The Impact of Multivesicular Release on the Transmission of Sensory Information by Ribbon Synapses

Abbreviated title: Multivesicular release at ribbon synapses

Ben James, Pawel Piekarz, José Moya-Diaz and Leon Lagnado*

Sussex Neuroscience, School of Life Sciences,

University of Sussex, Brighton BN1 9QG, UK

* Correspondence: l.lagnado@sussex.ac.uk

Number of pages: 36

Number of figures: 8

Abstract words: 248

Introduction words: 643

Discussion words: 1483

"The authors declare no competing financial interests."

Acknowledgements: We thank the Wellcome Trust for funding (221936/Z/20/Z).

Abstract

The statistics of vesicle release determine how information is transferred at the synapse but the classical Poisson model of independent fusion does not hold at the first stages of vision and hearing. There, ribbon synapses encoding analogue signals can coordinate the release of two or more vesicles as a single event. The implications of such multivesicular release (MVR) for spike generation are not known. Here we investigate the hypothesis that MVR might provide advantages over Poisson synapses. We used leaky integrate-and-fire models incorporating the statistics of release measured experimentally from glutamatergic synapses of retinal bipolar cells and compared these with models assuming Poisson inputs constrained to operate at the same average rates. We find that MVR can increase the number of spikes generated per vesicle while reducing interspike intervals and latency to first spike. The combined effect was to increase the efficiency of information transfer (bits per spike) over a range of conditions mimicing retinal ganglion cells of different size. MVR was most advantageous in neurons with short time-constants and reliable synaptic inputs, when a lower degree of convergence was required to trigger spikes. In the special case of a single input driving a neuron with short time-constant, as occurs at the first synapse in hearing, MVR increased information transfer whenever spike generation required more than one vesicle. This study demonstrates how presynaptic integration of vesicles by MVR has the potential to compensate for less effective summation post-synaptically to increase the efficiency with which sensory information is transmitted.

Introduction

The encoding of sensory information by spikes has been studied in detail (Gollisch and Meister, 2008; Kayser et al., 2009) but less is understood about the vesicle code that drives spiking. This is especially true in the early stages of vision and hearing where analogue voltage signals are transmitted through ribbon-type synapses (Lagnado and Schmitz, 2015; Neef et al., 2007). It has often been assumed that ribbon synapses recode graded signals by modulating the mean rate of a Poisson process in which vesicles fuse independently (Choi et al., 2005; Odermatt et al., 2012; Sterling and Laughlin, 2015) but an accumulation of evidence now indicates that the release of two or more vesicles can be tightly synchronized (Glowatzki and Fuchs, 2002; Hays et al., 2021; James et al., 2019; Singer et al., 2004). In retinal bipolar cells, for instance, the ratio of the variance to the mean of the exocytic response to a repeated stimulus is not one, as expected for a Poisson process, but about three (Moya-Díaz et al., 2022). Further, the fusion of multiple vesicles can be synchronized to within the resolution of electrophysiological recordings (<100 μs) by a process that has been termed *coordinated* multivesicular release to emphasize the lack of independence (Singer et al., 2004). At central synapses the release of two or more vesicles within milliseconds is also termed multivesicular release (MVR) (Auger et al., 1998; Christie and Jahr, 2006; Holler et al., 2021; Huang et al., 2010) but it may be that in these cases it simply reflects a large increase in release probability.

Whatever the mechanisms underlying MVR, the closer together in time vesicles are released at a single synaptic connection the more effectively they will summate on a postsynaptic dendrite to impact the spiking activity of the target neuron (Li and Ascoli, 2008). This simple idea has profound implications for understanding how information is transmitted in the brain because in principle it might allow for a sensory stimulus to be represented not just by the frequency of synaptic events (a temporal code) but also by their *amplitude* (Holler et al., 2021; James et al., 2019; Lisman et al., 2007). Direct evidence for amplitude coding at a synapse has been provided by improvements in the

optical imaging of glutamate in the retina, which provide the resolution to count vesicles as they are released from individual active zones. This approach reveals that bipolar cells transmit visual information both in the rate of vesicle release and in the number of vesicles released within each event, with larger and rarer events carrying the most information per vesicle (James et al., 2019).

We now need to understand how coordinated MVR contributes to the conversion of an analogue sensory signal into spikes. An experimental investigation of this question is difficult because it would require spikes to be recorded post-synaptically while also characterizing the influence of individual active zones monitored with a resolution of single vesicles. We have therefore started with a modelling approach in which a train of synaptic events is simulated based on experimentally measured statistics of vesicle release and used as the input to a Leaky Integrate-and-Fire (LIF) model of the post-synaptic neuron (Abbott, 1999; Burkitt, 2006a). We gather these statistics from the ribbon-type synapse of retinal bipolar cells across which visual information is transmitted to post-synaptic ganglion cells and then focus on the special case of transmission of auditory information from the ribbon synapses of hair cells. A comparison with simulations in which synapses release vesicles independently demonstrates that MVR can increase the efficiency with vesicles are used to transmit sensory information over a range of conditions mimicking retinal ganglion cells of different sizes. In the special case of auditory hair cells driving afferent fibers through a single synapse, MVR increased the efficiency of information transfer whenever spike generation required depolarization greater than that caused by a single vesicle.

Methods

Modelling Overview

Modelling the effects of MVR on spike generation involved the following basic steps:

1. Statistics of vesicle release in retinal bipolar cells responding to visual stimuli. Using iGluSnFR expressed in retinal bipolar cells of larval zebrafish, we measured the amplitude of synaptic events and their timing relative to a sinusoidal full-field stimulus of varying contrast (Fig. 1A). At each contrast we calculated i) the average rate of vesicle release over a 30 s period *R* (Fig. 1B; n = 60 synapses); ii) the probability of events composed of different numbers of quanta, P_{Qe} (Fig. 1C; n = 55 synapses), and iii) the "temporal jitter" of events of each amplitude, quantified as the standard deviation of event times relative to any particular phase of the sinusoidal stimulus (*TJ(Q)*, Fig. 1D-F; n = 60 synapses). Each statistic was calculated from at least 20 zebrafish.

2. Simulation of synapse output. Based on step 1, two types of simulated synaptic output were constructed: one in which stimulus contrast was represented as a "pure" rate-code in which all vesicles were released independently and a second "hybrid code" in which amplitude also varied (Fig. 2A). In both cases, the overall average rate of vesicle release *R* was fixed to values measured experimentally at a given contrast. The number of vesicles in an event (for the MVR case) or the number of released uniquantal events (for the rate case) was sampled from the distribution P_Q. For the MVR case, each event was then given a time by sampling from the Gaussian distribution of event times with mean and variance measured experimentally i.e by TJ(q) shown in Fig. 1F. For the rate case, a corresponding number of event times were sampled from the distribution of uniquantal event times.

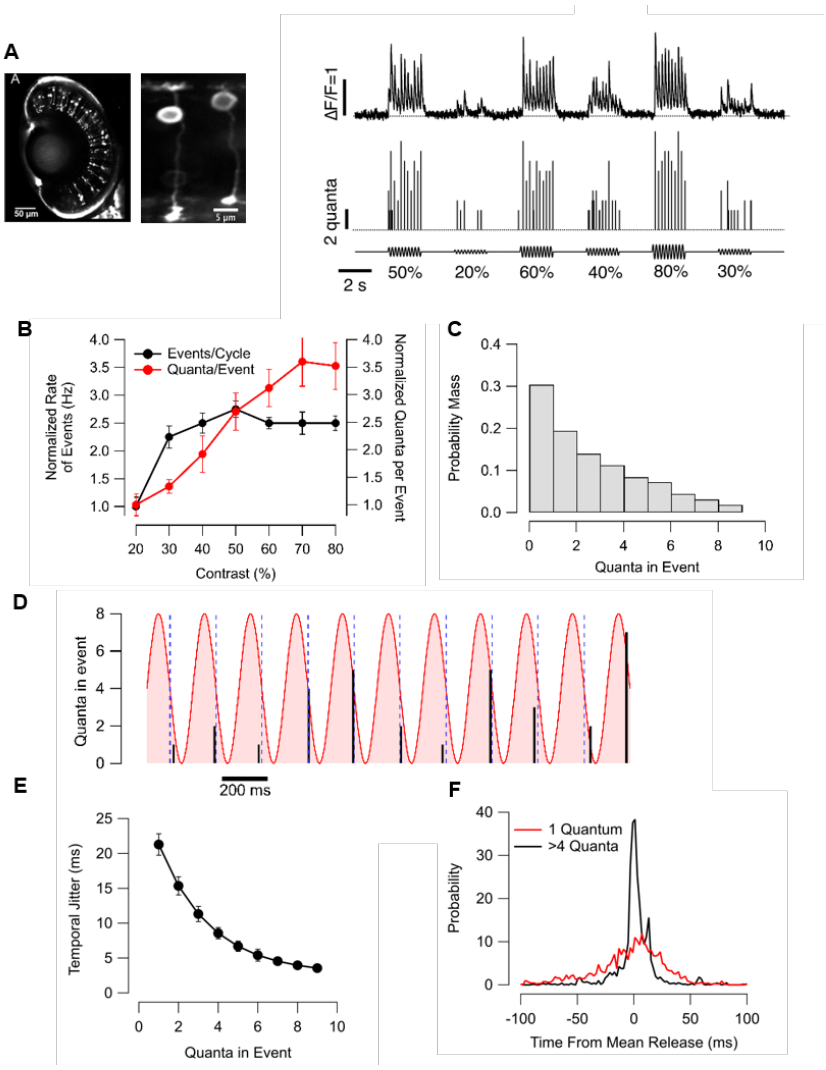

Figure 1. Multivesicular Release in the Retina of Zebrafish. A. Left: Multiphoton image of BCs in the retina of a larval zebrafish. They are visualized by the iGluSnFR reporter in the surface membrane. Middle: Two BCs at higher magnification. The mosaic expression of the reporter allows for distinguishing individual BCs and distinguishing glutamate release at different synapses. Right Top: Recorded fluorescence time series from an individual active zone. Bottom: Estimated quantal content of the time series above. Note the heterogeneity in the response amplitude, corresponding to variable quanta released per event. **B.** Normalized change in the rate of events (black) and average event amplitude in quanta (red) (n = 60 synapses). **C.** Distribution of event amplitudes during a 30 s application of a 5 Hz stimulus at 80% contrast (n = 55 synapses), truncated to only include events containing fewer than 11 quanta. **D.** A quantal time series (black lines) in response to a 5 Hz, 80% contrast full-field stimulus (red), with blue dashed lines present only for ease of visibility. Note that events composed of four or more vesicles occur in a temporally small window during each cycle, while lower quantal events are more temporally dispersed. **E.** The temporal jitter as a function of quanta in event for a 100% contrast stimulus (n = 60 synapses). **F.** Mean-shifted distribution of events containing 5 or more quanta (black) and of events containing one quantum (red) in response to a 30 s, 5 Hz, 100% contrast sinusoidal stimuli. Data from E.

The output of a synapse employing coordinated MVR was generated by randomly selecting a total of N events with amplitudes chosen from the distribution of quanta in an event P(Qe) (Fig. 2B). Each event was then given a time by sampling from the Gaussian distribution of event times with mean and variance measured experimentally i.e by TJ(Qe) shown in Fig. 1F. Each synapse was constrained to generate only one or zero events per cycle of the stimulus, with the probability of no event given by $p_Q(f, c)$. The output of each synapse was therefore described as a vector E of event times and a vector Qe of event amplitudes, both of length N.

