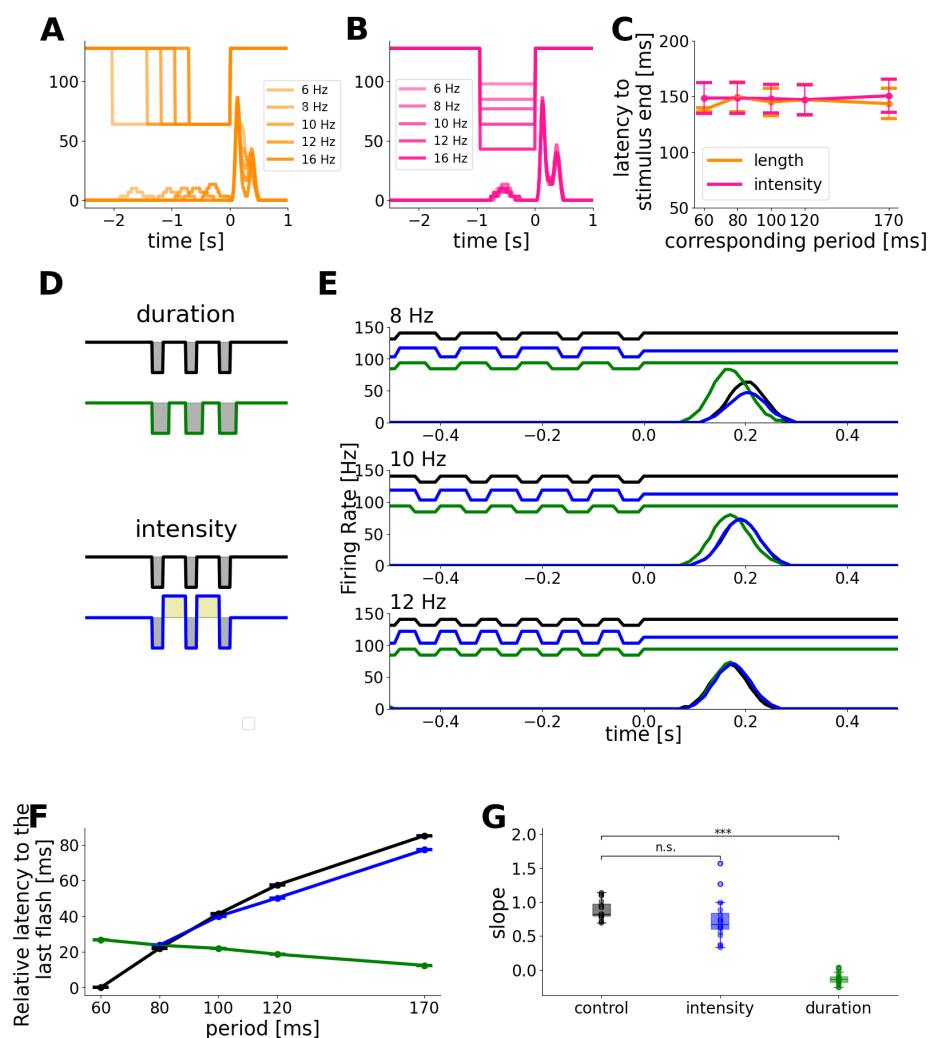


Figure 2: **Latency scaling of the OSR is selective to specific input patterns.** **A.** Firing rate responses of one example cell to dark step stimuli with different durations equivalent to the duration of flash trains with frequencies 6-16 Hz. **B.** Firing rate responses of one example cell to dark step stimuli with different light intensities equivalent to the mean light intensity of flash trains with frequencies 6 to 16 Hz. **C.** Mean \pm SEM between response latency and corresponding stimulus period for step stimuli with duration or luminance modulation comparable to the stimulus composed of periodic flashes, $n = 20$. **D.** Schematic description of the stimulus modifications for constant luminance across frequencies. "Duration" describes a modified stimulus in which the duration of each flash is set to half of the stimulus period. "Intensity" describes a stimulus modification in which the intervals between flashes are of brighter intensity to maintain a constant average luminance level **E.** Response traces to flash trains of 3 different frequencies of each modification shown in **D**. **F.** Mean \pm SEM scaling between response latency and stimulus period for all conditions, $n = 20$ cells.

G. Fitted slope between peak latency and stimulus period for $n = 20$ cells, illustrated as a boxplot (the band indicates the median, the box indicates the first and third quartiles and the whiskers indicate ± 1.5 interquartile range). Black/blue/green circles show individual datapoints. Control mean \pm SEM 0.89 ± 0.03 , intensity mean \pm SEM 0.74 ± 0.06 , duration mean \pm SEM -0.13 ± 0.01 Stimuli with intensity modulation lead to no significant change in slope ($p = 5.22e^{-2}$, two-sided Welch t-test) while duration modulations decrease the slope significantly ($p = 7.2e^{-23}$, two-sided Welch t-test) Source data are provided as a Source Data file.

