Rapid Neural Coding in the Retina with Relative Spike Latencies

Tim Gollisch* and Markus Meister†

Natural vision is a highly dynamic process. Frequent body, head, and eye movements constantly bring new images onto the retina for brief periods, challenging our understanding of the neural code for vision. We report that certain retinal ganglion cells encode the spatial structure of a briefly presented image in the relative timing of their first spikes. This code is found to be largely invariant to stimulus contrast and robust to noisy fluctuations in response latencies. Mechanistically, the observed response characteristics result from different kinetics in two retinal pathways ("ON" and "OFF") that converge onto ganglion cells. This mechanism allows the retina to rapidly and reliably transmit new spatial information with the very first spikes emitted by a neural population.

During natural vision, our gaze remains fixed for a mere fraction of a second. Sudden movements of the eye, called saccades, partition visual processing into short episodes (*1*, *2*). Each saccade exchanges the image that falls onto the retina; the new visual stimulus is then encoded into neural activity to be transmitted to the brain. Our visual system can analyze and classify a new complex scene in less than 150 ms (*3*), but the nature of the neural code that underlies this rapid visual processing has been elusive. Neurons in the vertebrate retina fire with remarkable temporal precision (*4*, *5*), so single spikes can, in principle, carry substantial information about visual stimuli. In order to assess how the retina transmits new visual information after a saccade, we investigated the responses of retinal ganglion cells to flashed visual images.

Spike trains were recorded simultaneously from many ganglion cells in the isolated salamander retina. The stimulus was a uniform gray field followed by appearance of a square-wave grating. Eight different shifted versions of the grating were used in a pseudo-random sequence. A ganglion cell typically responded to the appearance of the grating with a short burst of spikes (Fig. 1), and the vast majority of cells responded to most or even all of the stimuli. We characterized each burst by two numbers: the latency of the first spike after stimulus onset and the total spike count in the burst. For certain cell types, in particular fast and biphasic OFF cells (fig. S2), the spike count was very similar for all stimuli (Fig. 1B). By contrast, the spike latency for these cells varied across stimuli by as much as 40 ms (Fig. 1C). For repeats of the same stimulus, this latency was very reproducible, with a standard deviation of only 3 to 5 ms.

We calculated how much information the spike latency or the spike count conveys about which grating had been presented. Perfect identification of the stimulus among eight possibilities amounts to a maximum of 3 bits. The spike latency of a ganglion cell transmitted up to 2 bits of information on a single trial. The spike count provided considerably less information for the majority of all recorded cells (Fig. 1D). Subsequent brain regions may thus learn more about the stimulus from noting the time of the first spike after stimulus onset than by waiting for all spikes and noting the average firing rate.

In several sensory systems, shorter spike latencies result from stronger stimulation (*6*–*9*). This does not account for the present dependence of latency on spatial pattern. Stronger stimuli often generate higher spike counts, and indeed, gratings of higher contrast produced both more spikes and shorter latencies (fig. S3). By contrast, we observed a pronounced spatial tuning of the spike latency even when there were virtually no variations in spike count (Fig. 1); in some cases, shorter latencies even occurred in combination with fewer spikes (fig. S3).

Downstream brain centers can interpret the latency of a single neuron only if the onset time of the stimulus is known (*10*). If the new retinal image was initiated by an eye movement, then the brain does know the onset time, but it is unclear whether this motor information gets distributed to visual centers. We therefore asked what information can be extracted from visual signals alone by comparing latencies from neurons in the population (Fig. 2). For many pairs of ganglion cells, the difference between first spike times was strongly tuned with respect to the presented stimuli (Fig. 2C). In fact, the information contained in the latency difference reached values higher than 2 bits (Fig. 2D)—more than that from any single-cell absolute latency. One reason was the particular robustness of the latency difference to retinal noise. Each cell's latency underwent some trial-to-trial variation, but these fluctuations were often positively correlated in cells recorded simultaneously; when cell 1 fired earlier than usual, cell 2 tended to do the same (Fig. 2B). As a result, the latency difference (Fig. 2C) fluctuated less than expected from the noise in individual cell latencies (Fig. 2A). To assess the relevance of this compensation, we destroyed the noise correlations artificially by pairing the response of cell 1 with the response of cell 2 on the subsequent trial; this led to a substantial information loss of up to ~20% (Fig. 2D).

Stimuli of greater strength tend to produce shorter spike latencies in the sensory response. If the latencies of different neurons in the population are affected in similar fashion, downstream circuits might use the difference in spike latencies to extract stimulus quality independent of stimulus strength (*11*). We therefore presented the flashed gratings at different contrast levels. As expected, individual latencies increased at lower contrast. However, the shape of the latency tuning curve was well preserved at each contrast level (Fig. 2, E and F). Furthermore, the contrast-dependent shifts of the latency tuning curves were similar for different cells. As a result, the latency difference between two neurons was almost perfectly invariant to changes in contrast (Fig. 2G). In fact, a downstream decoder could recover almost all the spatial information without knowing anything about the contrast of the stimulus (Fig. 2H).

How can the observed latency code be explained in terms of neural mechanisms? We start by considering a standard framework for visual responses (*12*, *13*) and exploring its prediction for first-spike latencies. In this picture, the stimulus is first passed through a linear filter that summarizes retinal integration of the image over space and time (Fig. 3A). The "activation" signal emerging from the filter can be interpreted as the membrane potential of the ganglion cell. When this signal crosses a preset threshold, the model neuron fires a spike. For each ganglion cell, we measured the spatiotemporal filter in a separate reverse-correlation experiment (fig. S1) (*14*), whereas the threshold remained as a single free parameter. In using this model to process grating stimuli, one quickly finds that it cannot account for the observed responses. Because the stimulus is integrated linearly, a certain grating may elicit strong excitation, but then its sign-reversed counterpart will elicit inhibition and produce no spikes at all, counter to what was observed experimentally (Fig. 3, B and C).

Thus, one is forced to include nonlinear processing steps. Ganglion cells draw their excitatory input from bipolar cells (Fig. 3D). These have comparatively small receptive fields [<100 μm; (*15*)] and respond to light in essentially linear fashion (*16*), but transmission to retinal ganglion cells may involve a degree of rectification (*17*–*19*). Furthermore, bipolar cells come in two major types: ON bipolars are excited by an increase in light intensity and OFF bipolars by a decrease. Individual ganglion cells can receive inputs from both types (*8*, *20*–*24*). To include this structure of the inner retina into the model, we replaced the single spatiotemporal filter by a set of parallel filters that mimicked spatially local ON and OFF bipolar cells (Fig. 3E). Transmission from bipolar to ganglion cells was approximated by a half-wave rectifying function (Fig. 3F). Under these conditions, all stimuli led to

Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

*Present address: Max Planck Institute of Neurobiology, Am Klopferspitz 18, 82152 Martinsried, Germany.

†To whom correspondence should be addressed. E-mail: meister@fas.harvard.edu

excitatory activation of the ganglion cell (Fig. 3G). Moreover, the predicted timing of the first spike agreed remarkably well with the observed latency tuning curve (Fig. 3H) and outperformed alternative circuit schemes (fig. S6) for virtually all cells with substantial latency tuning (fig. S7).

Note that all the elements of this model are well-known components of retinal circuitry. Closer inspection reveals how the circuit accomplishes latency tuning. First, the rectifying synapses ensure that every stimulus excites the ganglion cell: any image change within the receptive field will activate some set of bipolar cells that transmit their excitation to the ganglion cell (*18*). The lasting and synchronous activation of both ON and OFF pathways by flashed gratings emphasizes their nonlinear summation and, thus, the need for separate filters in the model. By contrast, earlier studies of ganglion cells (*13*) were based on spatially homogeneous stimuli that only transiently activate one pathway at a time and, therefore, allow for simpler models (*14*). Second, the measured ON filters have slower kinetics than the OFF filters (Fig. 3E and fig. S5), such that ON stimuli affect ganglion cell spiking ~30 ms later than OFF stimuli. This is consistent with prior observations (*23*) and probably results from a signal transduction delay at the synapse between photoreceptors and ON bipolars (*25*, *26*). Thus, the proportion of light and dark stimulation within the receptive field determines the relative contribution of the ON and OFF pathways and modulates the time of the first threshold crossing.

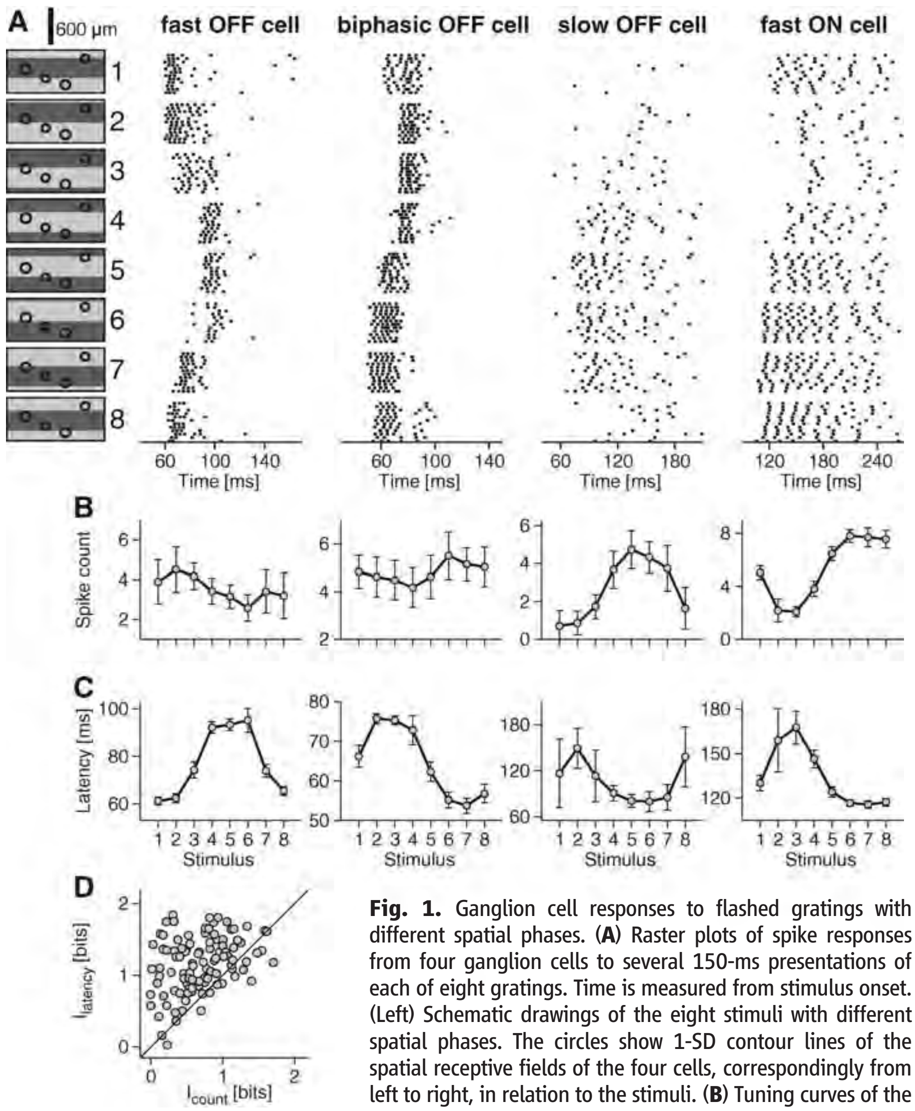

Fig. 1. Ganglion cell responses to flashed gratings with different spatial phases. (**A**) Raster plots of spike responses from four ganglion cells to several 150-ms presentations of each of eight gratings. Time is measured from stimulus onset. (Left) Schematic drawings of the eight stimuli with different spatial phases. The circles show 1-SD contour lines of the spatial receptive fields of the four cells, correspondingly from left to right, in relation to the stimuli. (**B**) Tuning curves of the elicited spike count. Here and in subsequent figures, all error bars show the standard deviation across trials with the same stimulus. (**C**) Tuning curves of the first-spike latency. "Fast OFF" and "biphasic OFF" cells typically showed strong tuning in the latency and only mild tuning in spike count; despite their names, these cell types receive input from both ON and OFF pathways (*19*). "Slow OFF" and "ON" cells, on the other hand, displayed good tuning in the spike count and often did not respond with spikes to all stimuli. The relatively long latencies are typical for cold-blooded animals. (**D**) Information about the stimulus identity contained in the spike count and in the latency, respectively, for all recorded cells. For a subdivision of the data by ganglion cell type, see fig. S2.