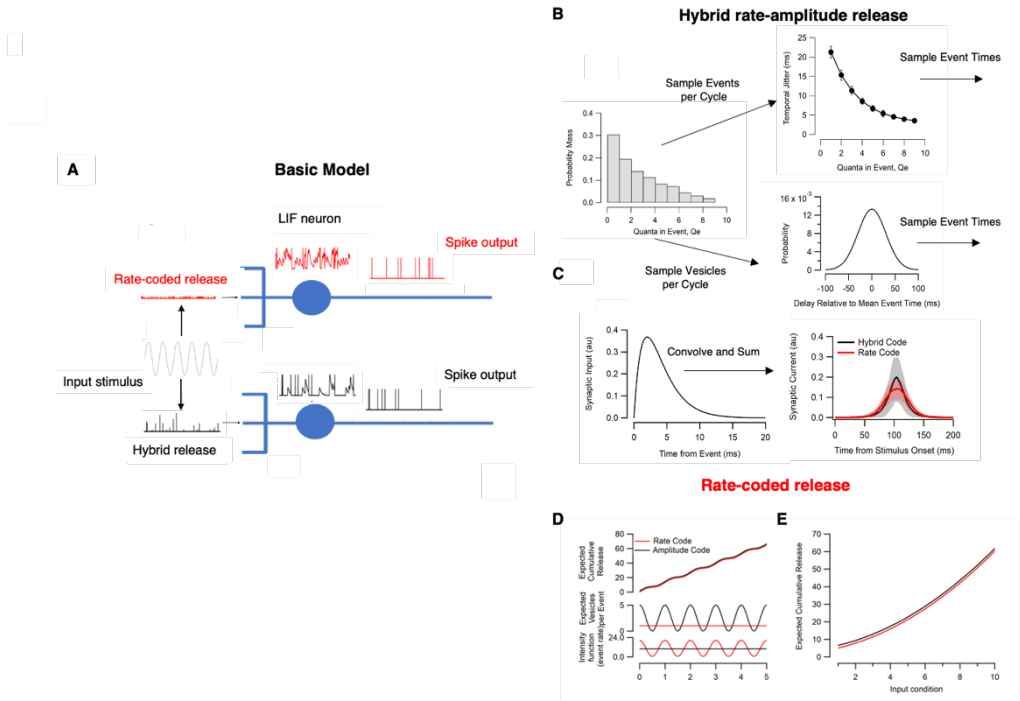

Figure 2. Overview of modelling. **A**. A leaky-integrate-and-fire neuron received either rate (top) or hybrid (bottom) inputs. **B.** Illustration of release simulation. The number of vesicles in an event (for the MVR case) or the number of released uniquantal events (for the rate case) is sampled from the distribution P_Q. For the MVR case, each event is then given a time by sampling from a Gaussian Distribution with variances set by the mean temporal precision of events, given by TJ(q). For the rate case, a corresponding number of event times are sampled from the distribution of uniquantal event times (bottom middle). **C.** The sampled event time and amplitude tuples are then convolved with a synaptic current filter. The resulting current is summed over the number of inputs to the cell. Plotted on the right are the average synaptic current input to a cell utilizing a rate code (red) and hybrid code (black). Note the increased temporal variability in the rate code (SD = 360 ms for the hybrid case vs. 685 ms for the rate case). **D**. Demonstration of the Poisson release model. Bottom: intensity function driving the rate of vesicular events. Middle: expected number of vesicles released in an event as a function of time. Top: Expected cumulative vesicles released as a function of time for rate and amplitude codes, offset for visualization. These functions have been equalized by design. **E.** Expected total vesicles released over a one second window as a function of stimulus 'contrast'.

175 The rate-code comparison was constructed by replacing each event consisting of
176 more than one quantum with an equivalent number of uni-quantal events, each event time
177 now sampled from the distribution of times of uniquantal events (Fig. 1F). The vectors
178 describing a pure rate synapse were therefore of variable lengths corresponding to the
179 total number of vesicles sampled within the 30 s sample period. The resulting vectors Qe
180 and E of event quanta and event times were then passed into a conductance-based
181 model LIF model (Fig. 1) and the spike output analysed.

182

183 ***3. Simulation of synapse output using marked Poisson Processes.*** For purely rate-
184 coded synapses, we used a Nonhomogeneous Poisson Process (NHPP) as the driving
185 force for vesicle release, in which the mean rate λ(t) varied sinusoidally in time. By
186 definition, simple Poisson Processes describe a situation in which all events occur
187 independently and with exponentially distributed interevent times and cannot, therefore,
188 be used to model MVR when two or more vesicles are released within a single synaptic
189 event. To construct an amplitude-modulated Poisson Process we turned to a modification
190 known as Poisson Splitting or Location-dependent thinning (Resnick, 1992), where events
191 can be assigned a type based upon their location in time. In this case events occur with a
192 constant rate but the number of vesicles in each event is chosen based upon when the
193 event occurred. We could therefore drive the LIF model neuron in two distinct regimes: a
194 pure rate case, where all information is encoded by modulating the mean rate of release
195 of individual vesicles and a hybrid case, where information is encoded by modulating both
196 the rate and quantal content of synaptic events.

197 A key aspect of our investigation was to construct the rate and hybrid models in such
198 a way that each released the same average number of vesicles, with differences found
199 only in the variabilities. In other words, vesicles were released individually in the "rate
200 code", but in packets of variable size in the "hybrid code". To simulate pure rate-coded
201 synaptic inputs case, we used a sinusoidal intensity function. The response of each
202 synapse to varying contrasts was simulated by increasing the amplitude of the sinusoid

203 between a total release of between 5 vesicles s^{-1} at input condition 1 (which we term 10%
204 contrast below) and 65 vesicles s^{-1} at input condition 10 (100% contrast; Fig. 2D and E).
205 for the rate case. Responses from multiple cells were assumed to come from the same
206 distribution, and as such we simulated the responses of multiple BCs onto a single RGC
207 by summing the response to all the individual synaptic inputs. To simulate hybrid coding
208 the probability of release was also modulated as a sinusoid with an amplitude generating
209 an average rate of vesicle corresponding to that measured experimentally at a given
210 stimulus contrast (Fig. 1A).
211
212 ***4. A conductance-based Leaky Integrate-and-Fire (LIF) model of the post-synaptic***
213 ***neuron.*** The vectors describing amplitudes and times of synaptic events were convolved
214 with a synaptic filter to estimate the time-course of the synaptic current S_{in} (Fig. 2C) which
215 was then passed through a Leaky Integrate-and-Fire model to calculate the change in
216 voltage and spike output. LIF models provide a computationally compact method for
217 exploring basic physiological parameters affecting spiking, including the number of
218 synaptic inputs, cell time-constant, threshold relative to resting potential and ion channel
219 conductances (Abbott, 1999; Burkitt, 2006a, b; Gerstner et al., 1997). Voltage is
220 governed by the differential equation:

221 $$I(t) = g_l\big(v_{rest} - v(t)\big) + \big(v_e - v(t)\big)g_e S_{in}(t) \quad \textbf{(1)}$$

222 Where g_l is leak conductance, g_e is excitatory conductance, v_e is the excitatory reversal
223 potential, and v_{rest} is the resting membrane potential. The synaptic input, S_{in}, is defined
224 by the convolution of an alpha current impulse function with the input s(t),

225 $$S_i n(t) = \alpha * s(t) = \sum_{i=1}^{n} \alpha(t - t_i) \quad \textbf{(2)}$$

226 Where s(t) is the sequence of Dirac delta function describing the timing of each event in a
227 realization of the Process, and

228 $$\alpha(t) = \left(\frac{t}{\tau_s}\right) - exp\left(\frac{-t}{\tau_s}\right) \quad \textbf{(3)}$$

229 describes the time course of synaptic current. Here, we used a value of τ_s = 2 ms
230 (Destexhe and Rudolph, 2004) which yields a total conductance of 2 nS per quanta. Once
231 the cell hits a threshold θ, it fires a spike and the voltage is immediately reset to the reset
232 potential v_{reset} , and remains at that value in a refractory period for τ_{ref} = 5 ms. There is
233 no explicit membrane time-constant ($\tau_m = RC$) in a conductance model but the time
234 course of the leak varies as the reciprocal of the leak conductance $\tau_l = g_l^{-1}$. Due to its
235 similarity with the current-based model, we will describe the model in terms of τ_l. The
236 ratio between the membrane capacitance and the leak channel conductance is the
237 membrane time-constant ($\tau_m = C\, g_l^{-1}$).
238

Parameter (symbol)	Simulated Value/s
Vesicles Required to Spike, k	10, 25, 50,100 vesicles
Leak Constant, τ_l	1, 5, 10, 20, 50 ms
Resting Potential, v_{rest}	-70, -65, -50, -55 mV
Excitatory Reversal Potential, v_e	0 mV
Spiking Threshold, θ	-50, -45, -40, -35 mV
Number of Inputs, N	50, 100, 250, 500
Refractory period, τ_{ref}	5 ms
Synaptic Time Constant, τ_s	2 ms
Stimulus Frequency, f	5, 10, 15 Hz
Stimulus Contrast, c	20, 60, 100%

239
240 **Table 1: Simulation Parameters.** Parameters used for simulations, with ranges based upon experimentally
241 recorded values (Destexhe and Rudolph, 2004; O'Brien et al., 2002; Pillow et al., 2005)
242
243 To compare how the MVR might shape the spike output of a postsynaptic cell across
244 a range of physiological values, we systematically varied the parameters shown in Table
245 1. Note that here, rather than representing the set of the excitatory conductances, g_e, we

computed the approximate leak conductance required for the cell to spike in response to k vesicles in a perfect integrator – with corresponding g_e values ranging in the 1-100 nS range. In the retina, bipolar cells transmit to RGCs of widely varying size and electrical properties. In mice, for instance, input resistances of RGCs can vary from dozens of MΩ to a GΩ, with values of the membrane time-constant ranging from less than 5 ms to over 50 ms (O'Brien et al., 2002). At another extreme, primary auditory afferents receive inputs from single active zones delivering the output from hair cells and have input resistances of GΩ leading to the suggestion that a single glutamatergic vesicle is sufficient to trigger a spike (Fuchs and Glowatzki, 2015; Glowatzki and Fuchs, 2002). Thus, the number of vesicles sufficient to generate a spike in a cell postsynaptic to a ribbon synapse (k) varies from one to hundreds.

5. *Analyzing the spike output of the model neuron.* We concentrated on three aspects of the spiking response of the model neuron: the spike count generated by each cycle of the 5 Hz stimulus, the temporal accuracy of spikes and the average rate of information transfer about the modulation in stimulus strength (i.e contrast).