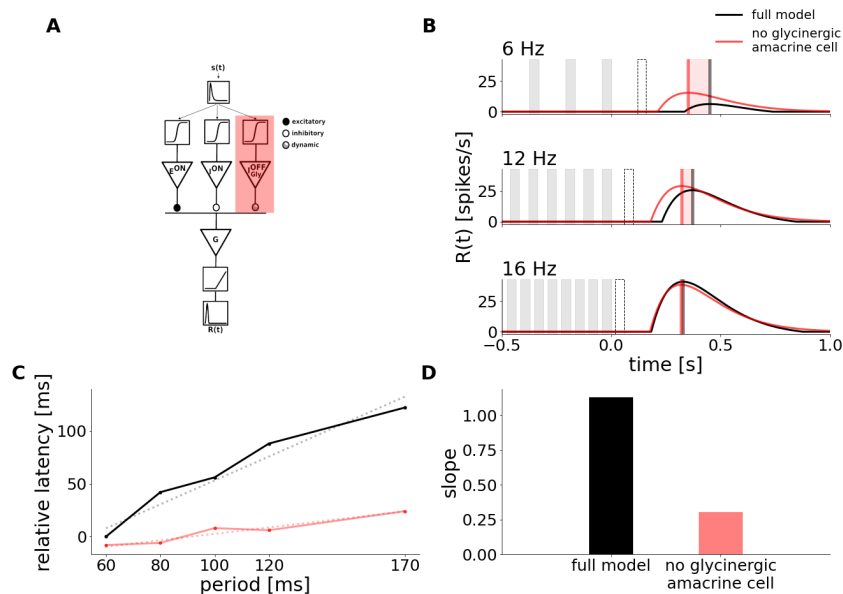


Figure 3: Mechanistic model replicates latency shift and strychnine experiment. **A.** Schematic description of the Model. It is composed of an ON excitatory input E^{ON} , an ON inhibitory input I^{ON} , and an OFF inhibitory input I_{Gly}^{OFF} representing a glycinergic amacrine cell. Each of those 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_{Gly}^{OFF} 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_{Gly}^{OFF} set to 0 (red, strychnine). The weight of I^{ON} was changed from -65 to -36 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 16 Hz stimulus. The slope of the latency shift decreased from 1.13 to 0.30 when I_{Gly}^{OFF} is set to 0 (dotted lines). **D.** Value of the slope fitted to latency shift in the full model and strychnine simulation. Source data are provided as a Source Data file.

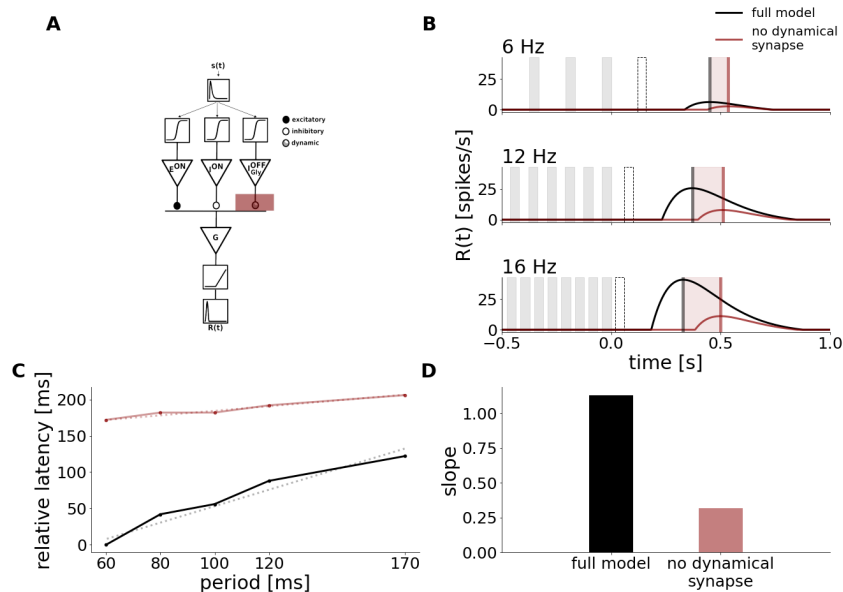


Figure 4: **Short-term plasticity is the crucial component for latency shift.** **A.** Schematic description of the Model, as in Figure 3. The shaded red area now represents removing the dynamical characteristic of the glycinergic synapse. **B.** Simulation of the model responses to flash trains of 3 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, as in Figure 3. **C.** Latency of the OSR plotted against stimulus period in control and strychnine simulation. The latency is expressed relative to the latency of the response to the 16 Hz stimulus in the control condition. The slope of the latency shift decreased from 1.13 to 0.31 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. Source data are provided as a Source Data file.