This hypothesis for latency coding relies intimately on the convergence of parallel neuronal pathways with intrinsic kinetic differences. If this picture is correct, removal of one of the pathways should lead to a breakdown of latency tuning. We therefore exposed the retina to 2-amino-4-phosphono-butyrate, a metabotropic glutamate receptor agonist that blocks neural transmission to ON bipolar cells (*26*). The results were as predicted: Fast OFF ganglion cells ceased responding to about half of the stimuli (fig. S8), consistent with a loss of all the ON filters of the model as shown in Fig. 3F.

Although grating stimuli are convenient for systematic investigations, they do not capture the complex statistics of natural scenes. We thus briefly flashed a photographic image onto the retina (Fig. 4A). Across repeated presentations, the image was shifted to many different locations. In this way, spikes from a single ganglion cell could be used to simulate a population of identical neurons with different receptive field locations. Responses to the natural image resembled those to the gratings. For fast OFF cells, almost all image presentations elicited spike bursts that varied in latency by about 40 ms (Fig. 4B), and this latency was systematically related to the stimulus. Indeed, by simply plotting the differential spike latencies as a gray-scale code, we obtained a rather faithful neural representation of the raw visual image (Fig. 4C). This demonstrates the high quality of the latency information. Subsequent brain regions could use this for local image computations; for example, a neuron that detects spike coincidence among multiple ganglion cells would be selective for contour lines or edges in the image.

The corresponding neural image created from spike counts (Fig. 4D) is more blurred and noisy, and the highest values are observed near edges in the stimulus. In the flat regions, the center-surround antagonism of ganglion cell receptive fields (fig. S1) reduces firing activity. But because the effect of the receptive-field surround is delayed relative to the center (*20*), it does not affect the first spike in a burst. In fact, latency and spike count may serve to encode complementary stimulus features, which could support a rapid scene analysis with subsequent refinement (*27*). Furthermore, in natural vision, the ongoing fixational eye movements after a saccade may well affect the spike count throughout fixation, but should have negligible effects on the timing of the first spike.

Altogether, our results suggest that a population code based on differential spike latencies can be a powerful mechanism to rapidly transmit a new visual scene. Rapid saccades are ubiquitous in animal vision (*1*). In salamanders, they result from turns of the head and constitute a vital part of the approach to a prey (*2*). During a saccade, many ganglion cells are strongly suppressed (*19*, *28*), such that the first spike after a saccade is easily recognized. The differential latency of these

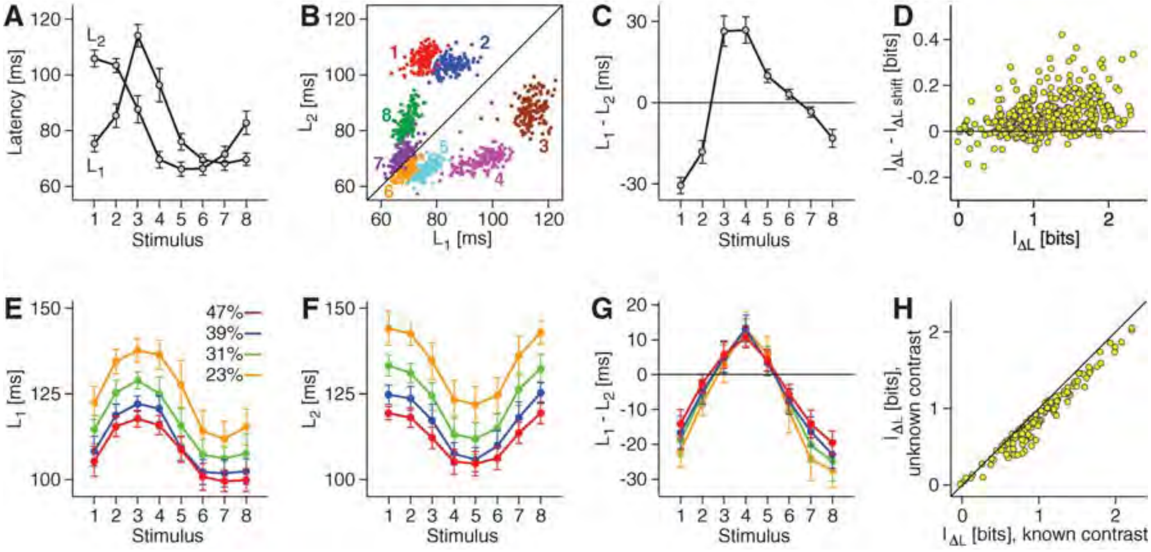

Fig. 2. (**A** to **D**) Encoding by relative latencies of pairs of ganglion cells. (A) Latency tuning curves for two simultaneously recorded fast OFF cells. (B) Scatter plot of latencies (L_1, L_2) for the two cells. The diagonally elongated distributions of the data show that L_1 and L_2 were positively correlated across trials with the same stimulus (*14*). (C) Tuning curve of the latency difference $L_1 - L_2$. (D) Information theoretical analysis of latency differences from simultaneously recorded cell pairs. The information $I_{\Delta L}$ about the stimulus contained in the latency difference $\Delta L = L_1 - L_2$ is plotted against the information loss $I_{\Delta L} - I_{\Delta L\ \mathrm{shift}}$ that occurred when L_2 was shifted by one trial with respect to L_1. (**E** to **H**) Contrast-invariant encoding by pairs of ganglion cells. (E) Latency tuning curves for a fast OFF cell whose responses were recorded for flashed gratings at different Michelson contrast levels. (F) Latency tuning curves of a second, simultaneously recorded fast OFF cell. (G) Tuning curves of the latency difference for the two neurons. (H) Information about the stimulus pattern, carried by latency differences whether the contrast level is known or not. All cell pairs that were recorded at the four different contrast levels were analyzed. The data points near the diagonal show that little information is lost by ignoring the stimulus contrast.

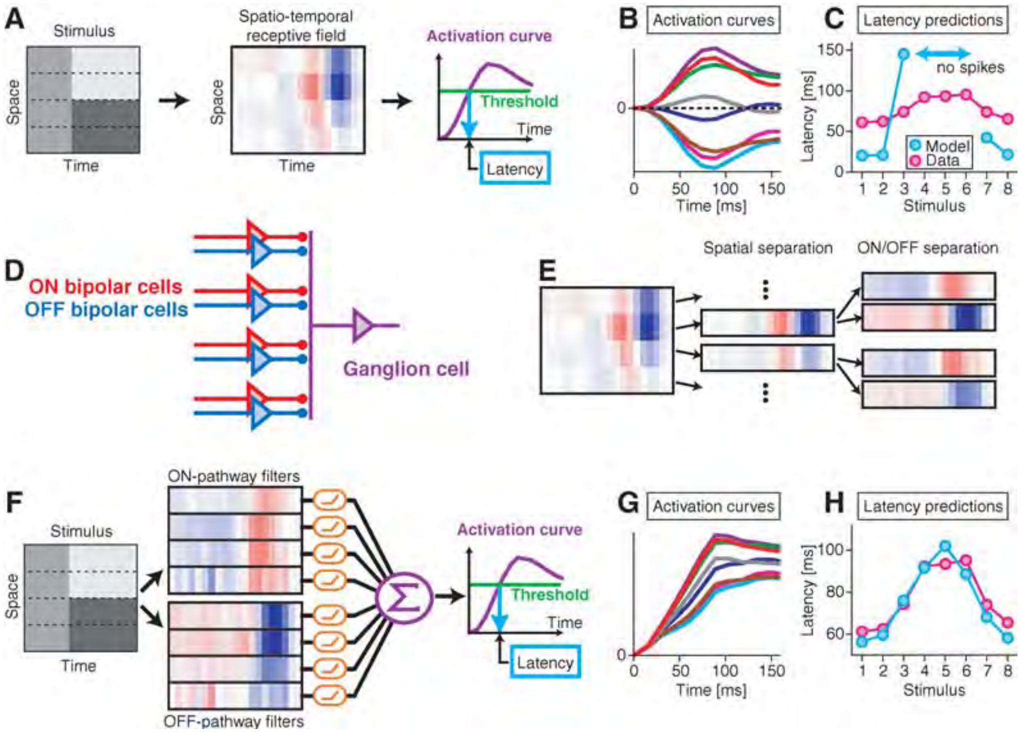

Fig. 3. Modeling the response latencies of retinal ganglion cells. (**A**) Standard framework for modeling ganglion cell responses. The stimulus (left) is gray illumination followed by a grating. This is convolved with a spatiotemporal filter (middle) representing the ganglion cell's receptive field (fig. S1). When the resulting activation curve exceeds a preset threshold, the first spike is fired (right). (**B**) Activation curves computed for each of the eight grating stimuli, by using the measured spatiotemporal filter for a sample fast OFF ganglion cell (first cell in Fig. 1). (**C**) Predicted and measured dependence of the latency on the stimulus. The threshold is the only free parameter of the model and was optimized from a χ^2 fit to the measured latency tuning curve. Several stimuli did not lead to positive activation and thus did not predict spikes. (**D**) Retinal interneuron pathways that motivate a revised model. Each small subregion of the receptive field activates both ON and OFF bipolar cells. The ganglion cell pools inputs across subregions and from both bipolar types. (**E**) Separation of the spatiotemporal filter into spatial subunits and subsequently into ON and OFF pathway contributions. [See (*14*) and fig. S5 for measurement of these contributions.] (**F**) Multi-pathway model of the response: Each subregion of the stimulus is passed through an ON filter and an OFF filter. The filter output is half-wave rectified and then pooled with all other outputs to yield the activation curve. (**G**) Activation curves computed for each of the eight grating stimuli, using the model in (F). Note that each stimulus produced excitation. (**H**) Measured latency tuning curves and predictions of the model in (F), after optimization of the threshold.