Temporal accuracy was quantified as “temporal jitter” (TJ, in ms), which was in turn calculated from the Vector Strength (VS), which takes values between zero and one, with zero being complete independence from the stimulus and one being perfect phase-locking (Baden et al., 2011; Goldberg and Brown, 1969). VS was computed for each quantal event type in response to a 30 s stimulus of varying contrast delivered at 5 Hz, with the equation:

$$VS_Q = \frac{1}{N_Q}\sqrt{\left(\sum_{i=1}^{N_Q} cos\left(\frac{2\pi t_{Q_i}}{T}\right)\right)^2 + \left(\sum_{i=1}^{N_Q} sin\left(\frac{2\pi t_{Q_i}}{T}\right)\right)^2} \quad \textbf{(4)}$$

where t_{Q_i} is the phase of the ith event of quantal content Q, and T is the stimulus period. VS can then be easily converted to Temporal Jitter (TJ) by:

$$TJ_Q = \frac{\sqrt{2(1-VS_Q)}}{2\pi f} \quad \textbf{(5)}$$

We calculated the mutual information (Shannon, 1948) between the stimulus S and either the spike times or the total number of spikes occurring over a period of 200 ms (N, corresponding to one cycle of the 5 Hz stimulus, 2 cycles of the 10 Hz stimulus or 3 cycles of the 15 Hz stimulus). Mutual information for spike counts was computed empirically from the joint and marginal distributions $p(s,n)$, $p(n)$, and $p(s)$ as:

$$I(S;N) = \sum_{s\in S, n\in N} p(s,n) log\left(\frac{p(s,n)}{p(s)p(n)}\right) \quad \textbf{(6)}$$

using the direct approach (Victor, 2006). Each stimulus s is has one of three contrasts and one of three frequencies, leading to a total of 9 different stimuli and a maximum of ≈ 3.17 bits of information. The total mutual information was calculated by discretising the response in a 200 ms time window into 5 ms bins, yielding a total of 40 response bins and 2^{40} possible responses (the vast majority of which are never observed). To reduce bias in the estimates, each simulation was run a total of 1000 times, with typical simulations generating under 80 unique responses. Keeping the ratio of the number of samples to the number of distinct responses large ensures any bias is small (Panzeri et al., 2007). The spike count information was computed similarly, using the distribution of spike counts **N**. Finally, the spike-time information was computed as the difference between the total information and the spike count information, $I(S;R)_{Time} = I(S;R)_{Total} - I(S;R)_{Count}$

The signal-to-noise ratio associated with a change in the rate of a Poisson process

Imagine the mean rate of a Poisson process, R, changes by a factor a. The signal, S, generated by comparing two observation times Δt will be the change in the mean number of events counted in each period (aR Δt - R Δt), and the variance in that signal will be the sum of the number of events counted in each period (aR Δt + R Δt). Defining the SNR in the same way as the discriminability (d') used in signal detection theory, we have

297 $$SNR = \frac{S}{\sqrt{variance}} \quad (\mathbf{7})$$

298 yields

299 $$SNR = \frac{(aR\Delta t - R\Delta t)}{\sqrt{(aR\Delta t + R\Delta t)}} \quad (\mathbf{8})$$

300 which can be rearranged to obtain the time period Δt required to obtain a given SNR, as

301 $$\Delta t = \frac{SNR^2}{R}\frac{(a+1)}{(a-1)^2} \quad (\mathbf{9})$$

302

303 **Multiphoton imaging of glutamate release**

304 Methods for *in vivo* imaging of glutamate release in the retina of zebrafish were as
305 described previously (James et al., 2019). Transgenic larval zebrafish expressing the
306 iGluSnFR reporter under the *ribeye a* promoter were imaged at 7-9 days post-fertilization
307 after embedding in a 3% low melting-point agarose in E2 on a glass cover-slip. To
308 prevent eye movements, the fish was injected with 1 nL of alpha-bungarotoxin behind the
309 eye. Imaging was carried out using a multiphoton microscope (Scientifica) equipped with
310 a mode-locked titanium–sapphire laser (Chameleon, Coherent) tuned to 915 nm and an
311 Olympus XLUMPlanFI ×20 water immersion objective (NA of 0.95). Fluorescence
312 emission was also collected using a sub-stage oil condenser (Olympus; NA of 1.4), which
313 was essential to achieve a SNR adequate for reliable detection of single vesicles. A total
314 of 42 zebrafish were used in this study and all measuremnts were made in the afternoon.

315 Light emission was filtered through GFP filters (HQe525/50, Chroma Technology)
316 before detection using GaAsP photomultipliers (H7422P-40, Hamamatsu). Photocurrents
317 were passed to a current-to-voltage converter, after which signals from the objective and
318 condenser were summed and digitized through ScanImage (v.3.6; Vidrio Technologies).
319 For this work, all time-series were acquired as 128 pixel line scans imaged at 1 kHz and
320 analysed in custom software written within IgorPro (Wavemetrics). Stimuli were generated
321 by an amber LED (peak wavelength of 590 nm) filtered through 590/10 bandpass filter
322 (Thorlabs) delivered through a light-guide placed near the fish's eye, with the mean

323 stimulus intensity of ~320 nW mm^{-2}. Detailed evidence that this approach can detect the
324 release of single vesicles from individual active zones has been presented previously
325 (James et al., 2019).
326

327 ## Results

328

329 **Synaptic coding of a sensory stimulus: gathering release statistics for an**
330 **empirically-driven model**

331 Our aim was to investigate how the statistical properties of vesicle release affected the
332 synaptic transfer of sensory information, comparing synapses employing coordinated
333 MVR with the conventional view of synapses operating as simple Poisson machines. The
334 starting point was to gather experimental measurements of release statistics from retinal
335 bipolar cells expressing the glutamate reporter iGluSnFR, and examples of such records
336 are shown in Fig. 1A. Both the rate and amplitude of glutamate transients varied widely
337 during sinusoidal modulations in light intensity (James et al., 2019).

338 We measured how changes in contrast altered three aspects of the vesicle code
339 highlighted in Fig 1: event rate R(Q), distribution of event amplitudes P(Qe), and variability
340 of event timing TJ(Q). R saturated at lower contrasts than Qe, such that contrasts greater
341 than ~50% were often encoded entirely by the number of vesicles released within an
342 event (Fig. 1B and C). A potential role of MVR in determining the temporal accuracy of
343 spiking in the post-synaptic neuron is immediately suggested by the observation that
344 multiquantal events were more tightly phase-locked to the stimulus than vesicles released
345 individually (Fig. 1D-F). Uniquantal events had a temporal jitter of 21.2 ± 1.5 ms, while
346 events composed of 9 quanta had a temporal jitter of just 2.7 ± 0.6 ms, which is very
347 similar to the precision of spikes observed in RGCs of the salamander retina responding
348 to a high-contrast stimulus (Berry et al., 1997a; Uzzell and Chichilnisky, 2004). The

349 relation between TJ and Qe is shown in Fig. 1E and a comparison of the distribution of
350 times for uniquantal and multiquantal events is shown in Fig. 1F.

351 Based on these measurements we constructed two types of synaptic input to the
352 model neuron: purely rate-code inputs in which all vesicles were released independently
353 (red in Fig. 2A) and hybrid inputs in which both rate and amplitude of events varied (black;
354 see Methods). To compare how these two coding strategies performed with equivalent
355 resources the average rates of vesicle release were fixed for both types of input at the
356 values measured experimentally at a given contrast and frequency (Fig. 1B). Measuring
357 the output of individual synapses to stimuli of different contrasts then allowed us to make
358 measures of the mutual information between the stimulus set and spikes generated in
359 model neurons.

360

361 **Synaptic coding strategy alters interspike intervals**

362 Examples of simulations comparing the effects of rate-coded and hybrid-coded synaptic
363 inputs are shown in Fig. 3 for two model neurons chosen to highlight the effect of cell size.
364 Bipolar cells transmit the visual signal to RGCs that vary in the dimensions of their
365 dendritic trees (determining the size of receptive fields) and membrane time-constants
366 (determining temporal integration of synaptic inputs). In the fovea of primates, for
367 instance, midget ganglion cells have small receptive fields and are electrically compact,
368 receiving 25-50 synaptic inputs from between one and three bipolar cells (Calkins et al.,
369 1994; Kolb and Marshak, 2003). But in the periphery of the retina, RGCs have larger
370 dendritic fields and receive inputs from hundreds of bipolar cells, with up to thousands of
371 synapses (Jacoby et al., 2000). We therefore begin by comparing spike outputs in a
372 "small" synthetic RGC (τ_l = 50 ms with 50 synaptic inputs) and a "big" RGC (τ_l = 5 ms
373 with 500 inputs) to stimuli that are spatially uniform, acting on all excitatory inputs equally.
374 These two example neurons have different contrast-response relation (Fig. 3A),
375 qualitatively similar to variations observed physiologically.

376

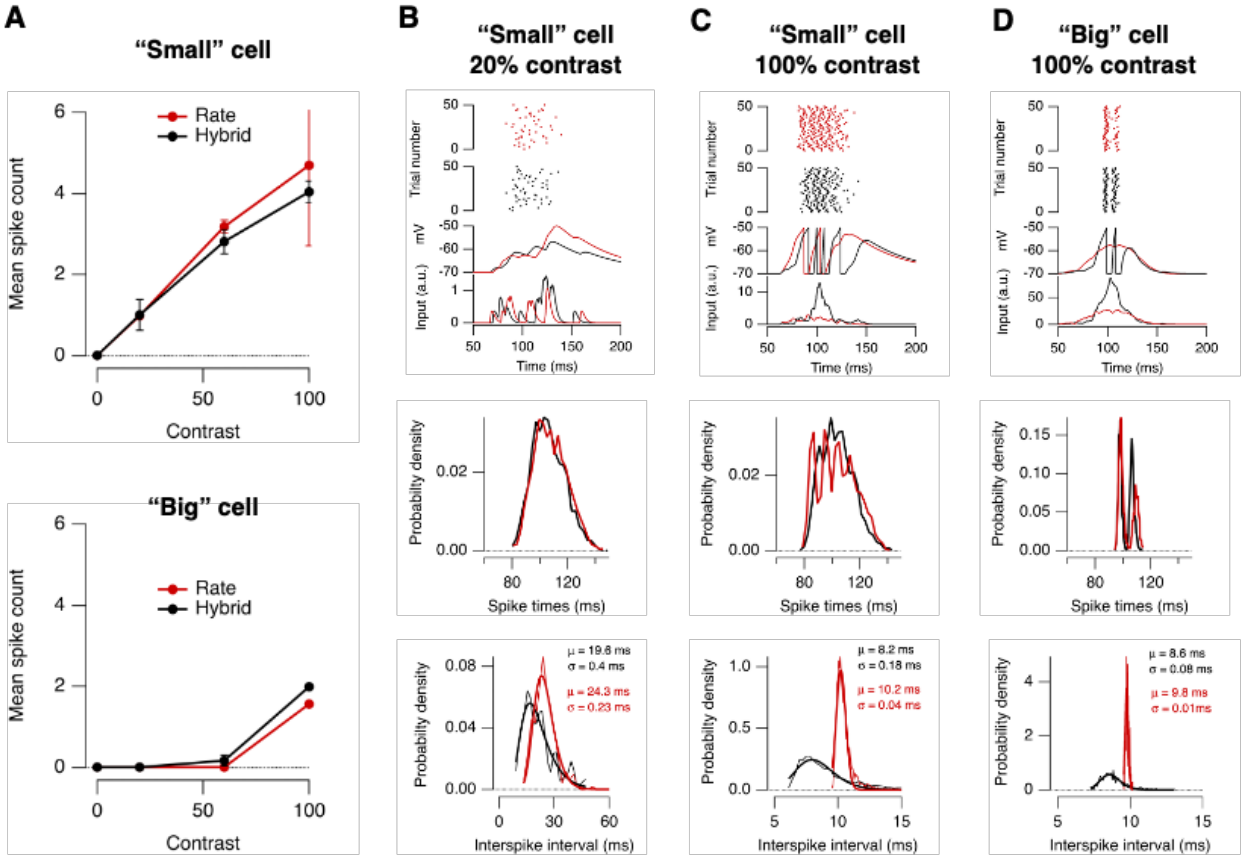

Figure 3. Comparison of vesicle codes on spike generation in a "small" cell and a "large" cell. **A**. Mean spike count as a function of contrast for a "small" cell (top) and "big" cell (bottom). The "small" cell was simulated with: inputs = 50, k = 10, τ_l = 50 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV. The "big" cell was simulated with: inputs = 500, k = 100, τ_l = 5 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV. Stimuus at 5 Hz. Red shows responses to synaptic inputs employing a pure rate code and black to inputs employing a hybrid rate-amplitude code. **B.** A small cell responding to 20% contrast. Top panel: raster plots of spike times (top), example voltage traces (middle), and example synaptic input (bottom), for hybrid and rate codes. One cycle of the sinusoidal stimulus is shown. Middle panel: distribution of spike times. Bottom panel: distribution of interspike intervals (ISI, thin trace) could be described as a lognormal function (bold trace) with values of μ and σ shown (equation 10). **C.** Same as B, but for 100% contrast. Note the lower mean but broader distribution of ISI's in the cell receiving hybrid inputs. **D.** Same as C, but for the big cell. Note that despite the much narrower distributions of spike times (middle) the ISI's were more broadly distributed in the cell receiving hybrid inputs (bottom).