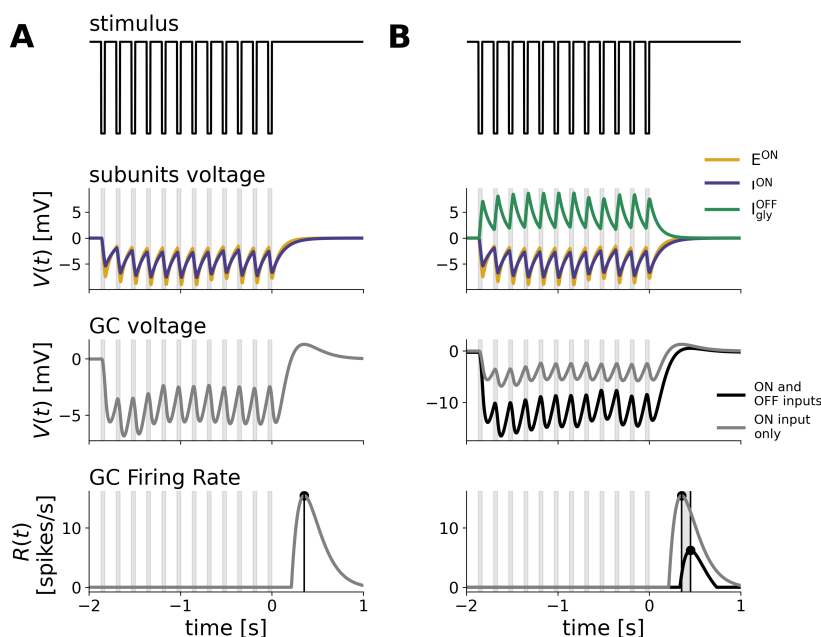


Figure 5: **ON components of the model produce a peak after stimulus end while the glycinergic OFF input shifts the latency.** From top to bottom: Stimulus intensity, Bipolar and Amacrine voltage responses, ganglion cell voltage and firing rate. **A.** Internal model responses of ON components to a 6 Hz dark flash stimulus without OFF inhibition. ON excitation and inhibition hyperpolarize in response to dark flashes. When both inputs are subtracted 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.** Effect of 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 $I_{\text{Gly}}^{\text{OFF}}$. Source data are provided as a Source Data file.

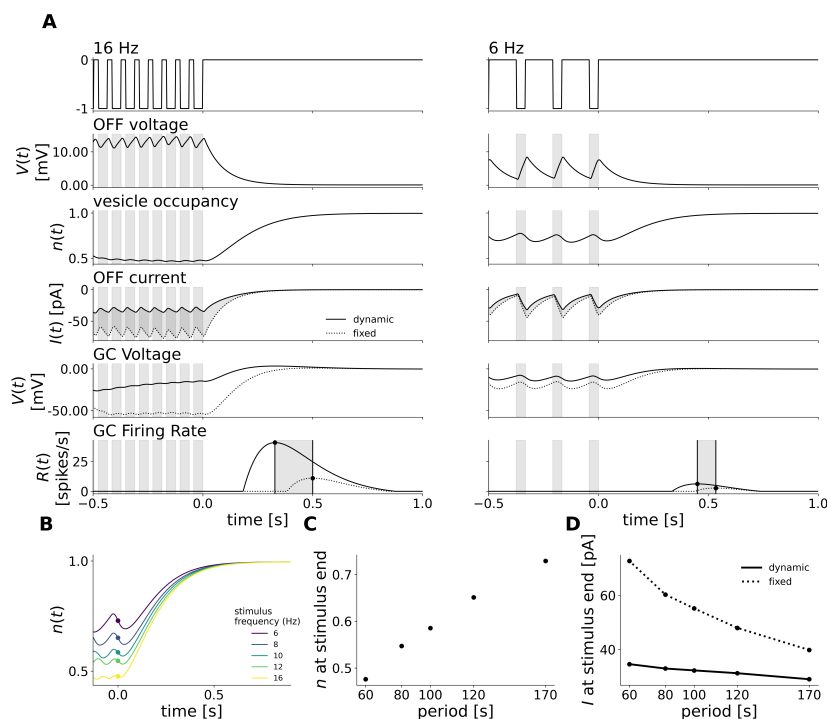


Figure 6: Synaptic depression scales OFF glycinergic input to stimulus frequency and thereby shifts the latency of the response. Responses shown are zoomed in around the end of the flash sequence. **A.** Impact of occupancy scaling on $I_{\text{Gly}}^{\text{OFF}}$ current input to G for a fast (16 Hz, left) and a slow (6 Hz, right) stimulus. From top to bottom: Stimulus intensity, $I_{\text{Gly}}^{\text{OFF}}$ 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. Current of unit X is computed with equation 5 in Methods. **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.** $I_{\text{Gly}}^{\text{OFF}}$ 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. Source data are provided as a Source Data file.

