Multivesicular Release Increases the Efficiency of Information Transmission at Sensory 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

The statistics of vesicle release determine how information is transferred in neural circuits. The classical model is of Poisson synapses releasing vesicles independently but ribbon synapses transmit early sensory signals by multivesicular release (MVR) when two or more vesicles are coordinated as a single synaptic event. To investigate the impact of MVR on the spike code we used leaky integrate-and-fire models with inputs simulating the statistics of vesicle release measured experimentally from retinal bipolar cells. Comparing these with models of independent release we find that MVR increases spike generation and the efficiency of information transfer (bits per spike) over a range of conditions that mimic retinal ganglion cells of different time-constant receiving different number of synaptic inputs of different strengths. When a single input drives a neuron with short time-constant, as occurs when hair cells transmit auditory signals, MVR increases information transfer whenever spike generation requires depolarization greater than that caused by a single vesicle. This study demonstrates how presynaptic integration of vesicles by MVR can compensate for less effective summation post-synaptically to increase the efficiency with which sensory information is transmitted at the synapse.

Introduction

The fundamental units by which information is encoded and transmitted in the nervous system are spikes and synaptic vesicles. Electrophysiology has allowed the spike code to be studied in detail (Gerstner et al., 1997; Quian Quiroga and Panzeri, 2009; Schwartz et al., 2006), especially in the context of the representation of sensory information (Carleton et al., 2010; Gollisch and Meister, 2008; Kayser et al., 2009), but less is understood about the vesicle code operating at excitatory synapses that drive spiking (Manwani and Koch, 2001; Rusakov et al., 2020; Zador, 1998).

Theoretical and computational approaches to understanding how synapses transfer sensory information have been based largely on the idea that changes in presynaptic potential are communicated by modulating the mean rate of a Poisson process in which vesicles at the active zone are triggered to fuse independently (Johnson, 1996; Pillow et al., 2005; Schwartz et al., 2006). When the release probability is low, this scenario can be simplified further to the idea of “one-spike-one-vesicle” to create a purely binary output in which a spike invading the active zone sometimes triggers release of a single vesicle and sometimes none (Lisman et al., 2007). More recently, however, a range of experimental evidence has demonstrated that synapses at many sites in the brain do not follow these simple behaviours but can also release two or more vesicles within milliseconds - a process termed multivesicular release (MVR). MVR has been observed in the hippocampus (Christie and Jahr, 2006), cerebellum (Auger et al., 1998) and somatosensory cortex (Holler et al., 2021; Huang et al., 2010) and is also a fundamental feature of ribbon-type synapses specialized to transmit signals in the first stages of vision and hearing, where it does not obey Poisson statistics (Glowatzki and Fuchs, 2002; Hays et al., 2021; James et al., 2019; Lagnado and Schmitz, 2015; Neves and Lagnado, 1999; Singer et al., 2004). A recent combination of electrophysiology with correlative light-and electron-microscopy has even led to the suggestion that MVR may contribute to synaptic transmission throughout the nervous system (Holler et al., 2021).

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. 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 visual information can be represented 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).

How does MVR impacts on spike generation in the target neuron? 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, b; Gerstner et al., 1997). We gather these statistics from the ribbon-type synapse of retinal bipolar cells across which visual information is transmitted to post-synaptic ganglion cells, but the results can also be generalized to the transmission of auditory information from the ribbon synapses of hair cells.

We find that over a range of conditions MVR increased the number of spikes per vesicle compared to a simple rate-code in which all vesicles were released independently. The relative efficiency of information transfer, quantified as bits carried by the spiking response per vesicle released, differed between the two synaptic coding strategies by amounts that depended on the time-constant of the target neuron and the reliability of the

synaptic inputs. MVR tended to be advantageous when i) the neurons time-constant was shorter, causing less post-synaptic summation of synaptic potentials at physiological rates of vesicle release, and ii) individual synaptic inputs were more reliable so that a lower degree of convergence was required to depolarize the target to threshold for a given stimulus strength. The general picture is that presynaptic integration of vesicles by MVR can compensate for less effective summation post-synaptically to increase the efficiency with which sensory information is transmitted at the synapse.

Methods

Modelling Overview

Modelling the effects of MVR on spiking 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); ii) the probability of events composed of different numbers of quanta, P_{Qe} (Fig. 1C), 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).