Using a stimulus of low contrast (20%), the big cell barely responded but the small cell generated an average of one spike per cycle of the 5 Hz stimulus (Fig. 3A). The

397 distributions of spike times were not significantly affected by the coding strategy of the
398 synaptic inputs (Fig. 3B, middle) but when two or more spikes were generated within a
399 cycle the distribution of interspike intervals (ISI) was shifted to significantly lower values
400 when input was provided by hybrid synapses employing MVR (Fig. 3B, bottom; $p < 0.02$;
401 Kolmogorov–Smirnov test). This distribution, Y(ISI), could be described as a lognormal
402 function of the form
403

404 $$\mathrm{Y(ISI)}=\mathrm{A}\ exp\left\{-\frac{(\ln(ISI)-\mu)^2}{2\sigma^2}\right\} \quad \textbf{(10)}$$

405

406 where μ and σ are the mean and standard deviation of a normal distribution in ln(ISI). A
407 lognormal distribution of ISI's is commonly found in RGCs (Levine, 1991) and was a
408 consistent feature of the spike output across a range of conditions (Fig. 3B-D). In the
409 small cell responding to 20% contrast, the ISI was reduced from 24.3 ± 0.2 ms with rate-
410 coded inputs to 19.6 ± 0.4 ms with hybrid inputs. Stimulating the small cell with 100%
411 contrast increased the average spike count by a factor of more than 4 (Fig. 3A) but the
412 same qualitative difference was observed (Fig. 3C, bottom): the mean ISI was significantly
413 smaller with hybrid inputs (8.2 ± 0.18 ms) compared to rate-coded inputs (10.2 ± 0.04 ms;
414 $p < 0.0001$, KS test). The big cell with ten times as many inputs and one-tenth the time-
415 constant generated about half as many spikes at 100% contrast, but again the mean ISI
416 was shorter with hybrid inputs (Fig.3D, bottom; $p < 10^{-20}$; KS test).
417 The precise timing of spikes in an RGC can code information about the stimulus (Ishii
418 and Hosoya, 2020; Meister and Berry, 1999) that neurons post-synaptic to RGCs can
419 decode (Kaplan et al., 1987; Usrey and Reid, 1999). In mammals, for instance,
420 recordings from RGCs and their target neurons in the lateral geniculate nucleus (LGN)
421 show that the efficiency with which an RGC spike drives an LGN spike depends on the
422 ISI: ISIs shorter than about 10 ms have the greatest efficiency but this decreases
423 progressively with ISI up to ~30 ms (Usrey et al., 1998). Such ISI-based filtering of retinal
424 spikes can shape receptive fields and may serve as a mechanism by which the

425 information carried in ISI's is decoded (Ishii and Hosoya, 2020; Rathbun et al., 2010). The
426 simulations in Fig. 3 therefore show that a hybrid vesicle code driving excitation of RGCs
427 has the potential to increase the visual information transmitted to post-synaptic targets by
428 shortening spike intervals.
429
430 **Coordinated MVR increases spike count over a range of conditions**
431 The spike response of RGCs will depend not only on the numbers of synaptic inputs (N)
432 that converge on the dendritic tree but also on the strength of these inputs. In our
433 modelling framework synaptic strength is directly related to k, the minimum number of
434 vesicles that would need to be summed instantaneously to depolarise the neuron to
435 threshold. Although we systematically varied both k and N (Table 1) we found the ratio
436 k/N to be a useful summary metric that reduced dimensionality of the data while providing
437 insight into the effects of both synapse strength and convergence. For instance,
438 depolarization of the target neuron to threshold might occur by integrating signals from
439 just a few synapses with a high probability of release (i.e reliable synapses, high k/N) or
440 more synapses with lower average response (i.e synapses of lower reliability with low
441 k/N). Medium-sized RGCs in guinea-pigs receive about 1000-2000 inputs, and a brief
442 burst of excitation comprising 3-65 vesicles triggers a burst of 1-6 spikes (Freed, 2005).
443 This example therefore provides an upper limit of k ~10 while k/N is ~0.01. In contrast, a
444 midget RGC might receive just 10 inputs. If the same number of vesicles are required to
445 trigger a spike, k/N = 1, and even if just one vesicle is sufficient, k/N =0.1.
446 Examples of the interaction between k/N and the time-constant of the target neuron
447 are shown in Fig. 4. Increasing τ_l always increased the spike count over a single stimulus
448 cycle, and this is shown in Fig. 4A (top) for an example neuron with N = 500 inputs
449 requiring the summation of K = 100 vesicles to trigger a spike (k/N = 0.2). The absolute
450 difference in spike count between hybrid inputs and rate-coded inputs are shown below,
451 demonstrating that for this value of k/N, the hybrid code produces more spikes than the
452 rate code when τ_l is less than ~15 ms. The interaction between τ_l and k/N over a range of

453 values is shown for hybrid coded inputs in Fig. 4B and for rate-coded inputs in Fig. 4C. A
454 clear pattern emerges: for either coding strategy, the lower the average number of
455 vesicles released per synapse to trigger a spike the larger the total spike count for a cell of
456 given time-constant. But the size of these changes were not the same and the difference
457 between the counts generated by hybrid- and rate-coded inputs is shown in Fig. 4D,
458 colour-coded such that increases are shown as depth of green and decreases as depth of
459 red. For a given τ_l, rate-coded inputs generate more spikes at lower k/N and hybrid inputs
460 were more efficient at higher k/N. In other words, the rate code generated more spikes
461 per vesicle than the hybrid code only when synapses were relatively unreliable. As the
462 average number of vesicles released per synapse increased, a regime was entered in
463 which the hybrid code generated spikes with greater efficiency, after which the coding
464 strategy did not affect the spike count. The basic insight provided by these results is that
465 where spike generation required more reliable synapses, it often became more effective if
466 summation began *presynaptically* by the process of coordinated MVR.
467

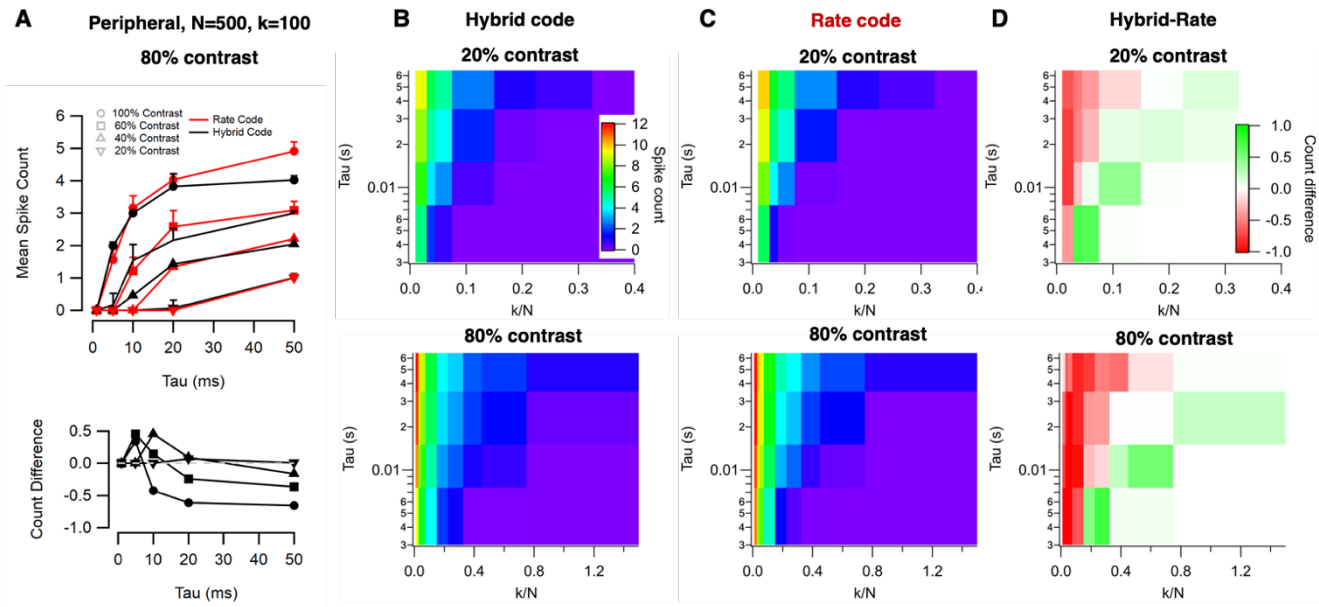

Figure 4. Hybrid coding increased spike count over a range of physiologically-relevant conditions

A. Top: Mean spike count as a function of leak time-constant (τ_l) for a cell receiving N=500 inputs and requiring k=100 vesicles to spike in response to a 5 Hz stimulus of varying contrast (other parameters as in the "large peripheral" cell in Fig. 3). This count is over a 200 ms time window so can be multiplied by 5 to obtain spike rates. Bottom: Spike Count Difference (Hybrid – Rate) from the data in top panel. For this ratio of k:N, the hybrid code produces more spikes than the rate code when τ_l < 15 ms. **B.** Top: heat map showing the spike count generated by hybrid inputs for cells of varying k/N and τ_l for a stimulus of 20% contrast. Counts were averaged over values of k and N shown in Table 1. Bottom: as top but for 80% contrast. **C.** As B but for rate-coded inputs. Note again that the average rate of vesicle release was the same as in B for a given contrast. **D.** The difference between the counts generated by hybrid- and rate-coded inputs in B and C. Note that, for a given τ_l, rate-coded inputs generate more spikes at lower k/N and hybrid inputs were more efficient at higher k/N.