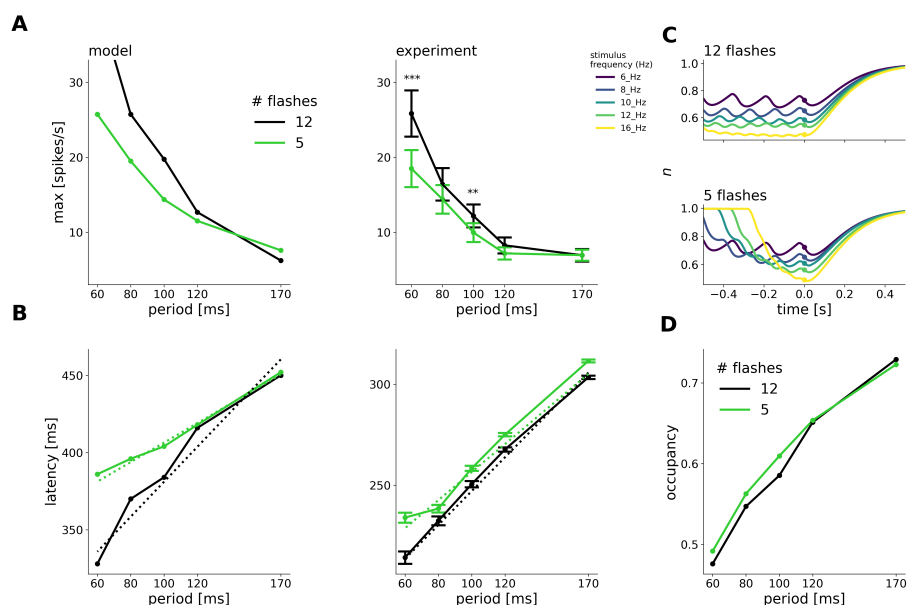


Figure 7: Latency shift decreases for shorter stimuli because of lacking steady state occupancy. **A.** Mean \pm SEM amplitude of the OSR against stimulus period for 12 and 5 flashes in the stimulus in simulations (left) and experiments (right). Amplitudes to 6 Hz and 10 Hz stimuli were significantly different according to two-sided t-test with Bonferroni-Holm correction^{Holm1979} (16 Hz: $p = 0.000006$, 10 Hz: $p = 0.001$). **B.** Latency against stimulus period in simulations (left) and experiments (right). Simulated slopes decreased from 1.13 after 12 flashes to 0.63 after 5 flashes in simulations. Experimental slopes decreased from 0.84 ± 0.02 to 0.69 ± 0.04 , mean \pm SEM. The latency after 12 and 5 flashes was not significantly different according to two-sided t-test with Bonferroni-Holm correction^{Holm1979} (6 Hz: $p = 0.2$, 8 Hz: 0.14, 10 Hz: $p = 0.044$, 12 Hz: $p = 0.031$, 16 Hz: 0.024). **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. Source data are provided as a Source Data file.