2. Simulation of synapse output using Gaussian statistics. 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.

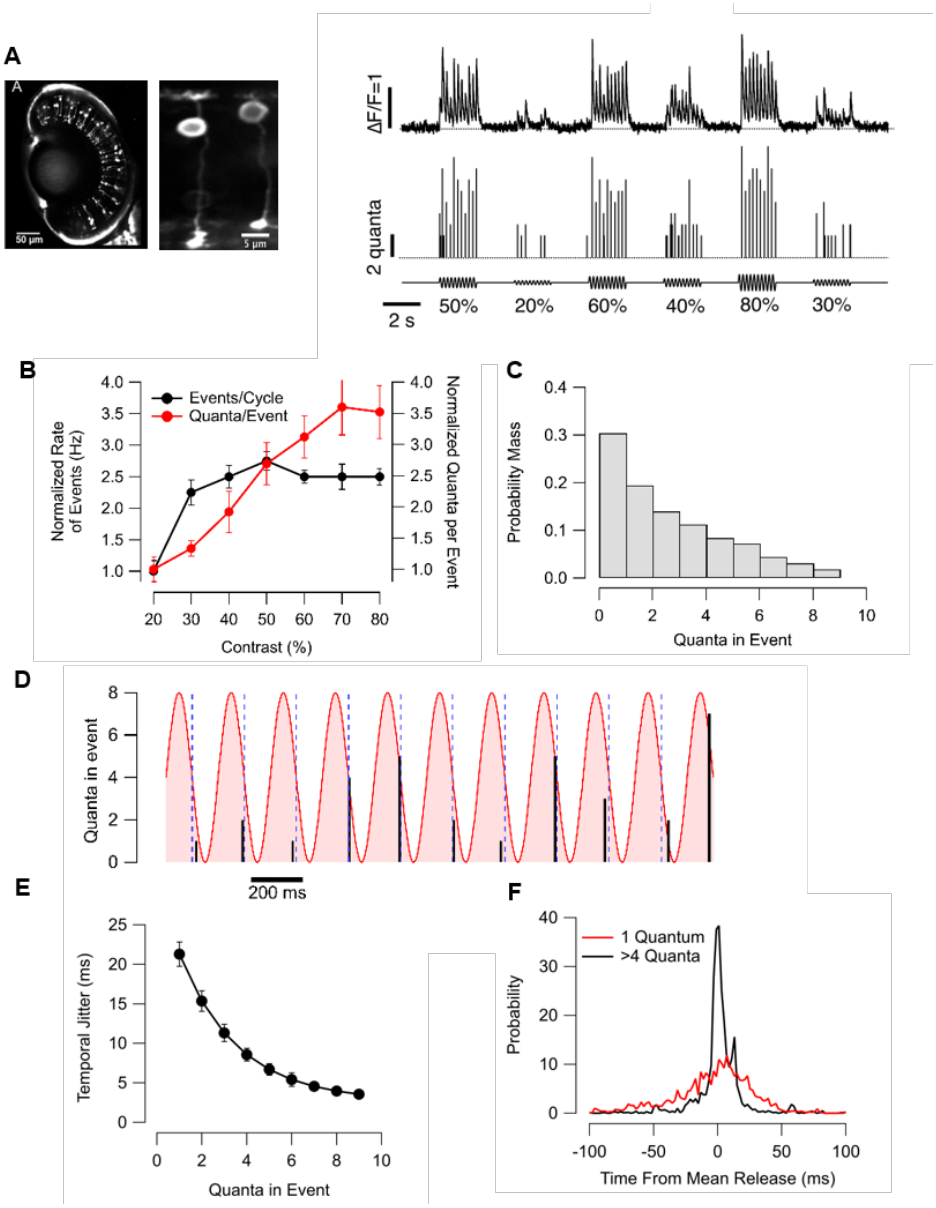


Figure 1. Multivesicular Release in the Retina of Zebrafish. A. Left: Multiphoton image of BCs in the retina of a larval zebrafish. They are visualized by the iGluSnFR reporter in the surface membrane. Middle: Two BCs at higher magnification. The mosaic expression of the reporter allows for distinguishing 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) (60 synapses). **C.** Distribution of event amplitudes during a 30 s application of a 5 Hz stimulus at 80% contrast (60 synapses), truncated to only include events containing fewer than 11 quanta. **D.** 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**. 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. Results compiled from a total of 60 synapses after mean shifting.

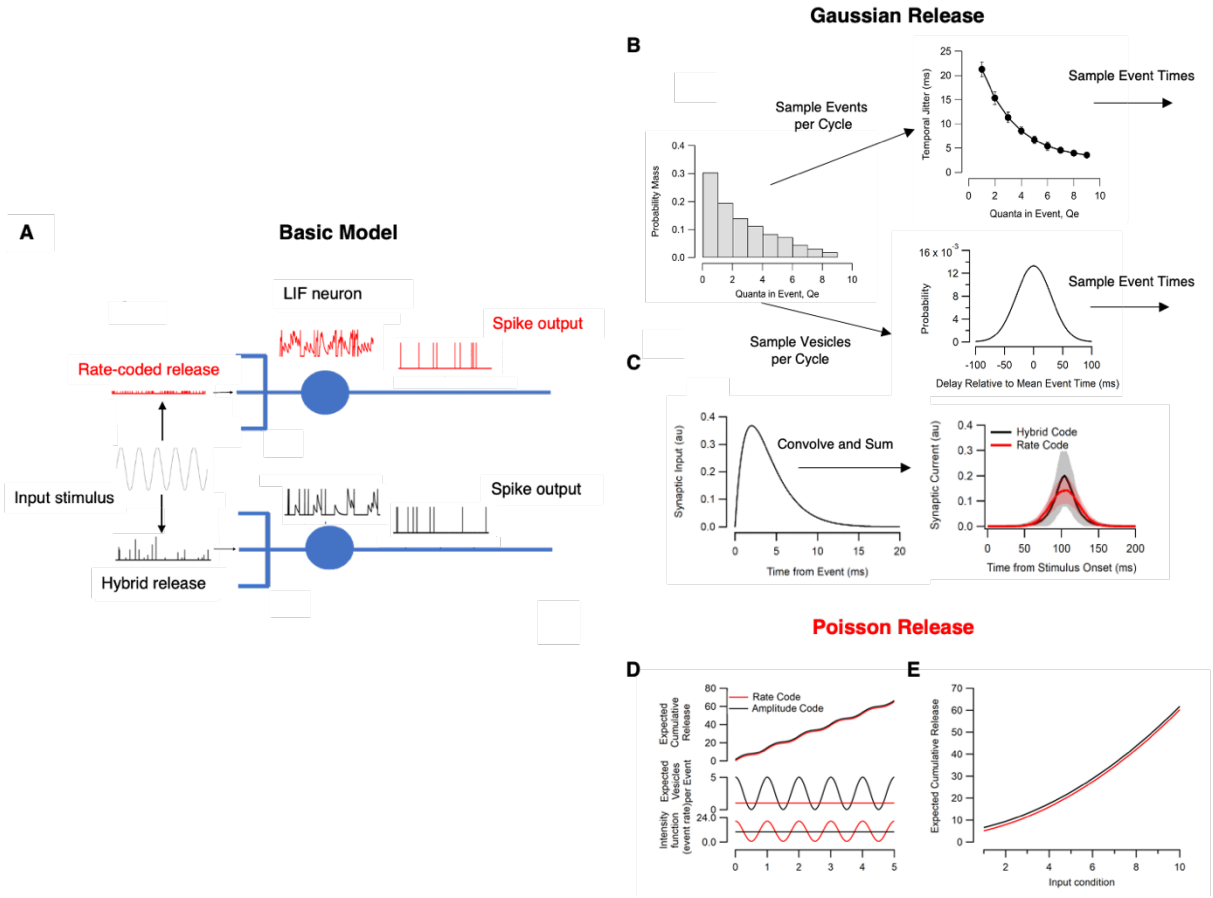


Figure 2. Overview of modelling. **A**. A leaky-integrate-and-fire neuron received either rate (top) or hybrid (bottom) inputs. **B.** Illustration of release simulation for the Gaussian Model. The number of vesicles in an event (for the MVR case) or the number of released uniquantal events (for the 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 for the hybrid 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'.