How does this comparison of the hybrid and rate codes relate to the physiology of the retina? Compared to other visual circuits, such as those in the cortex, the retinal output is strikingly reproducible over repeated trials of a stimulus and contrasts of 20-40% can trigger spikes in RGCs of a variety of species with almost perfect reliability (i.e close to 100% of trials; (Berry et al., 1997a; Gollisch and Meister, 2008)). Unsurprisingly, the reliablity of individual bipolar cell synapses was lower. At 20% contrast a 5 Hz stimulus triggered a response in 31 ± 3% of stimulus cycles (N = 48 synapses), rising to 64 ± 4% of cycles at 60% contrast (N = 56 synapses). Experimental measurements of the average number of vesicles released per stimulus at these contrasts (where N is a single synapse)

are shown in Fig. 1A-B: a stimulus of 20% contrast released an average of 2.32 ± 0.09 (N = 48 synapses) vesicles per synapse, while 60% contrast released 2.84 ± 0.14. Comparing these experimental measurements with the results of modelling indicates that synapses of bipolar cells are operating over a range of reliabilities where coordinated MVR provides an advantage over a pure rate-code by increasing the number of spikes generated per released vesicle.

Coordinated MVR reduces spike latency over a range of conditions

Visual information is contained not only in the number of spikes generated by a stimulus but also in their timing (Butts et al., 2011; Gollisch and Meister, 2010), which can be precise to within a few milliseconds in RGCs (Berry et al., 1997a; Uzzell and Chichilnisky, 2004), the LGN (Reinagel and Reid, 2002) and visual cortex (Buracas et al., 1998). In RGCs, the temporal precision of spiking allows the latency to the first spike to transmit more information about a stimulus than the total number of spikes, at least when a stimulus drives episodic responses (Gollisch and Meister, 2008). To assess the impact of MVR on the temporal precision of spike generation we therefore analyzed how τ_l and k/N affecting first spike latencies (Fig. 5).

First spike latency depended on both the membrane time constant τ_l and the reliability of the synaptic inputs k/N (Fig. 5A-D). All but the very fastest time constants produced shorter latency spikes utilizing a rate code compared to a hybrid code, with delays being reduced by up to 4.5 ms. Cells requiring fewer vesicles to spike benefit from the wider temporal window of vesicle release in univesicular compared to multivesicular events (as seen in Fig. 1F). However, the more reliable the inputs the weaker this effect (Fig. 5D). At an intermediate range of k/N hybrid inputs generated spikes with shorter latencies and then at higher k/N there was no dependence on coding strategy. The general picture is that the rate code generates spikes with shorter latency only in larger neurons in which spiking is driven by larger numbers of less reliable synaptic inputs -

conditions more likely to occur in brain regions such as the cortex than the retina or other circuits with more reliable synapses.

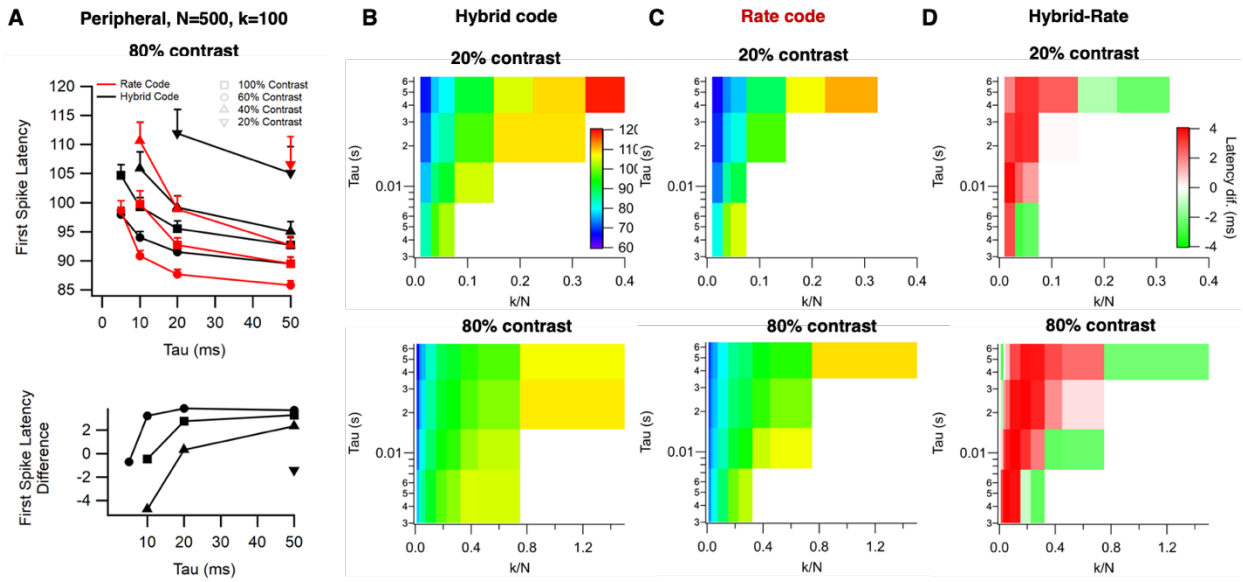

Figure 5. Hybrid coding reduced spike latency over a range of physiologically-relevant conditions

A. Top: Mean first spike latency as a function of leak time-constant (τ_l) for a cell receiving N=500 inputs and requiring k=100 vesicles to spike in response to a 5 Hz stimulus of varying contrast (other parameters as in the "large peripheral" cell in Fig. 3). Latencies were measured relative to an arbitrarily chosen phase of the sinusoidal stimulus so it is the change in these values that is most informative. Bottom: Latency Difference (Hybrid – Rate) from the data in top panel. For this ratio of k:N, the hybrid code produced at greater latency when τ_l > 15 ms. B. Top: heat map showing the average first spike latency generated by hybrid inputs for cells of varying k/N and τ_l for a stimulus of 20% contrast. Latencies were averaged over values of k and N shown in Table 1. Bottom: as top but for 80% contrast. C. As B but for rate-coded inputs. D. The difference in first spike latency generated by hybrid- and rate-coded inputs in B and C. For a given τ_l, rate-coded inputs generated faster responses at lower k/N while hybrid inputs generated faster responses at higher k/N (i.e when spikes are generated by activation of a larger proportion of inputs). Note the different ranges for k/N in upper and lower panels.

Coordinated MVR increases information transmission across a range of conditions

Having established that MVR impacts both spike count and timing we asked how these effects interact to alter the transmission of information (Borst and Theunissen, 1999). The mutual information was calculated between a set of nine stimuli (contrasts of 20%, 60% and 100% at frequencies of 5 Hz, 10 Hz and 15 Hz) and each of three measures of

the response in the model neuron: spike count, spike time and the total information within the spike sequence. A neuron that encodes a total of $\log_2(9) = 3.17$ bits of information over a 200 ms time window is discriminating perfectly between the nine stimuli.

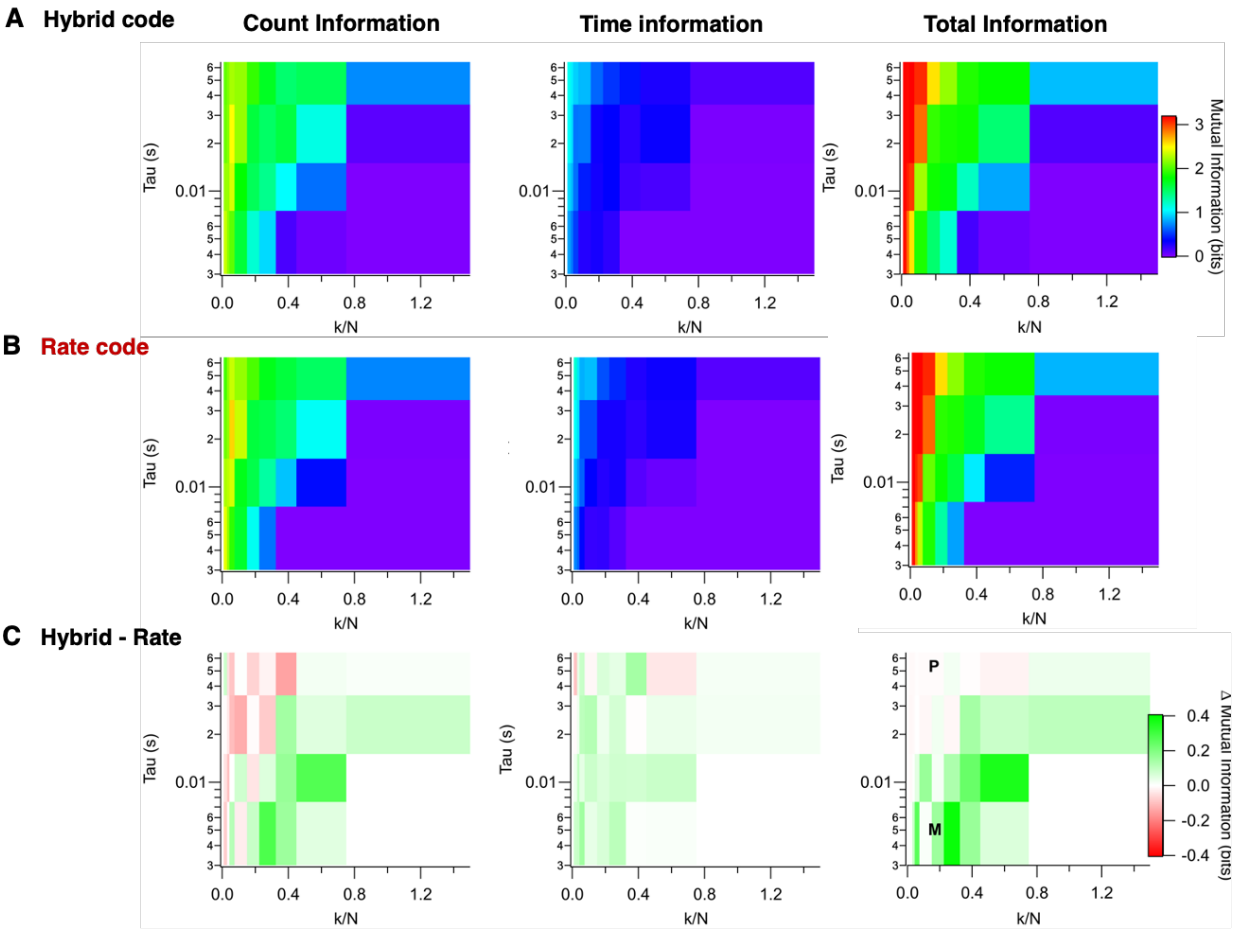

Figure 6. Hybrid coding increased information in the spike output over a range of physiologically-relevant conditions

A. Left: heat map showing the information transmitted in the spike count generated by hybrid synaptic inputs for cells of varying k/N and τ_I. Mutual Information was averaged over values of k and N shown in Table 1 and was measured over a time-window of 200 ms. These values can therefore be multiplied by 5 to obtain information rates as bits s^{-1}. Middle: Mutual Information in spike timing. Right: the total Mutual Information representing the sum of count and timing information. The colour scale bar applies to all three maps. Maximum possible information was 3.17 bits. More information was contained in the spike count compared to spike times. **B.** As A, but for rate-coded synaptic inputs operating at the same average release rates. **C.** The difference between A and B (hybrid - rate). The colour scale (right) shows combinations of parameters in which hybrid inputs generated more spikes in green. Zero difference is white and combinations of parameters in which rate-coded inputs generated more spikes are in red. Note that there are some situations in which count information is favoured by rate-coded inputs while time information is favoured by hybrid inputs (and vice versa) but total information is favoured by hybrid inputs over a wider range of conditions (right). The parameters for the specific model cells featured in Fig. 3 are indicated as M (midget) and P (peripheral) in the heat plot for the difference in total information.