spikes encodes fine spatial detail (Figs. 1 and 4), yet it is almost entirely invariant to the overall stimulus contrast level (Fig. 2). Furthermore, it is robust to retinal noise (Fig. 2), and it provides information about the pattern in the shortest possible time, namely, with the very first spikes. Many vertebrates have specific ganglion cells, often multiple types, that combine inputs from both ON and OFF pathways (*20*–*23*), as evidenced by their ON/OFF response characteristics or by their dendritic morphology that connects to both ON and OFF bipolar cells. These neurons are candidate carriers of a latency code (Fig. 3). Synapses in the early visual pathway are very efficient, such that short spike bursts are reliably transmitted from retina to cortex (*29*). Moreover, certain neurons in visual cortex are exquisitely sensitive to the coincidence of spikes on their

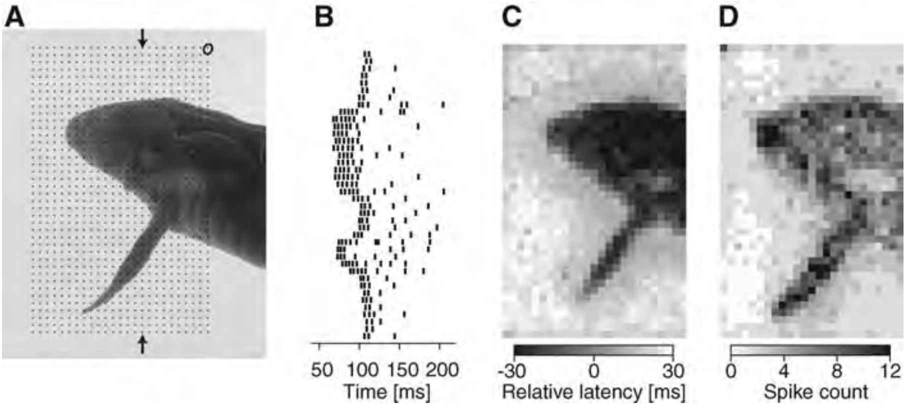

Fig. 4. Responses of a fast OFF ganglion cell to a flashed natural image. (For results from other cell types, see fig. S9.) (**A**) Photograph of a swimming salamander larva projected on the retina. The ellipse in the upper right corner shows a sample 1-SD outline of a ganglion cell receptive field. In each of 1000 presentations, the image was shifted slightly, and the grid of dots marks the resulting centers of the receptive field. Presentations were separated by gray illumination at the mean intensity of the photograph. The image onset produced luminance changes at most locations. (**B**) Spike trains of the ganglion cell for receptive-field locations along the column marked by the arrows in (A). (**C**) Gray-scale plot of the differential spike latency on single-trial presentations at the locations marked with dots in (A). The reference latency was chosen as the average value at all locations (*10*). (**D**) Corresponding gray-scale plot of the spike counts.

afferents (*30*), which is one possible readout mechanism for a latency code. Cortical neurons themselves carry substantial sensory information in their response latencies (*6*, *7*, *31*). Thus, it is conceivable that early aspects of sensory processing operate on the basis of the classification of spike latency patterns.

References and Notes

1. M. F. Land, *J. Comp. Physiol. A* **185**, 341 (1999).
2. C. Werner, W. Himstedt, *Zool. Jahrb. Abt. Allg. Zool. Physiol. Tiere* **89**, 359 (1985).
3. S. Thorpe, D. Fize, C. Marlot, *Nature* **381**, 520 (1996).
4. M. Meister, M. J. Berry, *Neuron* **22**, 435 (1999).
5. V. J. Uzzell, E. J. Chichilnisky, *J. Neurophysiol.* **92**, 780 (2004).
6. T. J. Gawne, T. W. Kjaer, B. J. Richmond, *J. Neurophysiol.* **76**, 1356 (1996).
7. D. S. Reich, F. Mechler, J. D. Victor, *J. Neurophysiol.* **85**, 1039 (2001).
8. M. Greschner, A. Thiel, J. Kretzberg, J. Ammermüller, *J. Neurophysiol.* **96**, 2845 (2006).
9. N. B. Sawtell, A. Williams, C. C. Bell, *Curr. Opin. Neurobiol.* **15**, 437 (2005).
10. S. M. Chase, E. D. Young, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5175 (2007).
11. J. J. Hopfield, *Nature* **376**, 33 (1995).
12. E. J. Chichilnisky, *Network* **12**, 199 (2001).
13. J. Keat, P. Reinagel, R. C. Reid, M. Meister, *Neuron* **30**, 803 (2001).
14. Materials and methods are available as supporting material on *Science* Online.
15. W. A. Hare, W. G. Owen, *J. Neurophysiol.* **76**, 2005 (1996).
16. S. A. Baccus, M. Meister, *Neuron* **36**, 909 (2002).
17. J. D. Victor, R. M. Shapley, *J. Gen. Physiol.* **74**, 671 (1979).
18. J. B. Demb, K. Zaghloul, L. Haarsma, P. Sterling, *J. Neurosci.* **21**, 7447 (2001).
19. M. N. Geffen, S. E. de Vries, M. Meister, *PLoS Biol.* **5**, e65 (2007).
20. F. S. Werblin, J. E. Dowling, *J. Neurophysiol.* **32**, 339 (1969).
21. F. M. de Monasterio, *J. Neurophysiol.* **41**, 1435 (1978).
22. F. R. Amthor, E. S. Takahashi, C. W. Oyster, *J. Comp. Neurol.* **280**, 97 (1989).
23. D. A. Burkhardt, P. K. Fahey, M. Sikora, *Vis. Neurosci.* **15**, 219 (1998).
24. J. L. Coombs, D. Van Der List, L. M. Chalupa, *J. Comp. Neurol.* **503**, 803 (2007).
25. J. F. Ashmore, D. R. Copenhagen, *Nature* **288**, 84 (1980).
26. X. L. Yang, *Prog. Neurobiol.* **73**, 127 (2004).
27. M. D. Menz, R. D. Freeman, *Nat. Neurosci.* **6**, 59 (2003).
28. B. Roska, F. Werblin, *Nat. Neurosci.* **6**, 600 (2003).
29. P. Kara, R. C. Reid, *J. Neurosci.* **23**, 8547 (2003).
30. W. M. Usrey, J. M. Alonso, R. C. Reid, *J. Neurosci.* **20**, 5461 (2000).
31. S. Panzeri, R. S. Petersen, S. R. Schultz, M. Lebedev, M. E. Diamond, *Neuron* **29**, 769 (2001).

32. We thank F. Engert and members of the Meister laboratory for advice. This work was supported by grants from the National Eye Institute (M.M.) and the Human Frontier Science Program Organization (T.G.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/319/5866/1108/DC1
Materials and Methods
Figs. S1 to S10
References

23 August 2007; accepted 11 December 2007
10.1126/science.1149639

Predicting Human Interactive Learning by Regret-Driven Neural Networks

Davide Marchiori[1] **and Massimo Warglien**[2]*

Much of human learning in a social context has an interactive nature: What an individual learns is affected by what other individuals are learning at the same time. Games represent a widely accepted paradigm for representing interactive decision-making. We explored the potential value of neural networks for modeling and predicting human interactive learning in repeated games. We found that even very simple learning networks, driven by regret-based feedback, accurately predict observed human behavior in different experiments on 21 games with unique equilibria in mixed strategies. Introducing regret in the feedback dramatically improved the performance of the neural network. We show that regret-based models provide better predictions of learning than established economic models.

The surge of interest in the neural bases of economic behavior (*1*–*3*) prompts the question of how well neural networks can model human interactive decision-making (*4*). This question implies two issues: the choice of the network architecture and the selection of input information to the network that has to be both economically and neurophysiologically motivated.

Interactive learning differs from individual learning in that, given *n* agents, each agent adapts to behaviors that are modified by the concurrent learning of the other *n*–1 agents. It has an obvious relevance in economic contexts, but (more generally) much of human learning that occurs in social contexts has an interactive nature. Experimental game theory has provided a large set of laboratory data on human interactive learning in repeated games (*5*), often contradicting the predictions of standard game theory. The need for models of interactive learning in games arises from the difficulties of ordinary game-solution concepts to explain both the trajectories and the long-run stationary state of experimentally observed human behavior in repeated games. Games with unique equilibria in mixed strategies are an especially interesting case, because Nash equilibrium not only fails to approximate behavior in early rounds but also is often a poor predictor of the stable behavior emerging in the long run.

Until now, two main modeling strategies have been used with some success in trying to fit and predict how humans learn in repeated games in a laboratory setting. One modeling strategy extends a classical paradigm of learning theory (i.e., rein-

[1] Interdepartmental Center for Research Training in Economics and Management (CIFREM), University of Trento, Italy. [2] Advanced School of Economics and Department of Business Economics and Management, Ca' Foscari University, Venezia, Italy.

*To whom correspondence should be addressed. E-mail: warglien@unive.it

Supporting Online Material

T. Gollisch & M. Meister, “Rapid Neural Coding in the Retina with Relative Spike Latencies”

Materials and Methods

Recording. Retinas were isolated from larval tiger salamanders, superfused with oxygenated Ringer's medium at room temperature, and placed ganglion cell layer down on a multi-electrode array, which recorded spike trains from many ganglion cells simultaneously, as described previously (*S1*).

In experiments to test the role of inputs from ON bipolar cells (Fig. S8), 200 µM of 2-amino-4-phosphono-butyrate (APB) was added to the Ringer’s medium. APB is a metabotropic glutamate receptor agonist that blocks light-induced activation of ON bipolar cells (*S2*). Although APB acts on multiple metabotropic receptors in the retina, the selective loss of ON responses is thought to derive from its action on the ON-bipolar dendrites (*S3*).

Stimulation. Visual images were projected onto the photoreceptor layer of the retina via a computer monitor with a frame rate of 66 Hz. White light was used with an average image luminance of $I_0 = 16\ \mathrm{mW/m^2}$, in the photopic range. A gray screen (750 ms) was followed by a square-wave grating (150 ms). The bar width of the grating was 330 µm on the retina, somewhat larger than a typical ganglion-cell receptive field center. The eight different grating versions were obtained by successively shifting the grating by one fourth of the bar width. The minimal shift between gratings was 82.5 µm, considerably smaller than most ganglion-cell receptive field centers. The light and dark bars had intensity values $I_0 + I_1$ and $I_0 - I_1$, and the quoted contrast values are Michelson contrast, I_1/I_0. The gratings were presented in pseudo-random order, mixing the spatial phases of the grating as well as the contrast levels in experiments with varying contrast.

To test responses to natural images, digital photos of a swimming salamander larva were converted to gray-scale, scaled to an apparent distance of about 5 cm, and displayed for 150 ms in the same way as the gratings. The same photograph was presented repeatedly with different spatial translations, so that for each presentation a given ganglion cell responded to a different patch of the photograph. Individual presentations were separated by 750 ms of gray illumination at the mean intensity of the photograph. These single-neuron data can be used to simulate responses from a population of identical neurons with different receptive field centers (*S4, S5*).

Spike train analysis. Spikes were sorted off-line by a cluster analysis of their shapes. Only units corresponding to well separated spike clusters with a clear refractory period were included in further analysis. In experiments with flashed gratings, spike latency was measured as the time of the first spike after stimulus onset. Spike count was measured as the number of spikes in a window of 220 ms following stimulus onset. The length of this window was chosen to generally include all spikes in the burst elicited by stimulus onset and to exclude spikes elicited by the offset of the 150-ms presentation of the grating. In the experiments of Fig. 1, 68 of 100 measured cells responded to each grating with at least one spike in at least 50% of the trials, and 54 of those responded with a spike in at least 90% of the trials for each grating.