The output of a synapse employing 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). The time of each event relative to the stimulus was then set with a temporal precision

147 dependent on its amplitude, defined by TJ(Qe). The output was constrained to be
148 composed of only one or zero events per cycle of the stimulus, with the probability of no
149 event given by $p_Q(f, c)$. The output of a hybrid synapse employing MVR was therefore
150 described as a vector E of event times and a vector Qe of event amplitudes, both of length
151 N. The rate-code comparison was constructed by replacing each event consisting of
152 more than one quantum with an equivalent number of uni-quantal events, each event time
153 now sampled from the distribution of times of uniquantal events (Fig. 2B). The vectors
154 describing a pure rate synapse were therefore of variable lengths corresponding to the
155 total number of vesicles sampled within the 30 s sample period. The resulting vectors Qe
156 and E of event quanta and event times were then passed into a conductance-based
157 model LIF model (Fig. 1) and the spike output analysed.

158

159 ***3. Simulation of synapse output using marked Poisson Processes.*** For purely rate-
160 coded synapses, we used a Nonhomogeneous Poisson Process (NHPP) as the driving
161 force for vesicle release, in which the mean rate λ(t) varied sinusoidally in time. By
162 definition, simple Poisson Processes describe a situation in which all events occur
163 independently and with exponentially distributed interevent times and cannot, therefore,
164 be used to model MVR when two or more vesicles are released within a single synaptic
165 event. To construct an amplitude-modulated Poisson Process we turned to a modification
166 known as Poisson Splitting or Location-dependent thinning (Resnick, 1992), where events
167 can be assigned a type based upon their location in time (Supplementary Figure 1). In this
168 case events occur with a constant rate but the number of vesicles in each event is chosen
169 based upon when the event occurred. We could therefore drive the LIF model neuron in
170 two distinct regimes: a pure rate case, where all information is encoded by modulating the
171 mean rate of release of individual vesicles and a hybrid case, where information is
172 encoded by modulating both the rate and quantal content of synaptic events.

173 A key aspect of our investigation was to construct the rate and hybrid models in such
174 a way that each released the same average number of vesicles, with differences found

175 only in the variabilities. In other words, vesicles were released individually in the "rate
176 code", but in packets of variable size in the "hybrid code". To simulate pure rate-coded
177 synaptic inputs case, we used a sinusoidal intensity function. The response of each
178 synapse to varying contrasts was simulated by increasing the amplitude of the sinusoid
179 between a total release of between 5 vesicles s^{-1} at input condition 1 (which we term 10%
180 contrast below) and 65 vesicles s^{-1} at input condition 10 (100% contrast; Fig. 2D and E).
181 for the rate case. Responses from multiple cells were assumed to come from the same
182 distribution, and as such we simulated the responses of multiple BCs onto a single cell by
183 summing the individual responses. To simulate hybrid coding we used a binomial
184 compounding function with a constant N=10, where the probability of release is a sinusoid
185 resulting in equivalent average release rates.
186
187 ***4. A conductance-based Leaky Integrate-and-Fire (LIF) model of the post-synaptic***
188 ***neuron.*** The vectors describing amplitudes and times of synaptic events were convolved
189 with a synaptic filter to estimate the time-course of the synaptic current S_{in} (Fig. 2C) which
190 was then passed through a Leaky Integrate-and-Fire model to calculate the change in
191 voltage and spike output. LIF models provide a computationally compact method for
192 exploring basic physiological parameters affecting spiking, including the number of
193 synaptic inputs, cell time-constant, threshold relative to resting potential and ion channel
194 conductances (Abbott, 1999; Burkitt, 2006a, b; Gerstner et al., 1997). Voltage is
195 governed by the differential equation:

196 $$I(t) = g_l\big(v_{rest} - v(t)\big) + \big(v_e - v(t)\big)g_e S_{in}(t) \quad \textbf{(1)}$$

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

200 $$S_i n(t) = \alpha * s(t) = \sum_{i=1}^{n} \alpha(t - t_i) \quad \textbf{(2)}$$

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

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

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

Parameter (symbol)	**Simulated Value/s**
Vesicles Required to Spike, k	10, 25, 50,100 vesicles
Leak Constant, τ_l	1, 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%

Table 1: Simulation Parameters. Parameters used for simulations, with ranges based upon experimentally recorded values (Destexhe and Rudolph, 2004; O'Brien et al., 2002; Pillow et al., 2005)

To compare how the MVR might shape the spike output of a postsynaptic cell across a range of physiological values, we systematically varied the parameters shown in Table 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 synapses (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

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

yields

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

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

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

Multiphoton imaging of glutamate release

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

Light emission was filtered through GFP filters (HQe525/50, Chroma Technology) before detection using GaAsP photomultipliers (H7422P-40, Hamamatsu). Photocurrents were passed to a current-to-voltage converter, after which signals from the objective and condenser were summed and digitized through ScanImage (v.3.6; Vidrio Technologies). For this work, all time-series were acquired as 128 pixel line scans imaged at 1 kHz and analysed in custom software written within IgorPro (Wavemetrics). Stimuli were generated