562 The heat plots in Fig. 6 demonstrate that spike count carried more of the total
563 information than spike time for both the synaptic coding strategies. But the relative
564 advantages of MVR and rate codes again depended on the time-constant of the target
565 neuron and the reliability of the inputs driving spikes. In a cell of given time-constant, an
566 increase in the reliability of inputs caused the hybrid code to become more efficient at
567 generating spikes (Fig. 4) with shorter latency (Fig. 5) resulting in more information
568 transmission (Fig. 6). In the small cell, for instance, the spike count over a 200 ms period
569 carried 0.07 bits more information when generated by hybrid inputs, while spike time
570 carried 0.08 bits more, and the total information difference was ~0.16 bits (equivalent to
571 an information rate of 0.82 bits s^{-1}). The largest difference in information rate in favour of
572 the hybrid code was 3.8 bits s^{-1} or 24% of the maximum possible information rate,
573 occurring when the post-synaptic cell required 25 vesicles to spike, had 250 inputs, and a
574 1 ms time constant.

575 The rate code carried marginally more of the total information only in larger cells in
576 which greater numbers of less reliable synapses converge. For the large neuron
577 highlighted in Fig. 3, for instance, we calculated mutual information differences (hybrid –
578 rate) of -0.12 bits for spike count and 0.11 bits for spike time, yielding a total information
579 difference of -0.01 bits. The largest difference in information rate in favour of the rate
580 code was only 0.24 bits s^{-1}, occurring when only 10 vesicles were required to generate a
581 spike, with 50 inputs, and a time constant of 50 ms.

582 Together, the simulations in Figs. 2-6 indicate that synapses which generate MVR
583 can often trigger more spikes per released vesicle compared to a classical Poisson
584 synapse in which all vesicles are released independently. In cells with time-constants less
585 than 15 ms, the tendency for MVR to enhance the transmission of information becomes
586 greater as k/N falls below ~0.8. These conditions are physiologically relevant. In the
587 peripheral retina, for instance, alpha ganglion cells have membrane time-constants less
588 than 5 ms and ten or fewer quanta released from thousands of synapses are sufficient to
589 generate a spike (Freed and Sterling, 1988; O'Brien et al., 2002). More morphologically

590 compact neurons, such as midget ganglion cells in the fovea of primates, only receive
591 inputs from synapses from a single midget bipolar cell and are likely to have longer time-
592 constants (Sinha et al., 2017).
593
594 **MVR increases the efficiency of information transmission across a range of**
595 **conditions**
596 Neurons can transmit more information by spiking at higher rates but this comes at a
597 relatively large energetic cost (Laughlin, 2001). The release of vesicles and generation of
598 spikes is the major consumer of energy in the brain with one estimate being of the order of
599 ~24,000 ATP molecules consumed per bit of information transmitted (Attwell and
600 Laughlin, 2001; Harris et al., 2012). The efficiency with which vesicles drive the spike
601 code is therefore likely to have been an evolutionary constraint on the vesicle code itself
602 (Sterling and Laughlin, 2015).
603 To investigate the potential effects of MVR on the efficiency of information
604 transmission we normalized the measurements of total information summarized in Fig. 6A
605 and B as bits per vesicle and bits per spike, as shown in Fig. 7A and B. Considering the
606 amount of information per vesicle, the relative advantages of hybrid and rate codes can be
607 seen to follow a clear pattern: the more reliably inputs responded (i.e the higher k/N) the
608 broader the range of cell time-constants over which hybrid-coded inputs improved the
609 efficiency with which vesicles transferred information through spikes in the target neuron
610 (Fig. 7A right). The rate-code was only slightly advantageous in neurons acting as the
611 best temporal integrators receiving the least reliably responding inputs. The highest
612 vesicle efficiency using either synaptic code was ~0.044 bits per vesicle.
613

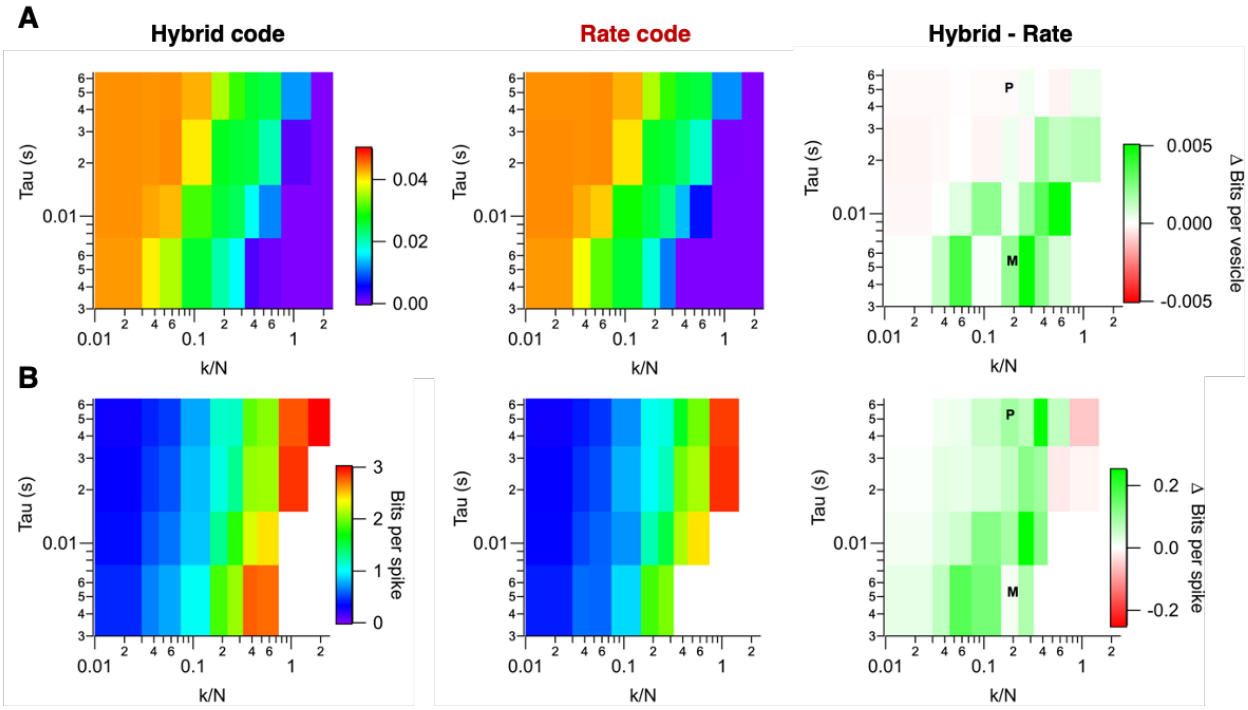

Figure 7. Hybrid coding increased the efficiency of information transmission over a range of conditions

A. Left: heat map showing the information transmitted per vesicle generated by hybrid synaptic inputs for cells of varying k/N and τ_l. Middle: the same for rate-coded synaptic inputs operating at the same average release rates. Right: The difference in the efficiency of information transmission considered as bits per vesicle (hybrid - rate). The colour scale (right) shows combinations of parameters in which hybrid inputs generated more spikes in green. The parameters for the specific model cells featured in Fig. 3 are indicated as M (midget) and P (peripheral) in the heat plot. **B.** As A, but considering efficiency as bits per spike. Note that the hybrid vesicle code increases the efficiency of information transmission across a range of conditions. The pure rate vesicle code was only advantageous in LIF neurons with the longest time-constants driven by the most reliably responding synaptic inputs.

A similar pattern was observed when we quantified the amount of information per spike: the more reliably inputs responded the broader the range of cell time-constants over which hybrid-coded inputs increased the amount of information per spike (Fig. 7B right). The highest spike efficiency was ~3 bits per spike. A comparison can be made with the information transmitted by spikes in RGCs responding to white-noise stimuli, where the most sluggish transmit ~3.5 bits/spike, while those that fire most briskly encode ~2 bits/spike (Koch et al., 2006). The results in Fig. 7 encapsulate what we believe is the most fundamental insight of this study – the potential of MVR to increase the efficiency with which vesicles are used to transmit information.

636 **The first synapse in hearing**

637 MVR is also a prominent property of the ribbon synapse of auditory hair cells, where the
638 fusion of multiple vesicles can be synchronized on timescales as short as 100 μs (Fuchs
639 and Glowatzki, 2015; Glowatzki and Fuchs, 2002). This presents a special case because
640 there is no convergence onto the target, each Type I afferent fiber receiving sensing
641 glutamate released from just one ribbon synapse. Experiments indicate that spikes in the
642 afferent fiber can be triggered by just a few vesicles and in some cases perhaps as a few
643 as one (Grabner and Moser, 2018; Rutherford et al., 2012). We therefore considered
644 two broad situations: i) k = 1 where *any* vesicular event, regardless of quantal content, is
645 capable of generating a spike, and ii) k > 1, where some (but not all) multivesicular events
646 are capable of generating a spike from rest.

647 The transmission of information was strongly dependent on both the number of
648 vesicles required to generate a spike, k, and the membrane time constant but again the
649 nature of this dependence differed for synapses employing a pure rate code compared to
650 those also employing MVR (Fig. 8A). Rate coding produced the maximum mutual
651 information in the simulations when k = 1, where the spike sequence in the afferent
652 reproduced the vesicle sequence leaving the synapse (Fig. 7A, red). In this situation a
653 multivesicular event is no more effective than a univesicular event so the hybrid code
654 produces fewer spikes per vesicle and reduces the efficiency of information transmission
655 (Fig. 8A, black).