Latency correlations. Fig. 2B illustrates that the spike latency fluctuates across trials with the same stimulus, but the latencies L_1 and L_2 of two neurons were often positively correlated. The average correlation coefficient for the latencies over all stimuli and all cell pairs was 0.25 ± 0.04 (SD). For comparison, the average spike count correlation coefficient was only 0.07. A shift control showed that the correlations in spike latency were not caused by slow trends over the course of the experiment: when L_2 was shifted by one trial with respect to L_1, the average latency correlation dropped to 0.05.

Receptive fields. Spatiotemporal receptive fields were measured by reverse correlation to a stimulus with parallel stripes that flickered randomly and independently with intensities drawn from a Gaussian distribution. The stripes were oriented parallel to the bars of the flashed gratings used in the main experiment and were arranged so that each bar of the grating covered exactly four of the stripes. Spatio-temporal receptive fields with a single spatial dimension were obtained by calculating the spike-triggered average (*S1*), i.e., the average sequence of stripe patterns that preceded a spike (see Fig. S1).

To classify neurons into cell types, the measured spatio-temporal receptive field was approximated as a product of a spatial and a temporal filter, which were estimated from a singular-value decomposition (*S6*). The temporal filter was then used for a cluster analysis (*S7*). We applied agglomerative maximum-linkage clustering: Initially, every filter formed its own cluster. In each iteration, the two clusters with the smallest maximal Euclidean distance between two of their elements were merged.

Although the temporal filters generally seemed to form a continuum, five cell types could be separated that led to a fairly good match with previous results (*S7*): fast ON, slow ON, biphasic OFF, fast OFF, and slow OFF. The latter two classes correspond to a combination of the monophasic, medium, and slow OFF cells in ref. (*S7*). Figure S2 shows the collections of temporal filters for the five cell classes and the analysis of latency and spike count tuning curves for these types. Note that both fast and slow ON cells have considerably slower dynamics than biphasic and fast OFF cells.

Many of the cells called “biphasic OFF” and “fast OFF” are now known to have sizeable responses to ON stimuli (*S8*). For most of these, however, the response characteristics are dominated by the OFF inputs. Furthermore, the shorter latency of the OFF pathway always leads to a primary peak with OFF characteristics in the reverse-correlation analysis (Fig. S1). For these reasons, we adhered to the “fast OFF” nomenclature introduced in prior studies.

For comparison with the flashed natural images, two-dimensional spatial receptive fields were obtained by reverse correlation with a flickering checkerboard stimulus and subsequently factoring the obtained spatio-temporal receptive field into a spatial and a temporal filter via singular-value decomposition. A Gaussian fit to the spatial receptive field is shown by the 1-SD contour line in Fig. 4A.

Information theory. The information I_{count} contained in the spike count about the identity of the presented stimulus was estimated according to $I_{count} = H(count) - H(count \mid stimulus)$, where $H(count)$ is the total entropy of the spike count distribution and $H(count \mid stimulus)$ is the mean entropy of the spike count given the stimulus, also called "noise entropy" (*S9*, *S10*). Entropies were calculated by computing the sampling frequencies f_k of obtaining k spikes in the response and using the formula $H = -\sum_k f_k \log_2 f_k$. Bias correction for finite sampling was performed by computing information values for several smaller fractions of the data and extrapolating to the limit of infinite data (*S11*) (see Fig. S10).

To estimate the information contained either in the absolute latency of a cell or in latency differences, we used the fact that for a given stimulus s, the distribution of latencies L could be well fitted by a Gaussian curve $g_s(L)$. Numerical integration of the formula

$$I = \sum_s \int dL\, p(s) p(L \mid s) \log_2 \frac{p(L \mid s)}{p(L)} \qquad (1)$$

was then applied to obtain information values (*S10*). Here, $p(s) = 1/8$ is the probability of stimulus s, $p(L \mid s) = g_s(L)$ is the probability density of latency L given stimulus s, and $p(L) = \sum_s p(s) \cdot p(L \mid s) = \sum_s 1/8 \cdot g_s(L)$ is the total probability density of latency L, obtained as the normalized sum of the Gaussian fits $g_s(L)$. For stimuli that included responses without any spikes, the absence of spikes was treated in the calculation as an additional symbol, effectively including a delta-function in $g_s(L)$ at infinite latency.

By performing the analysis on smaller fractions of data and extrapolating to infinite data, the obtained values were also corrected for sampling bias, which can occur from statistical errors in the fitted Gaussian curves, but the necessary corrections were typically only ~1% (see Fig. S10).

For the information theoretical assessment of contrast-invariant coding (Fig. 2H), we analyzed cell pairs whose latencies had been recorded at four different contrast levels. We computed the information $I_{\Delta L}$ that the latency difference ΔL transmits about the spatial phase of the grating, under conditions where contrast is either known or unknown. The latter (Fig. 2H, y-axis) assumes that the decoder of the latencies relies solely on ΔL: $I_{\Delta L}$ was computed by Eq. (1) after pooling the data from all contrast levels. The contrast level was then ignored in the calculation. The former (Fig. 2H, x-axis) assumes that the decoder can use the information about the contrast level of each stimulus; here $I_{\Delta L}$ was computed independently for each contrast and then averaged. As contrast and phase were assigned to the stimuli independent of each other, this provided a simple way of calculating the information about the grating phase contained in the combination of contrast level and latency.

For an alternative information theoretical assessment and comparison with information transmission by single-cell latencies, see Fig. S4.

Modeling. In order to test whether a model based on known properties of the retina's circuitry can account for the observed response latencies, we explored different ways in which ganglion cells can integrate sensory information over their receptive fields. This can be formulated mathematically as a cascade model, as depicted schematically in Fig. S6A. In this model framework, the flashed grating stimulus is integrated into an activation curve.

This stimulus integration is achieved by spatio-temporal filtering based on the ganglion cell's receptive field. Crossing of a threshold by the activation curve determines the time of the first spike, and thus the response latency. Because this analysis focuses on the timing of the first spike only, typical extensions of the filter/threshold model that use feedback dynamics to take the recent spiking history into account (*S12, S13*) need not be considered.

In the main text, we presented two models that were based on standard linear stimulus integration (Fig. 3A) and on spatially local ON and OFF filters with subsequent rectification (Fig. 3F), respectively. Here, we present these models as part of a more general framework in which stimulus integration can be either linear or nonlinear across spatial subfields and also across ON and OFF pathways, resulting in four candidate models shown in Fig. S6B.

In all four cases, the starting point to model stimulus integration for a given neuron is its spatio-temporal receptive field as measured by reverse correlation (Fig. S1). To model nonlinear spatial stimulus integration (Models 2 and 4), the receptive field was split up into local subfields. The size of the spatial subunit was chosen as the minimal shift in spatial phase of the grating, which was ~80 μm on the retina. This corresponds approximately to the receptive field size of bipolar cells (*S14*), the putative origin of spatial subunits (*S15*), and is therefore a reasonable spatial scale for the model subunits.

If the ON and OFF pathways are pooled linearly (Model 1 and 2), they effectively act as a single filter, whose time course corresponds to the receptive field. In order to incorporate nonlinear integration over ON and OFF pathways (Model 3 and 4), we needed to split up the receptive field into its ON and OFF components. For this purpose, we extended the method of ref. (*S8*) that identifies which spikes are triggered through the ON and the OFF pathways, as described below ("Obtaining the Model Parameters") and in Fig. S5. This yields separate ON and OFF filters for each spatial location in the receptive field (Fig. S5E).

For fully linear stimulus integration (Model 1), the activation curve was obtained by simply convolving the stimulus with the spatio-temporal receptive field of the cell (*S16*). Nonlinear integration of spatial subunits (Model 2) was considered by convolving the stimulus separately for each spatial location with the corresponding temporal component of the receptive field. Half-wave rectification and summation of these subfield activations then resulted in the activation curve for this model. For nonlinear integration of ON and OFF signals (Model 3), the outputs of the obtained ON and OFF filters were linearly summed over space, but separately for the ON and OFF pathway and then half-wave rectified before being combined. Finally, nonlinear integration over spatially local ON and OFF subfields (Model 4) was obtained by half-wave rectifying the output of each ON and OFF filter individually before summation.

More formally, the activation curve $a(t)$ was computed as follows: We denote the spatio-temporal receptive field of a neuron (cf. Fig. S5A) by $f_x(t)$, where the spatial index *x* enumerates the stripes on the retina for which the receptive field has been measured and *t* is the time relative to the spike. The ON and OFF fields are similarly denoted as $f_x^{(ON)}(t)$ and $f_x^{(OFF)}(t)$, respectively (cf. Fig. S5E). As explained below, these fields are normalized so that $f_x(t) = f_x^{(ON)}(t) + f_x^{(OFF)}(t)$. We use the notation $f_x^{(p)} * s_x(t)$ to denote the temporal convolution with the stimulus $s_x(t)$:

$$f_x^{(p)} * s_x(t) = \int_{-\infty}^{0} dt' \, f_x^{(p)}(t') \cdot s_x(t+t'), \tag{2}$$

where $p \in \{ON, OFF\}$. Furthermore, half-wave rectification is implemented by the function

$$N(x) = \begin{cases} 0 & x \le 0 \\ x & x > 0 \end{cases} . \tag{3}$$

With these conventions, the activation curves for the four models are:

$$\begin{aligned} &\text{Model 1:} \quad a_1(t) = \sum_x \sum_p f_x^{(p)} * s_x(t) \\ &\text{Model 2:} \quad a_2(t) = \sum_x N\left(\sum_p f_x^{(p)} * s_x(t) \right) \\ &\text{Model 3:} \quad a_3(t) = \sum_p N\left(\sum_x f_x^{(p)} * s_x(t) \right) \\ &\text{Model 4:} \quad a_4(t) = \sum_x \sum_p N\left(f_x^{(p)} * s_x(t) \right) \end{aligned} \tag{4}$$

Note that the four models for stimulus integration are all based on the same measured spatio-temporal field and only differ in how they employ this measurement for predicting the cell's activation. No free parameters are introduced by the half-wave rectification or by splitting up the measured receptive field in the ways described. Only the threshold value is a free parameter and was determined for each model individually by minimizing the χ^2 deviation between the predicted and the measured latency tuning curve.

Figure S6 shows how the four models can be interpreted in terms of retinal circuitry, together with the performance of the models for a sample cell. Linear pooling of bipolar inputs over spatial subregions (Model 2) as well as over ON and OFF inputs separately (Model 3) are biologically plausible, considering the diversity of branching patterns of ganglion cell dendritic trees. These can display branches that spread into local regions of the retina or ramify in different sublaminae of the inner-plexiform layer where synaptic terminals from either ON- or OFF-bipolars accumulate (*S17, S18*). In Models 2 and 3, rectification may result from dendritic processing in the ganglion cell after selective pooling of bipolar inputs. In Model 4, rectification could also occur presynaptically, for example in the voltage-dependent control of transmitter release (*S19*).