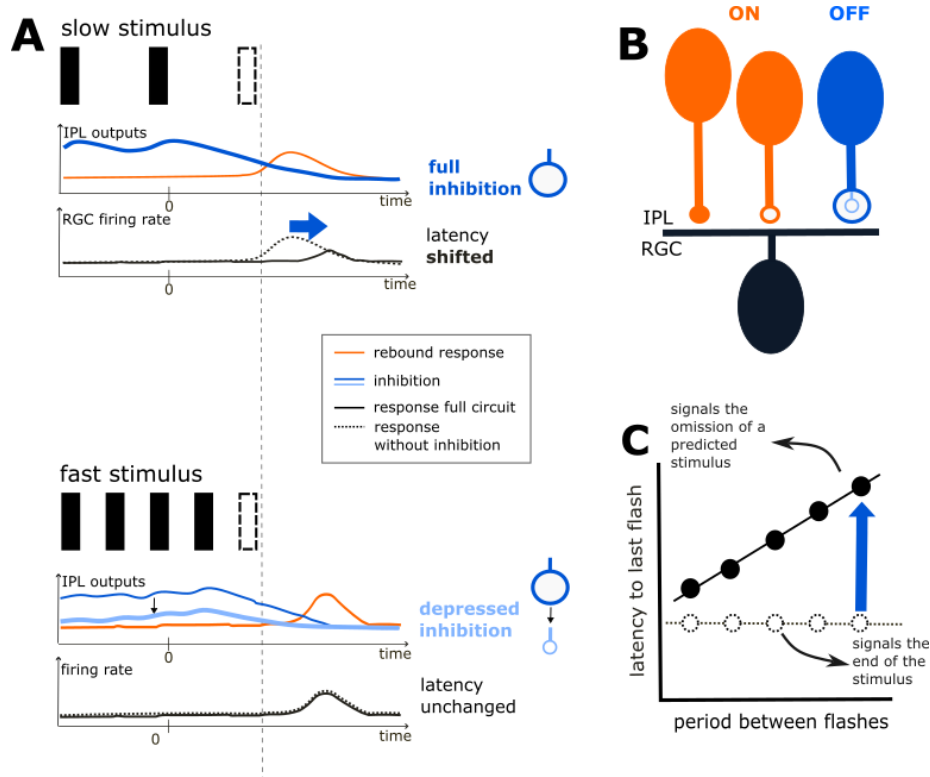


Figure 8: **Mechanistic neuronal circuit explains the latency shift of the Omitted Stimulus Response (OSR) via depressing inhibition.** **A.** Schematic responses of components shown in **B** to a slow-frequency and a fast frequency stimulus. Black rectangles represent dark flashes in a periodic sequence. The dotted rectangle represents an omitted flash at the end of the sequence. For both stimulus speeds, the upper panel shows inputs from the IPL. ON (orange) inputs are combined into one trace and OFF inhibitory input is shown in blue. ON inputs generate a peak response after stimulus end. OFF inhibitory input is activated by the dark flash sequence and then slowly decrease after stimulus end, overlapping with the ON peak. The lower panel shows the RGC firing rate. Without OFF inhibition (dotted black), the RGC emits a peak at the same time as the ON inputs. Adding OFF inhibitory inputs (solid black) generally decreases the this response and push the peak backwards. The impact of inhibition depends on the stimulus frequency due to synaptic depression. At slow frequency inputs, synaptic depression is not activated. Inhibitory inputs remain strong and shift the latency of the OSR (blue arrow) to a later time. Fast-frequency stimuli trigger strong synaptic depression, leading to a depletion of inhibitory inputs. The latency remains unchanged. **B.** Schematic description of the circuit. A retinal ganglion cell (RGC, in black), which emits an OSR with latency shift, receives 3 inputs in the IPL (inner plexiform layer). Orange units are ON inputs, which provide excitation (filled circle) and inhibition (empty circle). The blue unit provides OFF inhibition and has a dynamical synapse which depresses after activation.

C. Latency of the OSR to the last flash in the stimulus plotted against stimulus period. A RGC without depressing inhibitory inputs (dotted black) will emit a response at roughly the same time after stimulus end. It thus merely signals the end of the stimulus. With depressing inhibition (solid black), the latency is shifted backwards for slower frequencies (with longer periods), which leads to a 1-to-1 scaling between latency and stimulus period. Via this mechanism, the latency of the OSR with respect to the omitted flash remains constant across frequencies, indicating that the RGC signals the omitted stimulus. The latency can thus provide a prediction of the time when the flash should have occurred.