295 by an amber LED (peak wavelength of 590 nm) filtered through 590/10 bandpass filter
296 (Thorlabs) delivered through a light-guide placed near the fish's eye, with the mean
297 stimulus intensity of ~320 nW mm^{-2}. Detailed evidence that this approach can detect the
298 release of single vesicles from individual active zones has been presented previously
299 (James et al., 2019).
300
301
302 ## Results
303
304 **Synaptic coding of a sensory stimulus: gathering release statistics for an**
305 **empirically-driven model**
306 Our aim was to investigate how the statistical properties of vesicle release affected the
307 synaptic transfer of sensory information, comparing synapses employing MVR with the
308 conventional view of synapses operating as Poisson machines. The starting point was to
309 gather experimental measurements of release statistics from retinal bipolar cells
310 expressing the glutamate reporter iGluSnFR, and examples of such records are shown in
311 Fig. 1A. Both the rate and amplitude of glutamate transients varied widely during
312 sinusoidal modulations in light intensity.
313 We measured how changes in contrast altered three aspects of the vesicle code
314 highlighted in Fig 1: event rate R(Q), distribution of event amplitudes P(Qe), and variability
315 of event timing TJ(Q). R saturated at a lower contrasts than Qe, such that contrasts
316 greater than ~50% were often encoded entirely by the number of vesicles released within
317 an event (Fig. 1B and C). A potential role of MVR in determining the temporal accuracy of
318 spiking in the post-synaptic neuron is immediately suggested by the observation that
319 multiquantal events were more tightly phase-locked to the stimulus than vesicles released
320 individually (Fig. 1D-F). Uniquantal events had a temporal jitter of 21.2 ± 1.5 ms, while
321 events composed of 9 quanta had a temporal jitter of just 2.7 ± 0.6 ms, which is very
322 similar to the precision of spikes observed in RGCs of the salamander retina responding

323 to a high-contrast stimulus (Berry et al., 1997a; Uzzell and Chichilnisky, 2004). The
324 relation between TJ and Qe is shown in Fig. 1E and a comparison of the distribution of
325 times for uniquantal and multiquantal events is shown in Fig. 1F.

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

335

336 **The effects of multivesicular release in a "small" cell and "large" cell**

337 Bipolar cells transmit the visual signal to RGCs that vary in widely in the area of their
338 dendritic trees and, therefore, their receptive fields. In the fovea of primates, for instance,
339 midget ganglion cells are electrically compact and receive 25-50 synaptic inputs from
340 between one and three bipolar cells (Calkins et al., 1994; Kolb and Marshak, 2003). But
341 in the periphery of the retina, RGCs have larger dendritic fields and receive inputs from
342 hundreds of bipolar cells, with up to thousands of synapses (Jacoby et al., 2000). To
343 illustrate how MVR might impact on the spike output of these different post-synaptic
344 neurons we begin by comparing responses in a synthetic 'midget' RGC (τ_l = 5 ms with 50
345 synaptic inputs; Fig. 3A-E) and a larger 'peripheral' RGC (τ_l = 50 ms with 500 inputs; Fig.
346 3F-J) responding to a 5 Hz stimulus.