Note that the incorporation of separate filters for the ON and OFF pathway is essential for a good fit to the latency data. By contrast, other reports of ganglion cell light responses (*S12, S13*) built successful models with just a single filter. These studies used a very different visual stimulus: spatially homogeneous white noise. Such a stimulus produces transient activations of either the ON or the OFF pathway, but not both at the same time, which allows a single filter that combines the effect of both pathways to capture the neural responses to a good degree. By contrast, the flashed gratings activate ON and OFF pathways together, which reveals their nonlinear interaction.

We have restricted ourselves to modeling the first-spike latency in response to a flashed stimulus. Other approaches have aimed at capturing the whole time course of the firing rate (*S20, S21*) or even the occurrence of individual spikes throughout the response (*S12*). Doing so requires a detailed account of the spike generation mechanisms, in particular the effects of previous spikes, neural refractoriness, gain control, and adaptation. In principle, the type of model discussed in the present work could be used as a front end to an existing spike-generation model, such as that of ref. (*S12*). This would allow the predictions of response characteristics beyond the first-spike latency, but would require a larger number of parameters to be determined from the data, which is beyond the scope of the current experiments and a topic of ongoing research.

Obtaining the model parameters. In order to compare measured latencies to the predictions of the different model versions, the filters corresponding to the ON and OFF pathways and the threshold values had to be determined for each ganglion cell individually. In each case, the starting point was the measurement of the spatio-temporal receptive field as obtained from the spike-triggered-average analysis under stimulation with flickering stripes (Fig. S1).

As a first step, we determined which stimulus stripes (i.e., locations on the retina) are relevant for the subsequent analysis. Locations far outside the ganglion cell's receptive field do not contribute to the cell's activation, but can add noise to the model predictions because of stochastic fluctuations in the estimation of the receptive field. To reduce this noise source, we disregarded stimulus stripes that did not significantly affect the spike probability. Significance was established in the following way: For a given stripe, the similarity of each 300-ms stimulus segment to the spike-triggered average at this location was calculated as a dot-product between the stimulus segment and the spike-triggered average. For the 5% of stimulus segments with the highest similarity, the number of elicited spikes was counted. If this differed by more than 3 standard deviations from the average spike count obtained after randomizing spike times, the corresponding stripe was considered to significantly contribute to the spike probability. Typically, this procedure resulted in a broad region of significant stripes around the receptive field center plus occasional spurious, isolated stripes. In all four models, only these significant stripes were used to calculate the activation curves.

Having obtained the set of significant stripes, we needed to calculate the corresponding stimulus filters. For the model versions that integrated the ON and OFF pathways linearly (Model 1 and Model 2), the spike-triggered average for each significant stripe was directly used as a filter in the first model stage. For the model versions with nonlinear integration over ON and OFF pathways (Model 3 and Model 4), this spike-triggered average was split up into contributions from the two pathways. The technique for this is an extension of the methodology used in ref. (*S8*) and illustrated in Fig. S5. For each significant stripe, the sequences of luminance values of the 20 stimulus frames (300 ms) preceding each recorded spike were collected. These luminance sequences were considered as points in a 20-dimensional space and subjected to a principal component analysis, yielding the covariance matrix and the corresponding eigenvalues and eigenvectors.

The luminance sequences typically came in two clusters as exemplified in Fig. S5C, which we attributed to stimuli that activate either the ON or OFF pathway. To automatically separate the two clusters, we calculated the projection of each luminance sequence onto the first principal component. This vector denotes the dimension along which the distribution of data points has the largest variance and is therefore well suited to separate the clusters. The luminance sequences were then assigned to two clusters, depending on whether this projection was positive or negative. We then obtained two filter shapes by averaging the luminance sequences within each cluster. These two filters were assigned to the ON and OFF pathway, respectively; the filter with the larger integral over the first 150 ms was used as the ON filter. (Typically, this integral was positive for the ON filter and negative for the OFF filter.) Finally, the two filters were normalized so that their sum equaled the spike-triggered average computed over all spikes. This is uniquely possible and easily done because each filter is proportional to the sum of complementary subsets of luminance sequences, and the spike-triggered average is proportional to the sum of all luminance sequences. This normalization accounts for the relative contributions of the pathways to the ganglion cell's activation. The above procedure was performed for each stimulus stripe that had been found significant.

After thus obtaining all the filters of the models from the analysis of the responses under flickering stripes, the only remaining parameter was the threshold value for the activation curve. This was optimized for each neuron and each model version separately by minimizing the χ^2 deviation between the measured latencies and the model prediction. For the purpose of this fitting procedure, stimuli that did not result in any spikes within the observation window (either in the experiment or in the model prediction) were assigned the length of the observation window (220 ms) as a latency.

References

S1. M. Meister, J. Pine, D. A. Baylor, *J Neurosci Methods* **51**, 95 (1994).

S2. M. M. Slaughter, R. F. Miller, *Science* **211**, 182 (1981).

S3. X. L. Yang, *Prog Neurobiol* **73**, 127 (2004).

S4. F. Ratliff, H. K. Hartline, *J Gen Physiol* **42**, 1241 (1959).

S5. A. L. Jacobs, F. S. Werblin, *J Neurophysiol* **80**, 447 (1998).

S6. W. H. Press, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge; New York, 1989).

S7. R. Segev, J. Puchalla, M. J. Berry, *J Neurophysiol* **95**, 2277 (2006).

S8. M. N. Geffen, S. E. de Vries, M. Meister, *PLoS Biol* **5**, e65 (2007).

S9. F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, *Spikes: Exploring the Neural Code* (The MIT Press, 1999).

S10. A. Borst, F. E. Theunissen, *Nat Neurosci* **2**, 947 (1999).

S11. S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, W. Bialek, *Phys Rev Lett* **80**, 197 (1998).

S12. J. Keat, P. Reinagel, R. C. Reid, M. Meister, *Neuron* **30**, 803 (2001).

S13. J. W. Pillow, L. Paninski, V. J. Uzzell, E. P. Simoncelli, E. J. Chichilnisky, *J Neurosci* **25**, 11003 (2005).

S14. W. A. Hare, W. G. Owen, *J Neurophysiol* **76**, 2005 (1996).

S15. J. B. Demb, K. Zaghloul, L. Haarsma, P. Sterling, *J Neurosci* **21**, 7447 (2001).

S16. E. J. Chichilnisky, *Network* **12**, 199 (2001).

S17. F. R. Amthor, E. S. Takahashi, C. W. Oyster, *J Comp Neurol* **280**, 97 (1989).

S18. H. Wässle, *Nat Rev Neurosci* **5**, 747 (2004).

S19. R. Heidelberger, G. Matthews, *J Physiol* **447**, 235 (1992).

S20. R. W. Rodieck, *Vision Res* **5**, 583 (1965).

S21. J. D. Victor, *J Physiol* **386**, 219 (1987).

S22. F. S. Werblin, J. E. Dowling, *J Neurophysiol* **32**, 339 (1969).

S23. C. Enroth-Cugell, J. G. Robson, D. E. Schweitzer-Tong, A. B. Watson, *J Physiol* **341**, 279 (1983).

S24. O. Schwartz, J. W. Pillow, N. C. Rust, E. P. Simoncelli, *J Vis* **6**, 484 (2006).

S25. D. A. Burkhardt, P. K. Fahey, M. Sikora, *Vis Neurosci* **15**, 219 (1998).

S26. J. F. Ashmore, D. R. Copenhagen, *Nature* **288**, 84 (1980).

S27. S. Nawy, *J Neurosci* **19**, 2938 (1999).

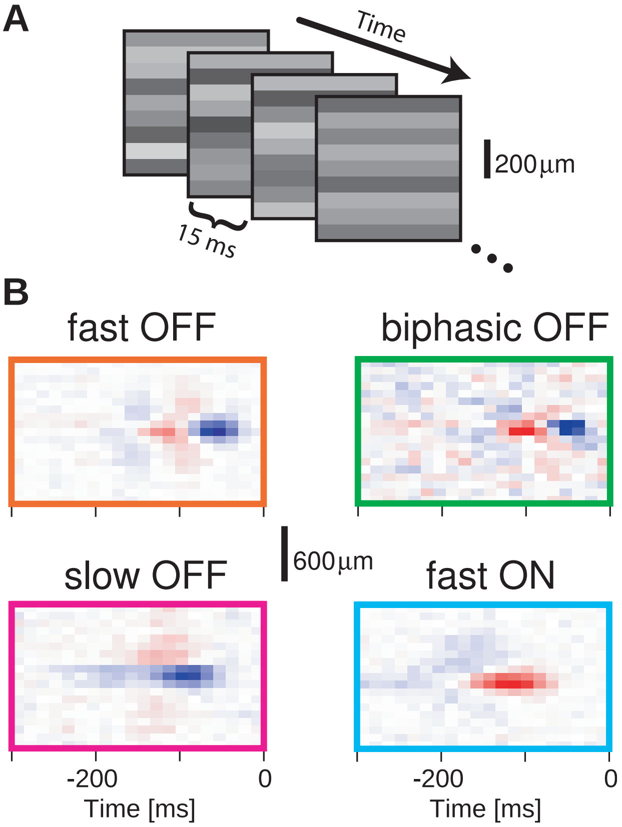

Fig. S1. Measurements of spatio-temporal receptive fields. (**A**) Schematic depiction of stimuli. At a frame rate of 66 Hz, each frame consisted of a gray-scale stripe pattern for which the intensity of each stripe was drawn randomly from a Gaussian distribution. The width of a single stripe spanned ∼80 μm on the retina. (**B**) Measured spatio-temporal receptive fields for the four sample cells shown in Fig. 1. The color code indicates the average stimulus that preceded a spike, where the abscissa denotes time and the ordinate space (in the direction orthogonal to the stripes in the stimulus). Blue regions correspond to lower than average light intensity, red regions higher than average. The receptive fields of the fast OFF and the biphasic OFF cells were biphasic over time. The slow OFF cell and the fast ON cell had more monophasic receptive fields that were broader over time and peaked at larger temporal lags. Most cells displayed the typical spatial center-surround structure. In the surround, the receptive field had the opposite sign from the center and showed a longer response delay by several tens of milliseconds. This delay of the antagonistic surround (*S22, S23*) is the likely reason why the first-spike latency values did not have the same suppression in the center of a homogeneous image surface as did the spike count (cf. Fig. 4).