642 Supplementary Figures

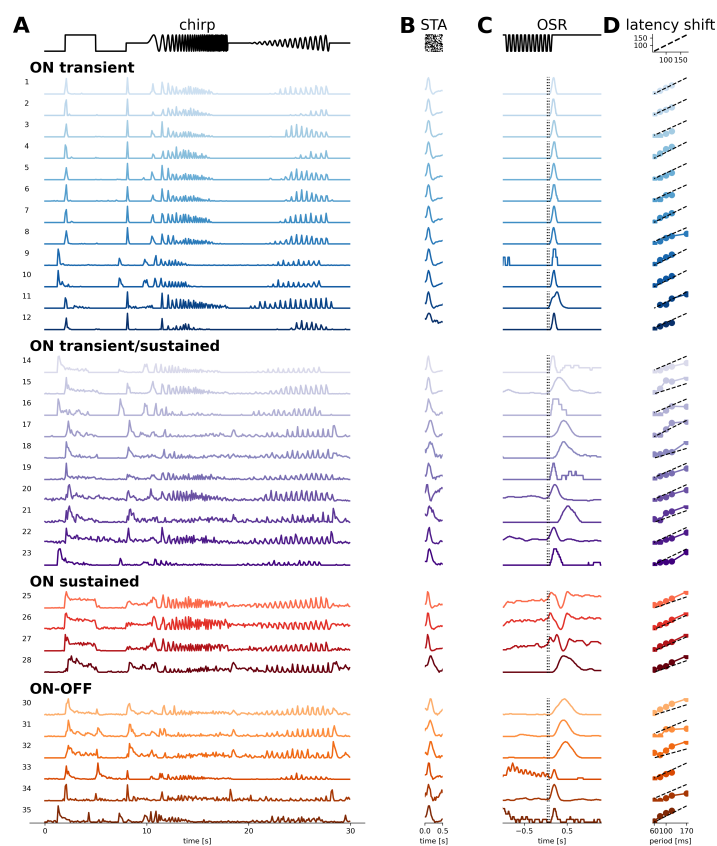


Figure S1: **ON cells that exhibit an OSR belong to various cell types.** Manual clustering of chirp responses from all ON cells with OSR yielded 4 broad cell types that show an OSR: ON transient, ON transient/sustained, ON sustained and ON-OFF. Shown are 35 cells with a response to the chirp stimulus and have a spike-triggered average to checkerboard stimulation. 5 other cells which are included in the statistics in the main paper but do not have a chirp response or STA are not shown in this figure. **A.** Response traces to a chirp stimulus. **B.** Temporal spike triggered average obtained after STA analysis to white noise-checkerboard stimulation. **C.** OSR after control stimulation with 12 Hz flash train (firing rate). **D.** Scaling between response latency and stimulus period for frequencies of 6,8,10,12,16 Hz. Black line has a slope of 1 for reference. Source data are provided as a Source Data file.

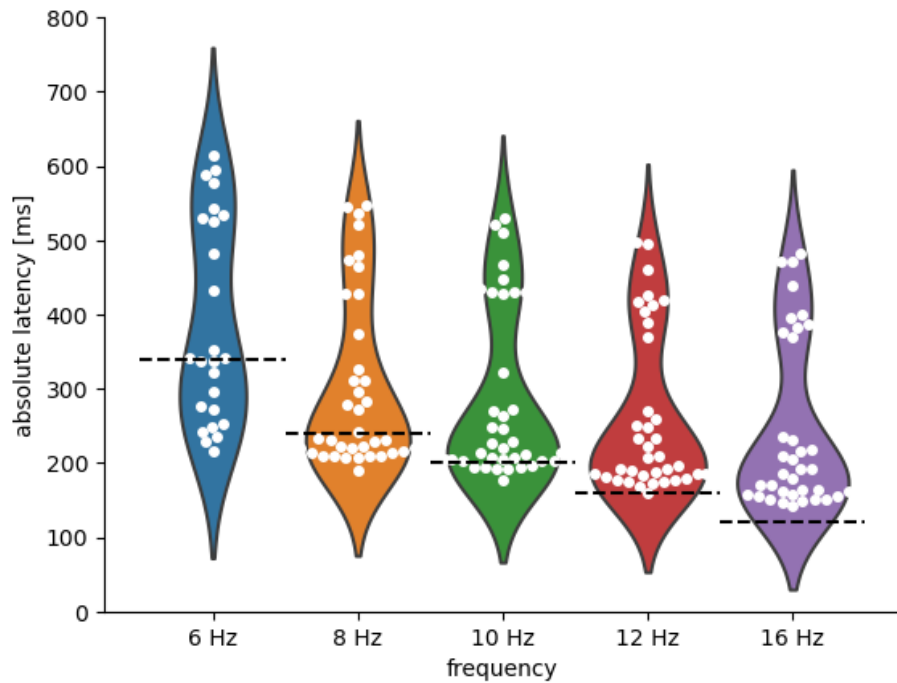


Figure S2: **Absolute latency of the OSR after all frequency tested of $n = 35$ cells included in our study.** Shaded colored areas show the density, individual datapoints are shown in light blue. Dotted black lines indicate $2T$ for each period $T = \frac{1}{f}$. 6 Hz: max = 615 ms , min 216 ms, median = 340 ms; 8 Hz: max = 547 ms , min 189 ms, median = 232 ms; 10 Hz: max = 529 ms , min 178 ms, median = 220 ms; 12 Hz: max = 497 ms , min 159 ms, median = 207 ms; 16 Hz: max = 483 ms , min 142 ms, median = 188 ms; Source data are provided as a Source Data file.