347

348

349

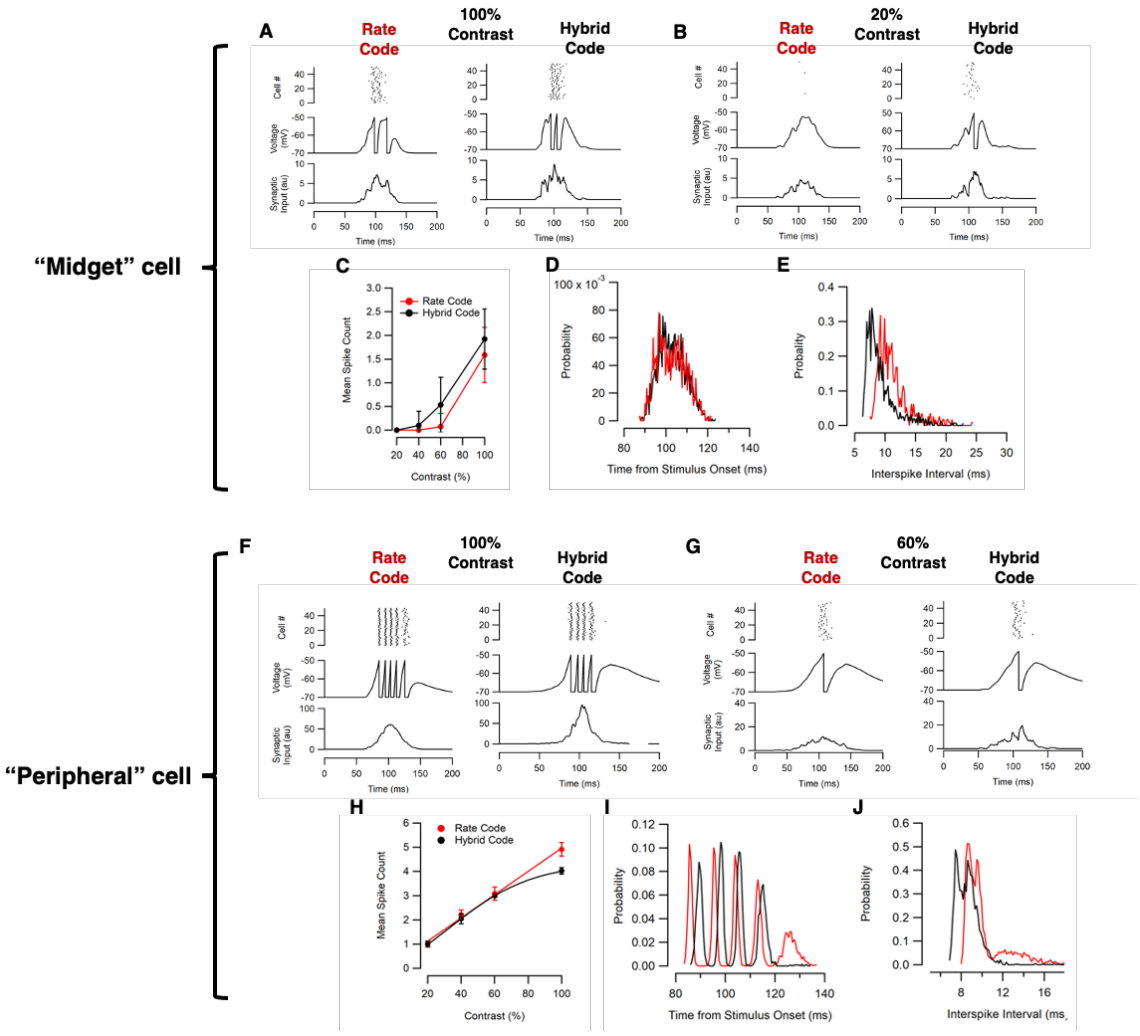


Figure 3. Effects of multivesicular release on spike generation in a "midget" cell and a "peripheral" cell.

A. Raster plots (top), example voltage traces (middle), and example synaptic input (bottom), for hybrid and rate code driven cell at 100% contrast at 5 Hz. One cycle of the sinusoidal stimulus is shown. This was a "small" cell simulated with: inputs = 50, k = 10, τ_l = 5 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV. The number of vesicles required to generate a spike (k) was 10. **B.** Same as **A**, but for 20% contrast. **C.** Mean spike count as a function of stimulus contrast (± standard deviation). Note the strongly supralinear relation. Synapses employing MVR begin to trigger reliable spiking at 40% contrast, while purely rate-coded synapses begin at >60%. **D.** Spike time distribution for rate (red) and hybrid (black) codes in response to a 5 Hz stimulus at 100% contrast as in **A**. **E.** Distribution of interspike intervals for 100% contrast, which were significantly shorter in the hybrid case (9.7 ± 2.5 ms) than the rate case (11.75 ± 2.62 ms, Paired Wilcoxon Rank Sum, $p < 0.05$). **F-J.** Corresponding to A-E, but for a "large" cell simulated with: inputs = 500, k = 100, τ_l = 50 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV, k = 100. The majority of rate code trials generate five spikes, while the majority of hybrid code trials generate only four spikes (F). The spike count in the cell driven by rate-coded synapses was proportional to stimulus contrast (slope of 0.03 spikes/percent contrast). Note the increased temporal precision of the rate code, with the first peak variance of 1.56 ± 0.23 ms compared to 6.95 ± 0.83 ms for the hybrid code.

369 In the midget cell, MVR increased the efficiency with which vesicles drove spikes as
370 well as the maximum frequency of spiking. Across all contrasts but the lowest, the cell
371 receiving inputs from hybrid synapses generated more spikes than the equivalent cell
372 receiving purely rate-coded inputs ($p < 0.013$, Wilcoxon Signed-Rank test; Fig. 3A-C). The
373 rate code failed to generate spikes reliably at all but the highest contrasts, while the hybrid
374 code drove responses starting at 40% contrast (Fig. 3C). The mean first spike latency
375 was not affected by coding strategy (Fig. 3D), but the distribution of interspike intervals at
376 100% contrast was significantly shorter in the hybrid case (9.7 ± 2.5 ms) than the rate
377 case (11.8 ± 2.6 ms, $p < 0.05$, Wilcoxon Signed-Rank. test; Fig. 3E).
378 Simulations of the peripheral RGC (Fig. 3F-J) demonstrated two key functional
379 differences compared with the small, both of which recapitulated experimental
380 observations in the retina of primates: spike rates were higher at any given contrast (Fig.
381 3H; Kaplan and Shapley, 1986) and the time to peak response was shorter (Fig. 3I; Sinha
382 et al., 2017). In this case, the synaptic coding strategy had little effect on the total number
383 of spikes generated by a stimulus (Fig. 3H) but it did affect temporal precision: the first
384 peak in the distribution of spike times had a standard deviation of 1.56 ± 0.23 ms with
385 rate-coded inputs compared to 6.95 ± 0.83 ms for synapses employing MVR (Fig. 3I).
386 Recapitulating observations in the midget cell, interspike intervals were shorter when the
387 larger neuron was driven by hybrid synapses (Fig. 3J).
388 These examples show how the impact of MVR on the spike code of the post-synaptic
389 neuron depends on its membrane time-constant: in a midget RGC with shorter τ_l MVR
390 tended to increase spike count but in a peripheral RGC with longer τ_l it tended to improve
391 the temporal precision of spikes. The maximum spike rate, however, was always higher
392 in the neuron driven by synapses exhibiting MVR.
393
394 **MVR increases spike count over a range of conditions**
395 RGCs with larger receptive fields and dendritic trees receive larger numbers of synaptic
396 inputs (N) and changes in the degree of convergence will be a key factor in determining

397 the efficiency of spike generation. A second key variable is the average strength of
398 synaptic responses, which in turn depends on the strength of the stimulus. In our
399 modelling framework synaptic strength is directly related to k, the minimum number of
400 vesicles that would need to be summed instantaneously to depolarise the neuron to
401 threshold. Although we systematically varied both k and N (Table 1) we found the ratio
402 k/N to be a useful summary metric that reduced dimensionality of the data while providing
403 insight into the effects of both synapse strength and the degree of convergence. For
404 instance, depolarization of the target neuron to threshold might occur by integrating
405 signals from just a few synapses with a high probability of response (i.e reliable synapses)
406 or more synapses with lower average response (i.e synapses of lower reliability).
407 Medium-sized RGCs in guinea-pigs receive about 1000-2000 inputs, and a brief burst of
408 excitation comprising 3-65 vesicles triggers a burst of 1-6 spikes (Freed, 2005). This
409 example therefore provides an upper limit of k ~10 and k/N is ~0.01. In contrast, a midget
410 RGC might receive just 10 inputs. If the same number of vesicles are required to trigger a
411 spike, k/N = 1, and even if just one vesicle is sufficient, k/N =0.1.