656

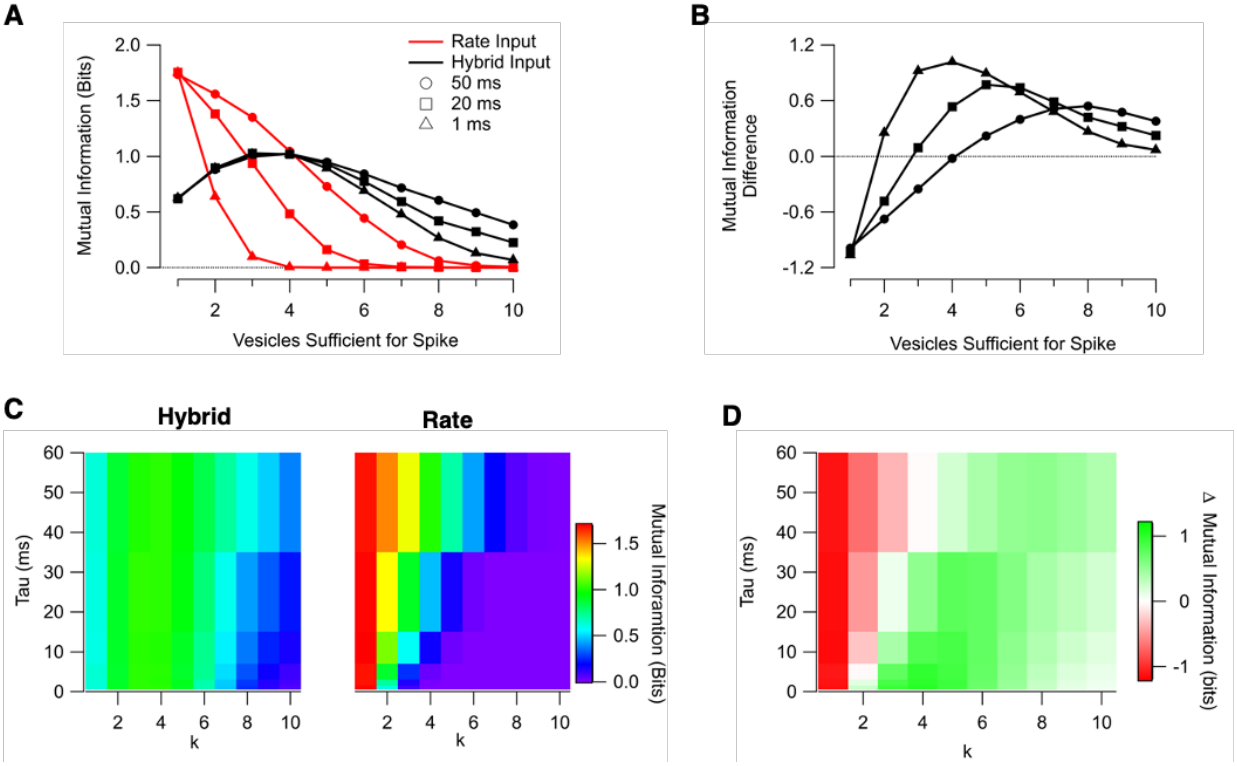

Figure 8. Spike Count Information for Amplitude (black) and Rate (red) inputs as a function of τ_I and k where N = 1. A. Mutual information for the rate input (dash) and amplitude input (solid) as a function of the vesicles sufficient to generate a spike. Values shown for τ_I of 50, 20, 10, and 1 ms. **B**. Mutual information difference (hybrid – rate) for the data shown in A. **C**. Mutual information (as **A**) in matrix form. **D.** Mutual information difference in matrix form.

As the gain of synaptic transmission decreased and more vesicles were required to generate a spike (k > 1), the picture began to change. Increasing k reduced the spike count and effectively abolished spiking at k > 3 when tau = 1 ms (Fig. 8A and B). Essentially, this is because the very short time-constant of the target neuron (mimicking an afferent fiber) did not allow for effective summation of vesicles. Hybrid coding, however, expanded the range of responses by increasing the number of spikes a stimulus could generate, maximizing at a value when most events at higher contrast individually generated a spike (approximately k = 4 for this range of conditions, as seen in Fig. 8B and D). This switch in the relative efficiency of the two synaptic coding strategies starkly illustrates a key functional advantage of the hybrid code: presynaptic integration of vesicles by MVR compensates for situations where the integration time of the post-synaptic neuron is short and limits summation post-synaptically.

676 The heatmap in Fig. 8D demonstrates that the increase in the efficiency of information
677 transmission caused by MVR was most dramatic in a target neuron with time-constants
678 less than about 10-15 ms in which a spike could be triggered by the arrival of 3-5 vesicles.
679 For $5 < k < 10$, MVR still utilized vesicles more efficiently than a pure rate-code (Fig. 8D),
680 despite the fact that not all MVR events were capable of generating a spike even at the
681 highest contrasts. MVR in rat hair cells can generate events equivalent to up to ~22
682 vesicles (Glowatzki and Fuchs, 2002) suggesting that the advantages of a hybrid code
683 might be exerted over a broad range of k in the auditory system.
684

685 **Discussion**

686 The motivation behind this study was to understand how the unusual statistics of vesicle
687 release from ribbon synapses affected the conversion of analogue sensory signals into
688 spikes. We compared two different regimes of synaptic coding known to operate in
689 different parts of the brain: independent release generating a pure rate code and a hybrid
690 code in which information is contained in both the rate and amplitude of synaptic events.
691 The results indicate that MVR can increase the efficiency with vesicles are used to
692 transmit sensory information over a range of conditions mimicking transmission of visual
693 signals to retinal ganglion cells of different sizes (Figs. 4, 6-7). In the special case of
694 auditory hair cells driving afferent fibers through a single synapse, MVR increased the
695 efficiency of information transfer whenever spike generation required depolarization
696 greater than that caused by a single vesicle (Fig. 8).

697 Two aspects of the process driving spikes interacted to determine the relative
698 efficiency of a pure rate-code compared to a hybrid code: the time-constant of the target
699 neuron and the reliability of the synaptic inputs. MVR tended to be advantageous when i)
700 the neurons time-constant was shorter, causing less post-synaptic summation of synaptic
701 potentials at physiological rates of vesicle release, and ii) individual synaptic inputs were
702 more reliable so that a lower degree of convergence was required to depolarize the target
703 to threshold for a given stimulus strength.

704 **What do ribbon synapses employ MVR to transmit sensory signals?**

705 Neurons represent sensory stimuli, at least in part, using a rate-code, where information is
706 contained in the number of spikes over a given time-window (Rieke and Warland, 1999).
707 But it has long been recognized that a potential drawback of a pure rate-code is the time-
708 scale on which it can transmit information: the more variable a neural response to a
709 stimulus the longer the observation time required to detect it to a given degree of certainty
710 (Gerstner et al., 1997; Stein et al., 2005). At the ribbon synapse of bipolar cells, for
711 instance, the maximum sensitivity to changes in contrast occurs around an average
712 release rate R of 20 vesicles s^{-1} and a 10% increase in contrast increases release by a
713 factor of a = 1.8 (James et al., 2019). If release events are Poisson distributed, signal
714 detection theory can be used to quantify the time-window over which the response would
715 need to be observed to differentiate it from ongoing release: a SNR of 2 would require
716 counting an extra 14 vesicles over an observation time (Δt) of ~0.9 s (equations 7-9 in
717 Methods). But vision operates much more rapidly. Our visual system can process a
718 scene in less than 150 ms (Thorpe et al., 1996) and neurons in the cortex of monkeys can
719 respond to images changing at intervals as short as 14 ms (Keysers et al., 2001).

720 The convergence of bipolar cell synapses onto RGCs is of course one features of the
721 retinal circuit that will shorten the time-scale of information transmission using a rate-code.
722 N equivalent Poisson inputs allows a given SNR to be achieved with a time-window that
723 varies as $1/\sqrt{N}$ (Faisal et al., 2008; Gerstner et al., 1997; Rusakov et al., 2020).
724 Continuing with the example above and considering a time-window of 100 ms, the output
725 of 50 synapses would provide a SNR of 2 in a time-window of 100 ms, but this would
726 require the release of $N*(1-a)*R*\Delta t$ = 80 extra vesicles. This is a very significant cost
727 given that synaptic transmission is a major source of energy consumption in the brain
728 (Attwell and Laughlin, 2001).

729 The potential problems with a pure rate code highlighted above are closely linked to
730 the idea that the firing rate of a neuron or release rate of a synapse encodes the sensory
731 variable continuously. In fact, if a varying stimulus is applied, most RGCs fire in bursts

732 from rest, driven by fusillades of vesicles released from bipolar cells (Berry et al., 1997b;
733 Freed, 2005; Keat et al., 2001). The information that RGCs convey is therefore
734 represented not just by the spike count but also by the precise timing of spikes (Gutig et
735 al., 2013; Meister et al., 1995; Victor, 2000) and it has been shown that such a temporal
736 coding strategy can significantly improve the speed of information transmission (Butts et
737 al., 2007; Gollisch and Meister, 2008). Crucially, our measurements of glutamate release
738 from individual ribbon synapses of bipolar cells demonstrate that these also signal visual
739 events in bursts, including MVR events of varying amplitude (Fig. 1A and C).

740 A consideration of different synaptic coding strategies should therefore also consider
741 the representation of temporally discrete events rather than the continuous encoding of a
742 stimulus variable such as contrast. It has been suggested that the onset of a burst in an
743 RGC provides information about when a visual event occurs, while the number of spikes
744 within the burst represent properties such as the magnitude of the event (Keat et al.,
745 2001). The temporal jitter in the first spike within a burst can be as low as a few
746 milliseconds (Berry et al., 1997a; Freed, 2005) and it is notable that larger MVR events
747 from bipolar cells achieve a similar temporal precision (Fig. 1E and F). This comparison
748 highlights one of the potential advantages of representing a stimulus using synaptic
749 symbols of different amplitude: an unexpected symbol immediately imparts new
750 information. In support of this idea, it has been shown that larger MVR events that are
751 more likely to trigger a spike are rarer and carry more information (James et al., 2019).
752

753 **Encoding auditory events**

754 The spike code carrying auditory information in cochlear afferents is determined only by
755 excitatory inputs from hair cells and is therefore likely to reflect the temporal precision of
756 ribbon synapses more directly than the spike code leaving the retina. The synaptic output
757 from hair cells has been measured *in vitro* with electrophysiological precision where it is
758 found that larger MVR events are more tightly phase-locked to a sinusoidal voltage-clamp
759 stimulus (Li et al., 2014). Notably, MVR events in hair cells can be as large as ~20 vesicle

760 equivalents when triggered using electrophysiology (Glowatzki and Fuchs, 2002) but it is
761 unclear if these can also be triggered by sound. Here we used the statistics of MVR
762 measured in bipolar cells *in vivo*, where the largest events evoked by the highest
763 contrasts are composed of 9-11 vesicles (James et al., 2019). A better understanding of
764 how the ribbon synapse of hair cells drive the spike code in auditory fibers will require
765 experimental measurements of the statistics of vesicle release in response to sounds of
766 different intensities and frequencies in hair cells with different temporal tuning.
767
768 **Future directions**
769 Our analysis of mutual information was more concerned with the "what" of the stimulus
770 (contrast), and spike count carried more of the total information than spike time for either
771 of the synaptic coding strategies (Fig. 7). This feature of the results might, however,
772 represent limitations of the LIF model in which we only considered excitation. The timing
773 of spikes in RGCs is also dependent on inhibitory inputs, especially feedforward inhibition
774 that increases the temporal precision of spikes and, crucially, also causes spikes to be
775 fired in bursts (Johnston and Lagnado, 2015; Murphy-Baum and Taylor, 2018). Network
776 models incorporating a wider range of biophysical parameters will help us understand how
777 excitation and inhibition interact to determine the temporal precision of the spike code
778 leaving the retina.
779 To put the statistics of vesicle release into a functional context we also need an
780 understanding of the spike code that it drives. The most decisive next step would be to
781 obtain direct experimental measurements of how effectively coordinated MVR events of
782 different size trigger spikes. The relatively simple organization of the first stage in hearing,
783 where spikes in afferents are driven by a single ribbon synapse, provides a beautiful
784 context in which to ask this question once the experimental tools are established. The
785 situation in the retina will be more complex because RGCs come in a wide variety of
786 morphological and functional types with different time-constants and number and strength
787 of synaptic inputs. In the peripheral retina, for instance, alpha ganglion cells have

membrane time-constants less than 5 ms and on average ten quanta are released from thousands of synapses for each spike generated (Freed, 2005). More morphologically compact neurons, such as midget ganglion cells in the fovea of primates, only receive inputs from a single midget bipolar cell and are likely to have much longer time-constants (Sinha et al., 2017). This variety of different functional contexts could also be explored with biophysical models that are more realistic than the LIF neuron that we have begun with here, potentially taking into account non-linearities such as saturation of post-synaptic glutamate receptors and active dendritic conductances.

References

Abbott, L.F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Res Bull *50*, 303-304.