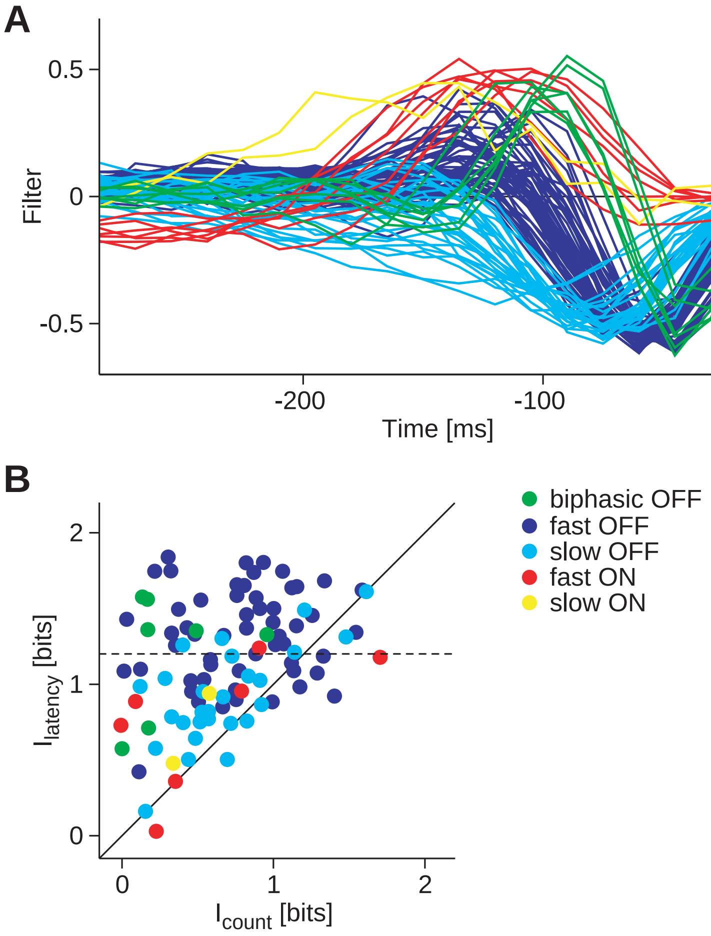

Fig. S2. Cell-type classification. (**A**) Temporal filters of all measured cells, color-coded according to cell type, which was determined by a cluster analysis. Two clusters correspond to fast and slow ON cells; the other three clusters are OFF cells with strongly biphasic filters (biphasic OFF cells), fast and moderately biphasic filters (fast OFF) and slow, monophasic filters (slow OFF). (**B**) Comparison of information values transmitted by the spike count and by the first-spike latency resolved for the five cell types for all cells for which a receptive field was reliably determined (98 of 100 cells). Data are the same as presented in Fig. 1D. The dashed line indicates the cutoff used in the analysis shown in Fig. S7B.

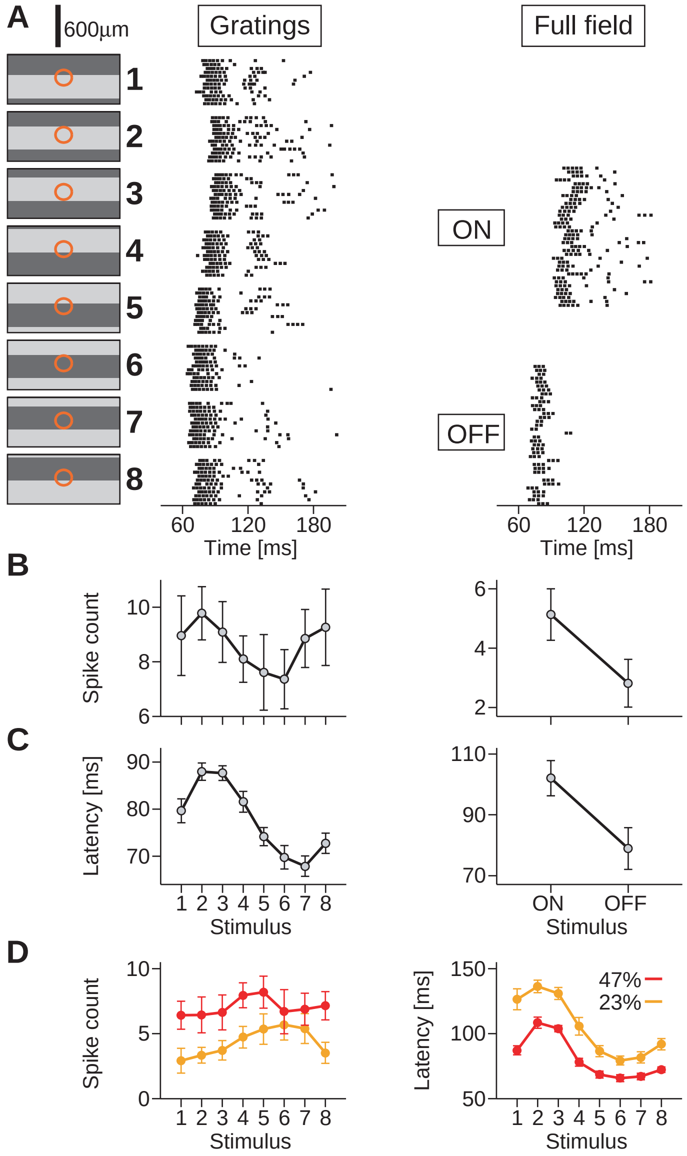

Fig. S3. (**A**-**C**) Responses to flashed gratings and full-field flashes for a sample neuron (biphasic OFF cell) that shows inverse tuning for spike count and latency: fewer spikes occurred together with shorter latencies. (**A**) Left and middle: Spike rasters in response to flashed gratings, presented as in Fig. 1A. Right: Response to homogeneous full-field ON and OFF flashes, presented for 150 ms in the same way as the gratings. (**B**) Spike-count tuning curves, displayed as in Fig. 1B. (**C**) Latency tuning curves, displayed as in Fig. 1C. For the responses to gratings as well as to full-field flashes, the tuning curves of this cell show shorter latencies when fewer spikes were elicited. (**D**) Spike-count and latency tuning curves from a different ganglion cell that was recorded for flashed gratings at different Michelson contrast levels (47% and 23%). For given contrast, the spike-count tuning curve is nearly flat, but changing the contrast (and thus the strength of stimulation) substantially affects spike counts.

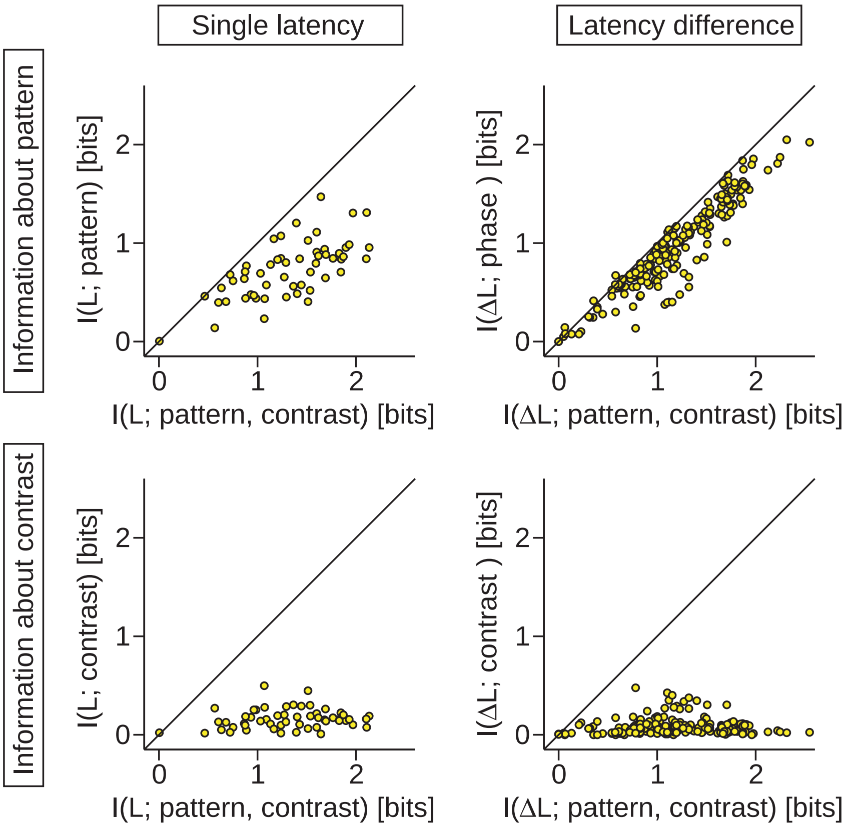

Fig. S4. Information transmission under varying contrast for single-cell latencies and latency differences between pairs of simultaneously recorded neurons. For neurons that were stimulated with eight different patterns (i.e. phases of the grating) at four different contrast levels (47, 39, 31, and 23%), we calculated the mutual information I that the latency L (left) or the latency difference ΔL (right) conveyed about the stimulus pattern (top) or the contrast level (bottom). These information values are plotted, for comparison, against the information that is transmitted about the combination of pattern and contrast. This shows that, for many cell pairs, nearly all information in the latency difference is about which pattern was presented, whereas hardly any information is transmitted about the contrast level.