412 Increasing τ_l always increased the spike count over the 200 ms time window, and this
413 is shown in Fig. 4A (top) for an example neuron with 500 inputs requiring the summation
414 of 100 vescles to trigger a spike (k/N = 0.2). The absolute difference in spike count
415 between hybrid inputs and rate-coded inputs are shown below, demonstrating that for this
416 value of k/N, the hybrid code produces more spikes than the rate code when τ_l is less
417 than ~15 ms. The pattern of interaction between τ_l and k/N is shown for hybrid coded
418 inputs in Fig. 4B and the rate-coded inputs in Fig. 4C: in both cases, the lower the
419 average number of vesicles released per synapse to trigger a spike the larger the total
420 spike count for a cell of given time-constant. But the size of these changes were not the
421 same and the difference between the counts generated by hybrid- and rate-coded inputs
422 is shown in Fig. 4D. For a given τ_l, rate-coded inputs generate more spikes at lower k/N
423 and hybrid inputs were more efficient at higher k/N (i.e when spikes were generated by
424 activation of a larger proportion of inputs).

This comparison revealed a clear pattern: the rate code outperformed the hybrid code only when synapses were relatively unreliable. As the average number of vesicles released per synapse increased, a regime was entered in which the hybrid code generated spikes with greater efficiency (i.e more spikes per released vesicle), after which the coding strategy did not affect the spike count. In other words, where spike generation required more reliable synapses, reflecting a higher k/N ratio, it often became more effective if summation began *presynaptically* by the process of MVR.

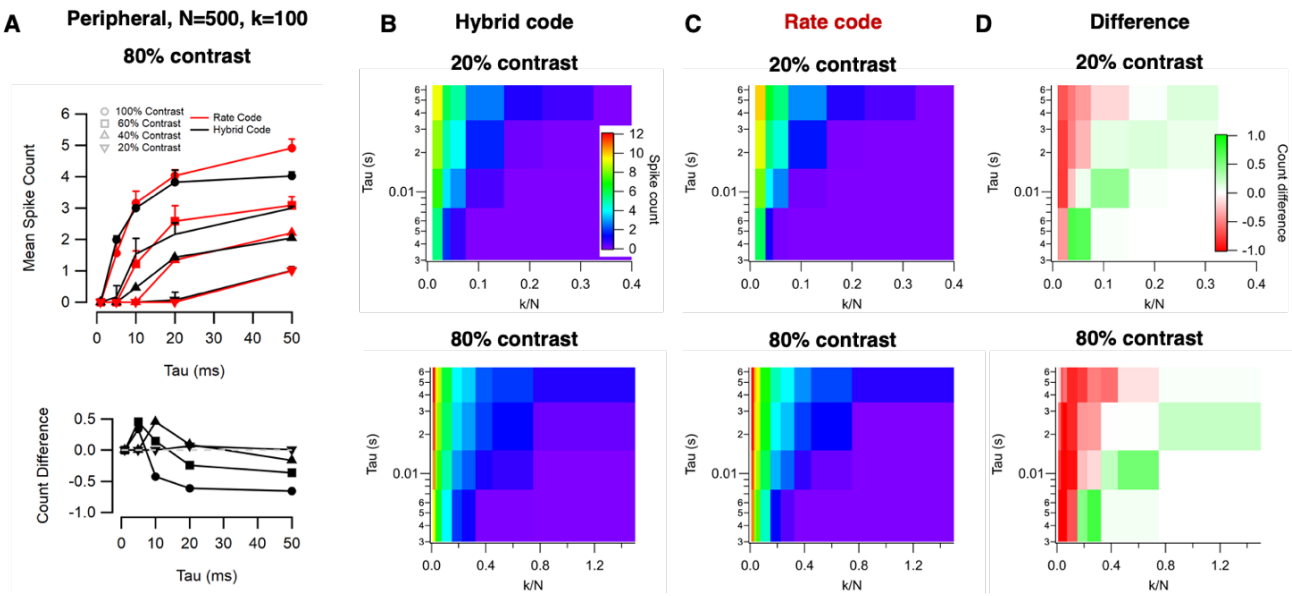


Figure 4. Hybrid coding increased spike count over a range of physiologically-relevant conditions
A. Top: Mean spike count as a function of leak time-constant $(\tau)_l$ for a cell receiving N=500 inputs and requiring k=100 vesicles to spike in response to a 5 Hz stimulus of varying contrast (other parameters 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 synapses transmitting the visual signal? Compared to other visual circuits, such as those

448 in the cortex, the retinal output is strikingly reproducible over repeated trials of a stimulus
449 and contrasts of 20-40% can trigger spikes in RGCs of a variety of species with almost
450 perfect reliability (i.e close to 100% of trials; Berry et al., 1997a; Gollisch and Meister,
451 2008)). Unsurprisingly, the reliablity of individual bipolar cell synapses was lower. At 20%
452 contrast a 5 Hz stimulus triggered a response in 31 ± 3% of stimulus cycles (N = 48
453 synapses), rising to 64 ± 4% of cycles at 60% contrast (N = 56 synapses). Experimental
454 measurements of the average number of vesicles released per stimulus at these contrasts
455 (where N is a single synapse) are shown in Fig. 1A-B: a stimulus of 20% contrast released
456 an average of 2.32 ± 0.09 (N = 48 synapses) vesicles per synapse, while 60% contrast
457 released 2.84 ± 0.14. Synapses transmitting the visual signal to the inner retina are
458 therefore operating over a range reliabilities where MVR provides a distinctive advantage
459 over a pure rate-code by increasing the number of spikes generated per released vesicle.
460 The consequences on the information carried by the spike response will be discussed
461 below (Fig. 7).
462 The flip side of the results in Fig. 4D is that a pure rate-code will be more effective
463 than MVR in situations where the neuron is driven by synapses where k/N less than ~0.15
464 i.e where convergence of a larger number of less reliable synapses is required for
465 depolarization to threshold. In mammals, this is the situation when the visual signal
466 reaches the primary visual cortex: excitatory synapses onto pyramidal cells have an
467 average release probability of ~0.17 vesicles per presynaptic spike (Allen and Stevens,
468 1994; Smetters and Zador, 1996) and these synapses operate with a high coefficient of
469 variation. The unreliability of cortical synapses is therefore a major source of noise that
470 impacts on information transfer (Branco and Staras, 2009; Moreno-Bote, 2014; Rusakov
471 et al., 2020; Allen and Stevens, 1994; Smetters and Zador, 1996).
472
473 **MVR reduces spike latency over a range of conditions**
474 Visual information is contained not only in the number of spikes generated by a stimulus
475 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).