Attwell, D., and Laughlin, S.B. (2001). An energy budget for signaling in the grey matter of the brain. Journal of Cerebral Blood Flow & Metabolism *21*, 1133-1145.

Auger, C., Kondo, S., and Marty, A. (1998). Multivesicular release at single functional synaptic sites in cerebellar stellate and basket cells. J Neurosci *18*, 4532-4547.

Baden, T., Esposti, F., Nikolaev, A., and Lagnado, L. (2011). Spikes in retinal bipolar cells phase-lock to visual stimuli with millisecond precision. Curr Biol *21*, 1859-1869.

Berry, M.J., Warland, D.K., and Meister, M. (1997a). The structure and precision of retinal spike trains. Proceedings of the National Academy of Sciences *94*, 5411-5416.

Berry, M.J., Warland, D.K., and Meister, M. (1997b). The structure and precision of retinal spike trains. Proc Natl Acad Sci U S A *94*, 5411-5416.

Borst, A., and Theunissen, F.E. (1999). Information theory and neural coding. Nat Neurosci *2*, 947-957.

Buracas, G.T., Zador, A.M., DeWeese, M.R., and Albright, T.D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron *20*, 959-969.

Burkitt, A.N. (2006a). A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. Biol Cybern *95*, 1-19.

Burkitt, A.N. (2006b). A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. Biol Cybern *95*, 97-112.

Butts, D.A., Weng, C., Jin, J., Alonso, J.M., and Paninski, L. (2011). Temporal precision in the visual pathway through the interplay of excitation and stimulus-driven suppression. J Neurosci *31*, 11313-11327.

Butts, D.A., Weng, C., Jin, J., Yeh, C.-I., Lesica, N.A., Alonso, J.-M., and Stanley, G.B. (2007). Temporal precision in the neural code and the timescales of natural vision. Nature *449*, 92-95.

Calkins, D.J., Schein, S.J., Tsukamoto, Y., and Sterling, P. (1994). M and L cones in macaque fovea connect to midget ganglion cells by different numbers of excitatory synapses. Nature *371*, 70-72.

Choi, S.Y., Borghuis, B.G., Rea, R., Levitan, E.S., Sterling, P., and Kramer, R.H. (2005). Encoding light intensity by the cone photoreceptor synapse. Neuron *48*, 555-562.

Christie, J.M., and Jahr, C.E. (2006). Multivesicular release at Schaffer collateral-CA1 hippocampal synapses. J Neurosci *26*, 210-216.

Destexhe, A., and Rudolph, M. (2004). Extracting Information from the Power Spectrum of Synaptic Noise. Journal of Computational Neuroscience *17*, 327-345.

Faisal, A.A., Selen, L.P., and Wolpert, D.M. (2008). Noise in the nervous system. Nat Rev Neurosci *9*, 292-303.

Freed, M.A. (2005). Quantal encoding of information in a retinal ganglion cell. J Neurophysiol *94*, 1048-1056.

Freed, M.A., and Sterling, P. (1988). The ON-alpha ganglion cell of the cat retina and its presynaptic cell types. J Neurosci *8*, 2303-2320.

Fuchs, P.A., and Glowatzki, E. (2015). Synaptic studies inform the functional diversity of cochlear afferents. Hear Res *330*, 18-25.

Gerstner, W., Kreiter, A.K., Markram, H., and Herz, A.V. (1997). Neural codes: firing rates and beyond. Proc Natl Acad Sci U S A *94*, 12740-12741.

Glowatzki, E., and Fuchs, P.A. (2002). Transmitter release at the hair cell ribbon synapse. Nat Neurosci *5*, 147-154.

Goldberg, J.M., and Brown, P.B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. J Neurophysiol *32*, 613-636.

Gollisch, T., and Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. Science *319*, 1108-1111.

Gollisch, T., and Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron *65*, 150-164.

Grabner, C.P., and Moser, T. (2018). Individual synaptic vesicles mediate stimulated exocytosis from cochlear inner hair cells. Proc Natl Acad Sci U S A *115*, 12811-12816.

Harris, Julia J., Jolivet, R., and Attwell, D. (2012). Synaptic Energy Use and Supply. Neuron *75*, 762-777.

Hays, C., Sladek, A., Field, G., and Thoreson, W. (2021). Properties of multi-vesicular release from mouse rod photoreceptors support transmission of single photon responses. eLife *10*.

Holler, S., Kostinger, G., Martin, K.A.C., Schuhknecht, G.F.P., and Stratford, K.J. (2021). Structure and function of a neocortical synapse. Nature *591*, 111-116.

Huang, C.H., Bao, J., and Sakaba, T. (2010). Multivesicular release differentiates the reliability of synaptic transmission between the visual cortex and the somatosensory cortex. J Neurosci *30*, 11994-12004.

Ishii, T., and Hosoya, T. (2020). Interspike intervals within retinal spike bursts combinatorially encode multiple stimulus features. PLoS Comput Biol *16*, e1007726.

Jacoby, R.A., Wiechmann, A.F., Amara, S.G., Leighton, B.H., and Marshak, D.W. (2000). Diffuse bipolar cells provide input to OFF parasol ganglion cells in the macaque retina. J Comp Neurol *416*, 6-18.

James, B., Darnet, L., Moya-Diaz, J., Seibel, S.H., and Lagnado, L. (2019). An amplitude code transmits information at a visual synapse. Nat Neurosci *22*, 1140-1147.

Johnston, J., and Lagnado, L. (2015). General features of the retinal connectome determine the computation of motion anticipation. eLife *4*, e06250.

Kaplan, E., Purpura, K., and Shapley, R.M. (1987). Contrast affects the transmission of visual information through the mammalian lateral geniculate nucleus. The Journal of physiology *391*, 267-288.

Kayser, C., Montemurro, M.A., Logothetis, N.K., and Panzeri, S. (2009). Spike-Phase Coding Boosts and Stabilizes Information Carried by Spatial and Temporal Spike Patterns. Neuron *61*, 597-608.

Keat, J., Reinagel, P., Reid, R.C., and Meister, M. (2001). Predicting Every Spike: A Model for the Responses of Visual Neurons. Neuron *30*, 803-817.

Keysers, C., Xiao, D.-K., Földiák, P., and Perrett, D.I. (2001). The speed of sight. Journal of cognitive neuroscience *13*, 90-101.

Koch, K., McLean, J., Segev, R., Freed, M.A., Berry, M.J., 2nd, Balasubramanian, V., and Sterling, P. (2006). How much the eye tells the brain. Curr Biol *16*, 1428-1434.

Kolb, H., and Marshak, D. (2003). The midget pathways of the primate retina. Doc Ophthalmol *106*, 67-81.

Lagnado, L., and Schmitz, F. (2015). Ribbon Synapses and Visual Processing in the Retina. Annu Rev Vis Sci *1*, 235-262.

Laughlin, S.B. (2001). Energy as a constraint on the coding and processing of sensory information. Curr Opin Neurobiol *11*, 475-480.

Levine, M.W. (1991). The distribution of the intervals between neural impulses in the maintained discharges of retinal ganglion cells. Biological Cybernetics *65*, 459-467.

Li, G.L., Cho, S., and von Gersdorff, H. (2014). Phase-locking precision is enhanced by multiquantal release at an auditory hair cell ribbon synapse. Neuron *83*, 1404-1417.

Li, X., and Ascoli, G.A. (2008). Effects of synaptic synchrony on the neuronal input-output relationship. Neural Comput *20*, 1717-1731.

Lisman, J.E., Raghavachari, S., and Tsien, R.W. (2007). The sequence of events that underlie quantal transmission at central glutamatergic synapses. Nat Rev Neurosci *8*, 597-609.

Meister, M., and Berry, M.J., 2nd (1999). The neural code of the retina. Neuron *22*, 435-450.

Moya-Díaz, J., James, B., Esposti, F., Johnston, J., and Lagnado, L. (2022). Diurnal changes in the efficiency of information transmission at a sensory synapse. bioRxiv, 2021.2009.2012.459944.

Murphy-Baum, B.L., and Taylor, W.R. (2018). Diverse inhibitory and excitatory mechanisms shape temporal tuning in transient OFF α ganglion cells in the rabbit retina. The Journal of physiology *596*, 477-495.

Neef, A., Khimich, D., Pirih, P., Riedel, D., Wolf, F., and Moser, T. (2007). Probing the mechanism of exocytosis at the hair cell ribbon synapse. J Neurosci *27*, 12933-12944.

O'Brien, B.J., Isayama, T., Richardson, R., and Berson, D.M. (2002). Intrinsic physiological properties of cat retinal ganglion cells. J Physiol *538*, 787-802.

Odermatt, B., Nikolaev, A., and Lagnado, L. (2012). Encoding of luminance and contrast by linear and nonlinear synapses in the retina. Neuron *73*, 758-773.

Panzeri, S., Senatore, R., Montemurro, M.A., and Petersen, R.S. (2007). Correcting for the sampling bias problem in spike train information measures. J Neurophysiol *98*, 1064-1072.

Pillow, J.W., Paninski, L., Uzzell, V.J., Simoncelli, E.P., and Chichilnisky, E.J. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. J Neurosci *25*, 11003-11013.

Rathbun, D.L., Warland, D.K., and Usrey, W.M. (2010). Spike timing and information transmission at retinogeniculate synapses. J Neurosci *30*, 13558-13566.

Reinagel, P., and Reid, R.C. (2002). Precise firing events are conserved across neurons. J Neurosci *22*, 6837-6841.

Rieke, F., and Warland, D. (1999). Spikes: exploring the neural code (MIT press).

Rusakov, D.A., Savtchenko, L.P., and Latham, P.E. (2020). Noisy Synaptic Conductance: Bug or a Feature? Trends Neurosci *43*, 363-372.

Rutherford, M.A., Chapochnikov, N.M., and Moser, T. (2012). Spike encoding of neurotransmitter release timing by spiral ganglion neurons of the cochlea. J Neurosci *32*, 4773-4789.

Shannon, C. (1948). A Mathematical Theory of Communication. Bell System Technical Journal *27*, 379-423.

Singer, J.H., Lassova, L., Vardi, N., and Diamond, J.S. (2004). Coordinated multivesicular release at a mammalian ribbon synapse. Nat Neurosci *7*, 826-833.

Sinha, R., Hoon, M., Baudin, J., Okawa, H., Wong, R.O.L., and Rieke, F. (2017). Cellular and Circuit Mechanisms Shaping the Perceptual Properties of the Primate Fovea. Cell *168*, 413-426.e412.

Stein, R.B., Gossen, E.R., and Jones, K.E. (2005). Neuronal variability: noise or part of the signal? Nature Reviews Neuroscience *6*, 389-397.

Sterling, P., and Laughlin, S. (2015). Principles of neural design (MIT press).

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. Nature *381*, 520-522.

Usrey, W.M., and Reid, R.C. (1999). Synchronous activity in the visual system. Annual review of physiology *61*, 435-456.

Usrey, W.M., Reppas, J.B., and Reid, R.C. (1998). Paired-spike interactions and synaptic efficacy of retinal inputs to the thalamus. Nature *395*, 384-387.

Uzzell, V.J., and Chichilnisky, E.J. (2004). Precision of spike trains in primate retinal ganglion cells. J Neurophysiol *92*, 780-789.

Victor, J.D. (2006). Approaches to Information-Theoretic Analysis of Neural Activity. Biol Theory *1*, 302-316.