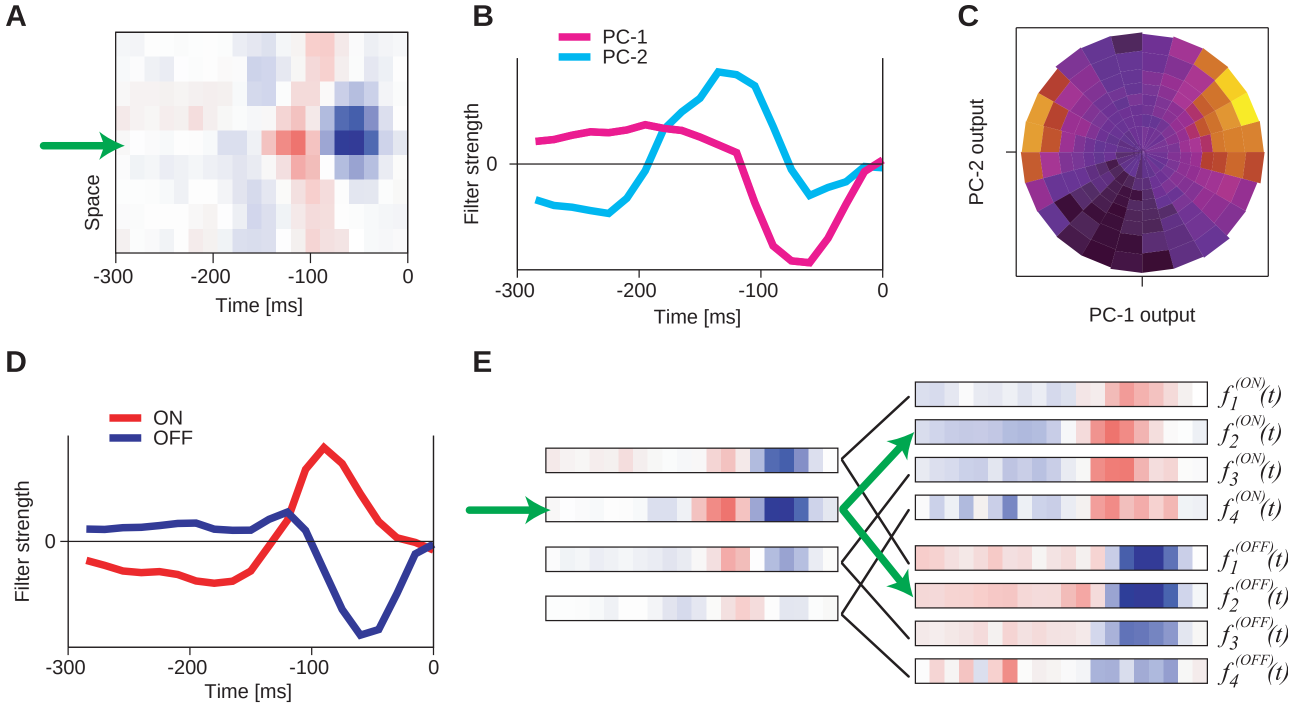

Fig. S5. Separation of the receptive field into ON and OFF filters that were used for modeling the ganglion-cell latencies. (**A**) Spatio-temporal receptive field of a fast OFF ganglion cell (same as in Fig. 3) measured by reverse correlation with flickering stripe patterns. Each row of the receptive field corresponds to the temporal receptive field at a particular spatial location. The goal is to subdivide this temporal receptive field into its ON and OFF components. The following example shows this separation for the location marked by the arrow. (**B**) The two most relevant temporal filters for the selected location. To obtain these filters, the stimulus sequences at this location that preceded spikes (the "spike-triggered stimulus ensemble") were collected and subjected to a principal component analysis (*S24*). The first two principal components PC-1 and PC-2 denote the stimulus features for which the variance of the spike-triggered stimulus ensemble differed most from the variance of the full stimulus ensemble. (**C**) Density plot of the spike probability as a function of the stimulus projection onto PC-1 and PC-2. The stimuli that elicited spikes formed two clusters, located at high and low values of the projection onto PC-1. The second principal component contributed little to the separation of the clusters, so that the clusters could be well separated according to their projection onto PC-1. (**D**) Recomputation of the spike-triggered average stimulus, separately for each of the clusters from (C). One has a typical OFF pathway shape, the other a typical ON pathway shape (cf. Fig. S2A). Note that the OFF pathway is faster, with a smaller time to peak than the ON pathway, which is typical for these ganglion cells (*S25*) and can be attributed to the slower dynamics of the synapse between photoreceptor and ON-bipolar as compared to OFF-bipolar (*S3, S26, S27*). (**E**) By repeating this process for each row of the receptive field in (A), one obtains separate ON and OFF temporal filters for each location x, denoted as $f_x^{(ON)}(t)$ and $f_x^{(OFF)}(t)$. These are used to calculate the model activation curves, see Eq. (4). The filters marked by arrows correspond to the sample location treated in panels (B)-(D).

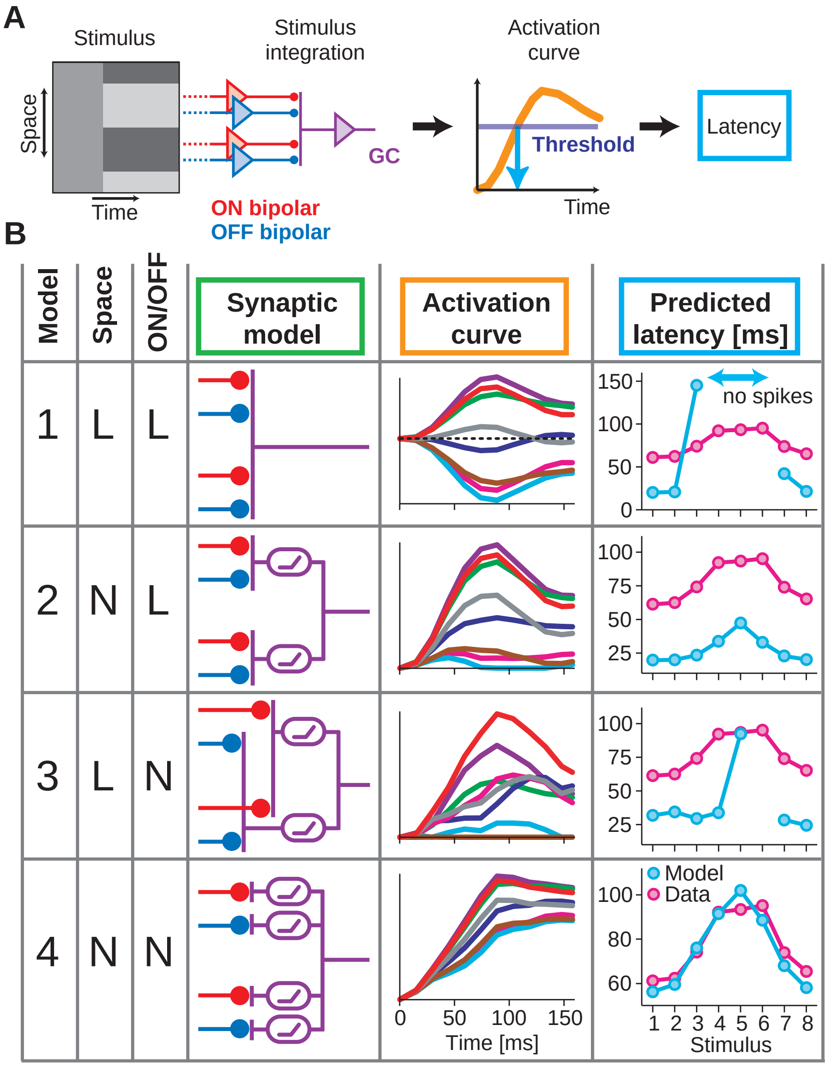

Fig. S6. Modeling the response latencies of retinal ganglion cells. (**A**) General model framework. The stimulus is uniform gray illumination followed by a square-wave grating. Processing in bipolar cell circuits according to one of the four models specified below results in an activation curve over time after stimulus onset, which can be thought of as the ganglion cells membrane potential. When the potential crosses a set threshold, the first spike is fired. The threshold is the only free parameter of the model and was optimized from a χ^2 fit to the measured latency tuning curve. (**B**) Four model versions of stimulus integration. For each model, summation over space and summation over ON and OFF pathways were either linear (L) or nonlinear (N). At each spatial location, the stimulus was integrated over time by ON and OFF bipolar cells according to the ON and OFF components of the spatio-temporal receptive field of the ganglion cell (Fig. S5). Model 1: Linear summation over ON- and OFF-bipolars and over space. Model 2: ON and OFF bipolar signals are summed linearly, then half-wave rectified before pooling over space. Model 3: Linear summation over space, separately within the ON and OFF pathways, followed by rectification and pooling. Model 4: Rectification of each bipolar cell output, followed by pooling. The fully linear Model 1 and the fully nonlinear Model 4 correspond to the two schemes compared in the main text, Fig. 3A and Fig. 3F, respectively. The activation curves for the eight applied stimuli, shown for one sample ganglion cell (same as in Fig. 3), reflect the degree of nonlinearity in the model. On the right, best fits of the latency tuning curve for each model are compared to the data for the sample cell.

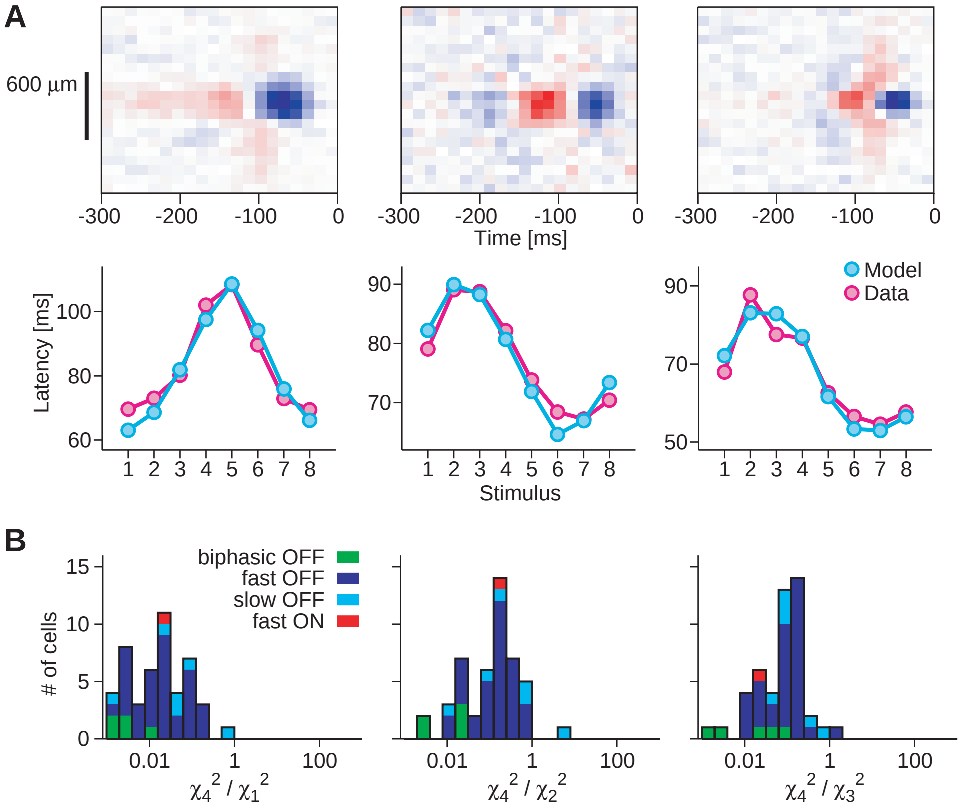

Fig. S7. Evaluation of model performance. (**A**) Fits of the latency tuning curve for three additional sample cells, a fast OFF cell (left) and two biphasic OFF cells (middle and right). All three cells have temporally biphasic receptive fields as shown in the top row. The model fits compared to the measured latency tuning curves are shown below for the model with fully nonlinear stimulus integration (Model 4 in Fig. S6). None of the other three tested models gave a satisfactory fit. (**B**) Comparison of goodness of fit for the four model versions of Fig. S6. For each model version, we investigated the performance in fitting the latency tuning curves of all cells that showed strong latency tuning. We used the 47 recorded neurons that transmitted more than 1.2 bits about the stimuli in their latency (mostly fast OFF cells, cf. Fig. S2B), and we calculated the ratios of the χ^2 values for Model 4 ($\chi_4{}^2$) and the other three models ($\chi_1{}^2$ to $\chi_3{}^2$). The histograms show that for almost all of these neurons, Model 4 clearly outperformed the other models.