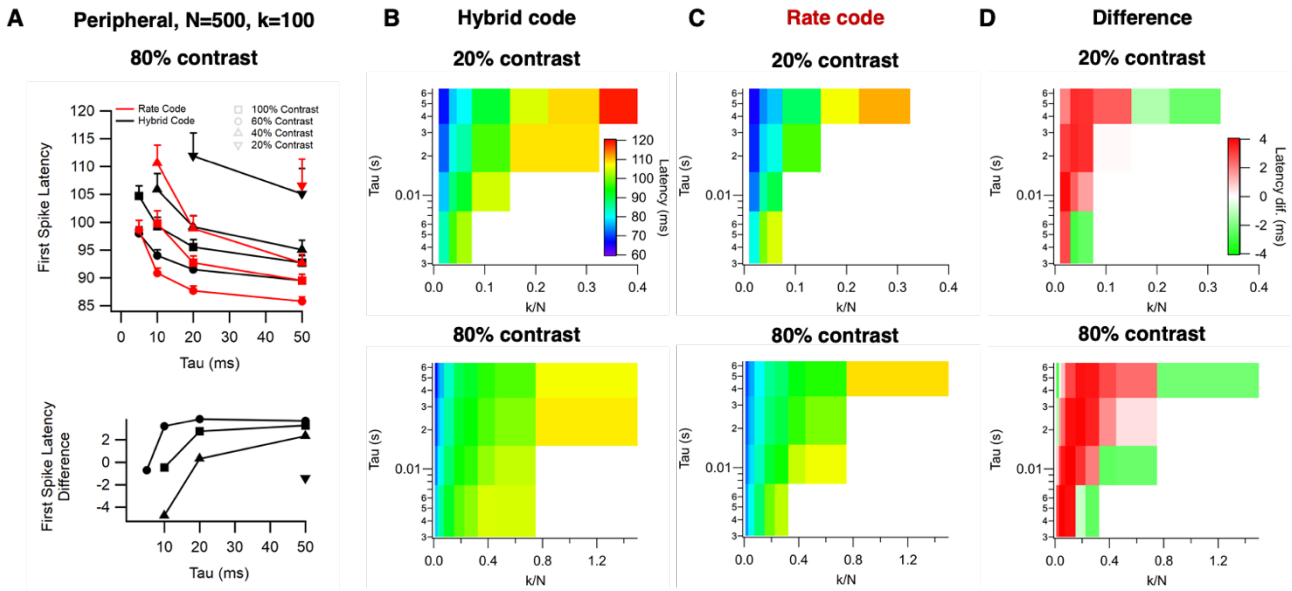


Figure 5. Hybrid coding reduced spike latency over a range of physiologically-relevant conditions
A. Top: Mean first spike latency as a function of leak time-constant $(\tau)_l$ for a cell receiving N=500 inputs and requiring k=100 vesicles to spike in response to a 5 Hz stimulus of varying contrast (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.

498 First spike latency depended on both the membrane time constant τ_1 and the
499 reliability of the synaptic inputs k/N (Fig. 5A-D). All but the very fastest time constants
500 produced shorter latency spikes utilizing a rate code compared to a hybrid code, with
501 delays being reduced by up to 4.5 ms. Cells requiring fewer vesicles to spike benefit from
502 the wider temporal window of vesicle release in univesicular compared to multivesicular
503 events (as seen in Fig. 1F). However, the more reliable the inputs the weaker this effect
504 (Fig. 5D). At an intermediate range of k/N hybrid inputs generated spikes with shorter
505 latencies and then at higher k/N there was no dependence on coding strategy.

506 The general picture is that the rate code generates spikes with shorter latency only in
507 larger neurons in which spiking is driven by larger numbers of less reliable synaptic inputs
508 - conditions more likely to occur in brain regions such as the cortex than the retina or other
509 circuits with more reliable synapses.

510

511 **MVR increases information transmission across a range of conditions**

512 Having established that MVR impacts both spike count and timing we asked how these
513 effects interact to alter the transmission of information (Borst and Theunissen, 1999).
514 The mutual information was calculated between a set of nine stimuli (contrasts of 20%,
515 60% and 100% at frequencies of 5 Hz, 10 Hz and 15 Hz) and each of three measures of
516 the response in the model neuron: spike count, spike time and the total information within
517 the spike sequence. Supplementary Figure 2 shows the probability distributions of quanta
518 per event, P(Qe), and temporal jitter functions, TJ(Q), while Supplementary Figures 3 and
519 4 show responses to the stimulus of the example midget and peripheral cells in Fig. 3. A
520 neuron that encodes a total of $\log_2(9) = 3.17$ bits of information over a 200 ms time
521 window is discriminating perfectly between the nine stimuli.