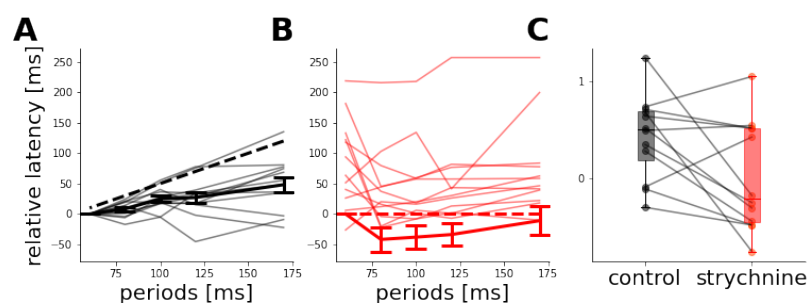


Figure S3: **Spiking onset is not a consistent quantity of the OSR latency shift.** **A.** Response onset latency against stimulus period in control conditions in all cells that show an OSR in their peak timing. Mean 0.39 ± 0.11 , compared to thick dotted line which has a slope of 1. All latencies are referenced to the latency to the fastest frequency in control condition. **B.** Response onset latency against stimulus period in strychnine conditions in all cells that show a latency shift in their peak timing. Mean 0.15 ± 0.36 , thick dotted line has a slope of 0. All latencies are referenced to the latency of the fastest frequency in control condition. **C.** Slope quantification, no significant difference slope in response onset between control and strychnine, $p = 0.056$. Source data are provided as a Source Data file.

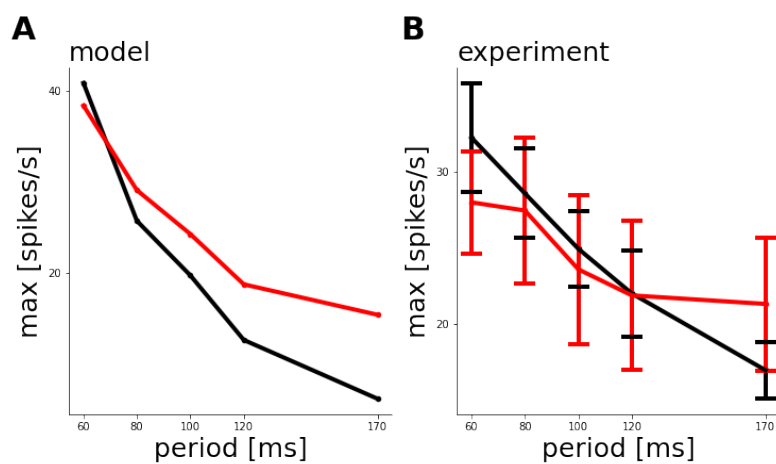


Figure S4: **Peak amplitude to different stimulus frequencies are not systematically impacted by strychnine.** **A.** Amplitude of the OSR against stimulus period for control and strychnine conditions in simulations. **B.** Same for experimentally measured amplitudes, not significantly different after Bonferroni-Holm correction^{Holm1979} (6 Hz: $p = 0.26$, 16 Hz: $p = 0.975$). Source data are provided as a Source Data file.

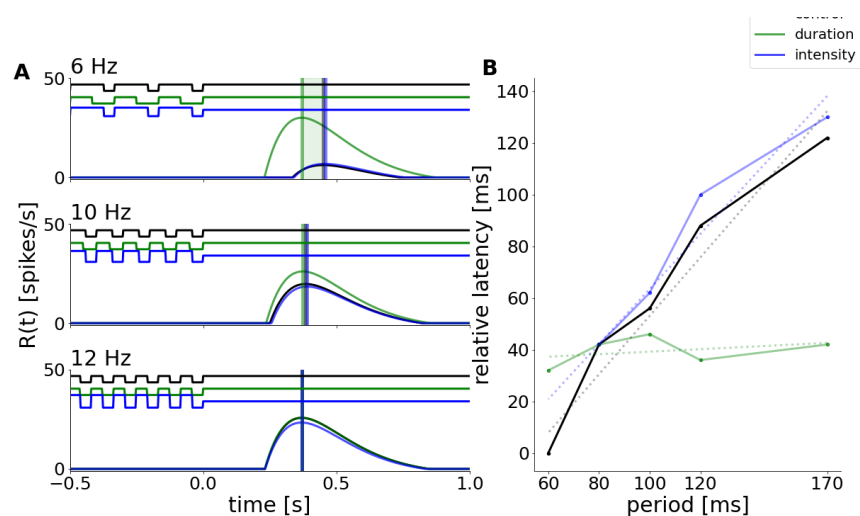


Figure S5: **Model predicts the selectivity of latency scaling to transient periodic flashes.** **A.** Simulated response traces of the model to flash trains of 3 different frequencies in control condition (black) and with modulations of flash-duration (green) and inter-flash brightness (blue). Stimuli are shown in the upper panels, simulated firing rate below. Vertical lines highlight the peak time-point, shaded areas show the latency shift compared to the control response. **B.** Latency of the OSR plotted against stimulus period for the 3 stimulus types. Solid lines are datapoints, dotted lines show linear fit between latency and period. Slope control: 1.13, slope duration modulation: 0.05, slope intensity modulation: 1.06.