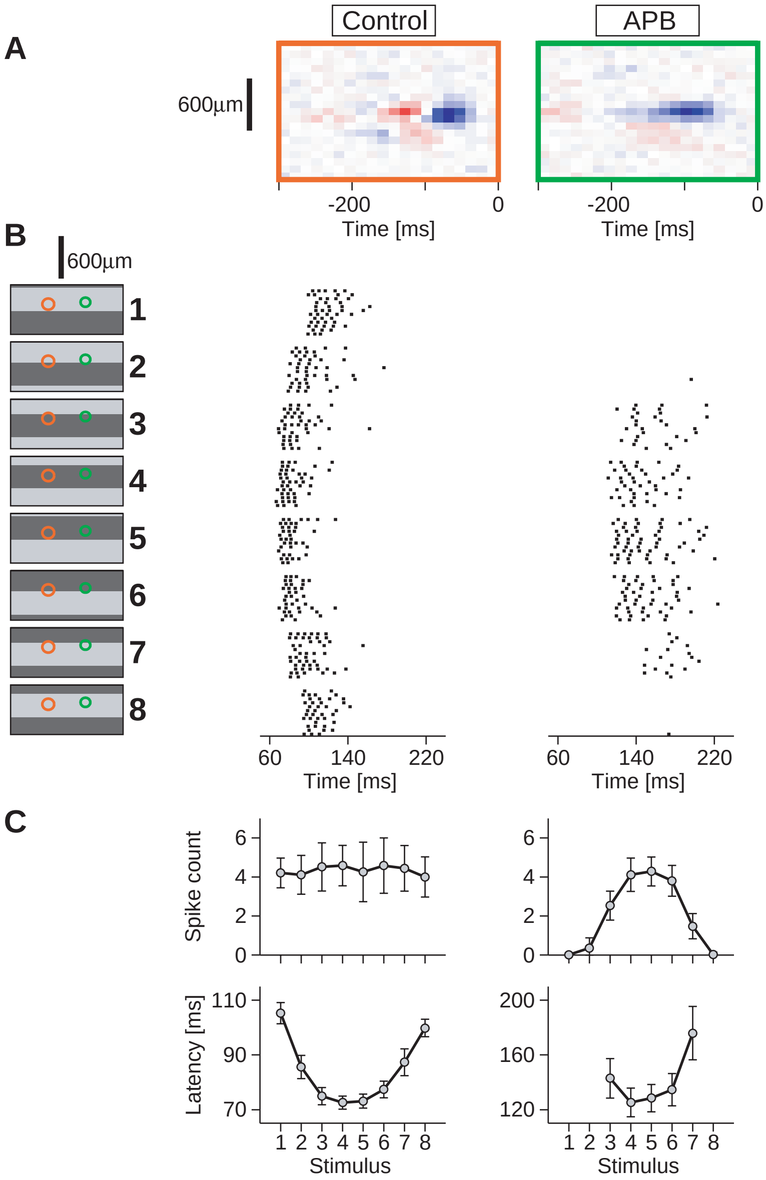

Fig. S8. Effect on the latency code of pharmacologically blocking activation of ON bipolar cells for a sample fast OFF ganglion cell. (**A**) Spatio-temporal receptive field before (Control) and after exposure of the retina to 2-amino-4-phosphono-butyrate (APB). With ON bipolar cells blocked by APB, the strong, delayed ON component of the receptive field center disappeared. (**B**) Responses to flashed gratings, displayed as in Fig. 1. After application of APB, several stimuli did not elicit spikes any more; specifically those for which a bright bar fell on the receptive field center. (**C**) Spike count and latency tuning curves. With ON inputs blocked, the cell displayed pronounced tuning in the spike count, which had been absent before.

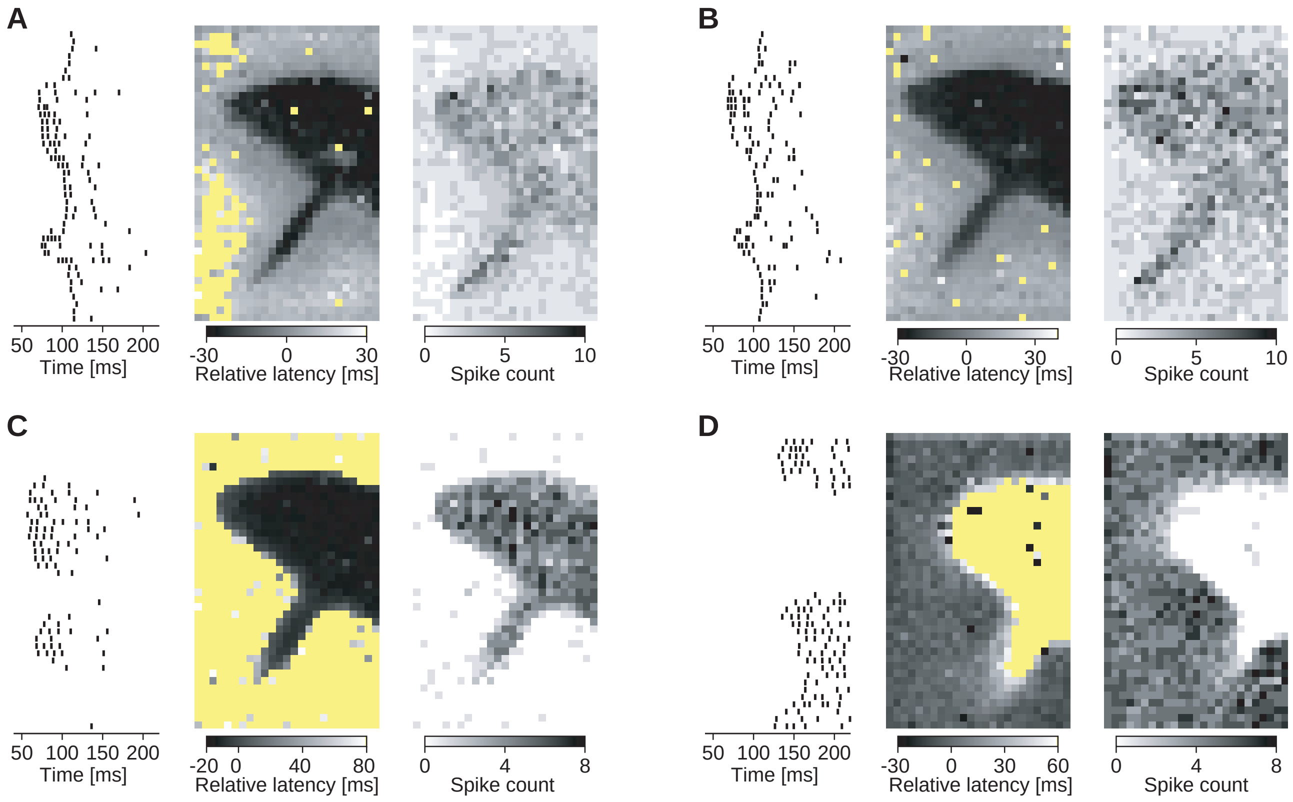

Fig. S9. Responses from four additional ganglion cells to the flashed natural photograph. Each panel shows sample spike trains and gray-scale-coded neural images from the relative latency and the spike count, presented as in Fig. 4. For the relative latency, the reference for each neuron was the average latency over all locations. (**A**, **B**) Responses from two fast OFF cells. (**C**) Responses from a slow OFF cell. (**D**) Responses from a fast ON cell. Yellow-colored pixels in the latency images correspond to locations for which the neuron did not fire. These are sparse for fast OFF cells and abundant for slow OFF and ON cells. For fast OFF cells, the neural image from the latencies always had higher quality than from the spike count. The slow OFF cell and the ON cell, on the other hand, yielded reliable neural images from the spike count, whereas most of the structure in their latency images comes from trials without any spikes.

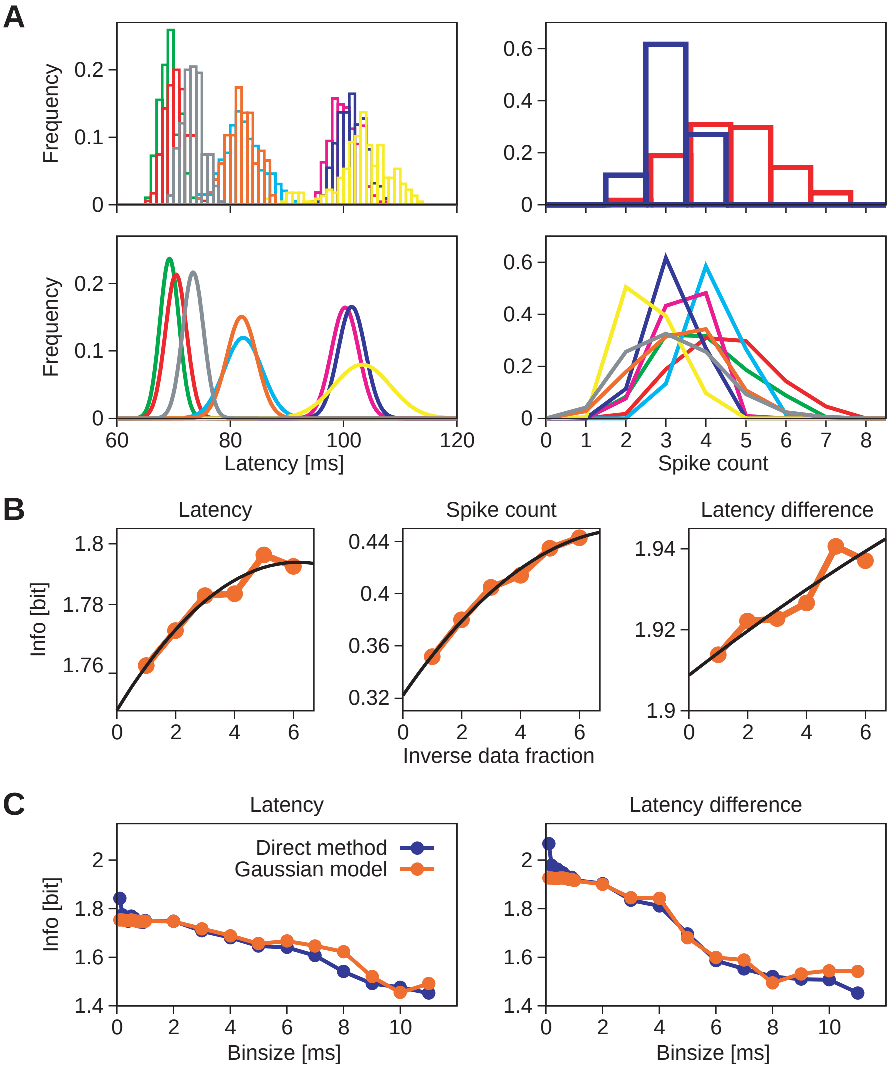

Fig. S10. Information theoretical analysis of latencies and spike counts. (**A**) Distribution of latencies (left column) and spike counts (right column) corresponding to the eight different stimuli for a sample cell. For the latencies, histograms binned at 1 ms are shown in the top panel, and the Gaussian fits (computed without binning) used for the information calculation are shown below. The data can be well fitted by Gaussian curves for each individual stimulus. Spike counts, on the other hand, assume discrete values, which were assembled into histograms without a parametric fit. Two full histograms are shown in the top panel, and all eight histograms are shown as line graphs in the bottom panel. (**B**) Bias correction for finite sampling of information values, illustrated with the information about stimulus identity contained in the latency (left) and spike count (middle) of a single cell as well as in the latency difference of two simultaneously recorded cells (right). In each case the information was calculated for the complete data set, as well as after segmenting the data into two halves, three thirds, and so on (*S11*). The average information value obtained in each segmentation is plotted against the inverse fraction of the data used, and a second-order polynomial fit was applied to extrapolate the curve to an inverse data fraction of zero, which corresponds to the case of infinite data. This corrects the values for bias induced by finite sampling. For latency information, using Gaussian fits of the latency distributions, the corrections were typically ∼1% or smaller; for spike count information, where no parametric fit was applied, corrections were ∼10% or smaller. (**C**) Comparison of information values for latencies obtained from Gaussian fits (panel (A), bottom left) and by the direct method of histogramming in fixed time bins (panel (A), top left). Values are corrected for sampling bias as in panel (B). For very small bin size, the direct method, which relies on large counts, can break down, yielding inappropriately large information values, whereas the method using a parametric fit remains stable. The Gaussian fit introduces a conservative estimate of information, because it ignores differences in the stimulus-conditional distributions beyond mean and variance that may still contribute to information transmission.