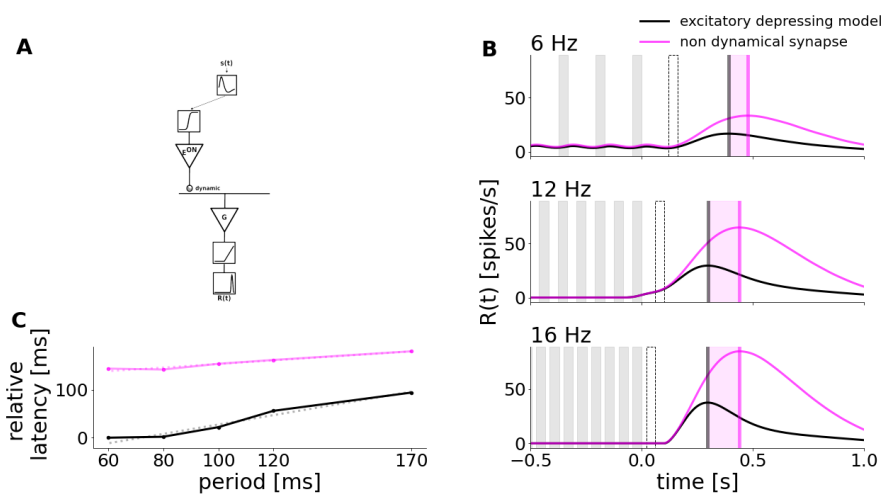


Figure S6: **Model with a biphasic ON bipolar cell can simulate the latency shift of the OSR via excitatory plasticity.** **A.** Schematic description of the model with one biphasic ON unit that has a plastic synapse. **B.** Response traces to 3 stimulus frequencies with the purely excitatory model with plasticity (black) compared to a model without plasticity (magenta). **C.** Scaling between response latency and stimulus period in both model variations. Slope full model: 0.98, slope without plasticity: 0.37. Parameter used for the simulation : $\tau_{OPL} = 0.1$, $\tau_{OPL2} = 0.1$, τ_{EON} , $\tau_G = 0.05$, $a_{ON} = 14.0$, $b_{ON} = 0.0$, $w_{EON} = 200.0$, $S_{EON} = 0.5$, $\theta_{EON} = 0$, $k_{rel} = 50.3$, $k_{rec} = 1.0$, $\beta = 13.6$, $\theta_G = 0.0$, $s_G = 2200$. Source data are provided as a Source Data file.

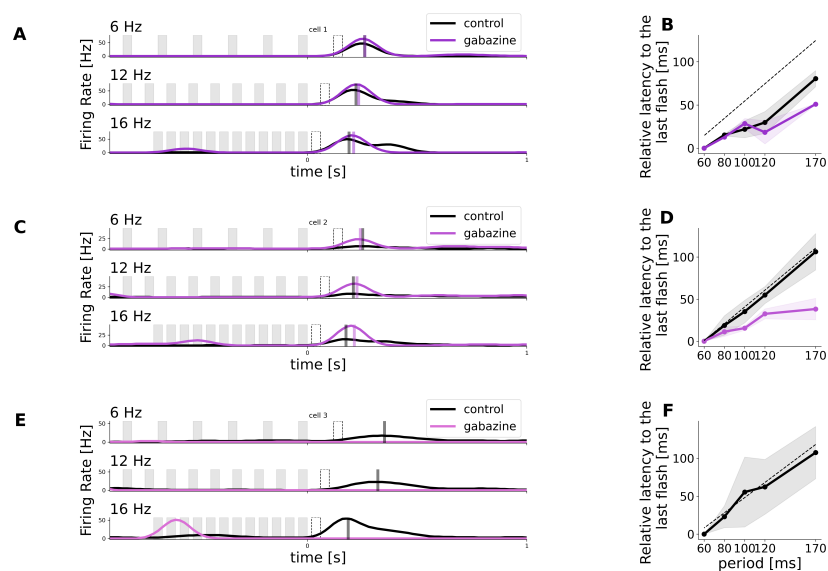


Figure S7: Gabazine has various effects on the OSR. **A.,C.,E.** Response traces to flash trains of 3 different frequencies in control (black) and gabazine (shades of purple) conditions. **B.,D.,E.** Mean relative latency to the last flash of the stimulus plotted against stimulus period for control (black) and gabazine (shades of purple) conditions. Dotted line shows slope of 1. **A.,B.** Example cell where OSR remains largely unaffected by gabazine. (Slope control = 0.71, gabazine = 0.4, $n = 2$ cells) **C.,D.** Example cell where OSR slope decreases with gabazine. (Slope control = 0.97, slope gabazine = 0.35, $n = 4$ cells). **E.,F.** Example cell where OSR disappears though gabazine. (Slope control = 0.96, $n = 4$ cells). Source data are provided as a Source Data file.