522

523

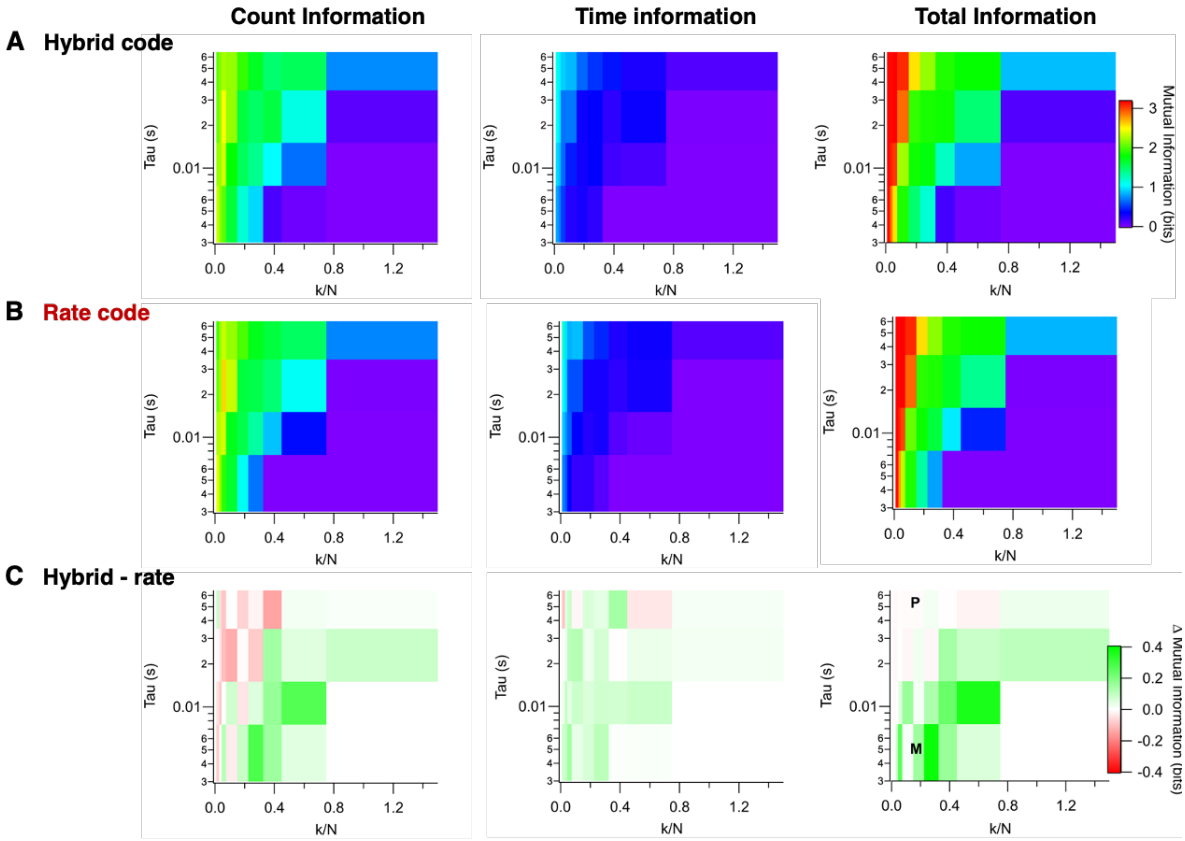


Figure 6. Hybrid coding increased information in the spike output over a range of physiologically-relevant conditions

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.

The heat plots in Fig. 6 demonstrate that spike count carried more of the total information than spike time for both the synaptic coding strategies. But the relative advantages of MVR and rate codes again depended on the time-constant of the target neuron and the reliability of the inputs driving spikes. In a cell of given time-constant, an

545 increase in the reliability of inputs caused the hybrid code to became more efficient at
546 generating spikes (Fig. 4) with shorter latency (Fig. 5) resulting in more information
547 transmission (Fig. 6). In the midget cell, for instance, the spike count over a 200 ms
548 period carried 0.07 bits more information when generated by hybrid inputs, while spike
549 time carried 0.08 bits more, and the total information difference was ~0.16 bits (equivalent
550 to an information rate of 0.82 bits s^{-1}). The largest difference in information rate in favour
551 of the hybrid code was 3.8 bits s^{-1} or 24% of the maximum possible information rate,
552 occurring when the post-synaptic cell required 25 vesicles to spike, had 250 inputs, and a
553 1 ms time constant.

554 The rate code carried marginally more of the total information only in lager cells in
555 which larger numbers of less reliable synapses converge. For the peripheral ("large")
556 RGC highlighted in Fig. 3, for instance, we calculated mutual information differences
557 (hybrid – rate) of -0.12 for spike count and 0.11 for spike time, yielding a total information
558 difference of -0.01 bits. The largest difference in information rate in favour of the rate
559 code was only 0.24 bits s^{-1}, occurring when only 10 vesicles were required to generate a
560 spike, with 50 inputs, and a time constant of 50 ms.

561 Together, the simulations in Figs. 2-6 indicate that synapses which generate MVR
562 can often trigger more spikes per released vesicle compared to a classical Poisson
563 synapse in which all vesicles are released independently. In cells with time-constants less
564 than 15 ms, the tendency for MVR to enhance the transmission of information becomes
565 greater as k/N falls below ~0.8. These conditions are physiologically relevant. In the
566 peripheral retina, for instance, alpha ganglion cells have membrane time-constants less
567 than 5 ms and ten or fewer quanta released from thousands of synapses are sufficient to
568 generate a spike (Freed and Sterling, 1988; O'Brien et al., 2002). More morphologically
569 compact neurons, such as midget ganglion cells in the fovea of primates, only receive
570 inputs from synapses from a single midget bipolar cell and are likely to have even shorter
571 time-constants (Sinha et al., 2017).

572 **MVR increases the efficiency of information transmission across a range of**
573 **conditions**

574 Neurons can transmit more information by spiking at higher rates but this comes at a
575 relatively large energetic cost (Laughlin, 2001). The release of vesicles and generation of
576 spikes is the major consumer of energy in the brain with one estimate being of the order of
577 ~24,000 ATP molecules consumed per bit of information transmitted (Attwell and
578 Laughlin, 2001; Harris et al., 2012). The efficiency with which vesicles drive the spike
579 code is therefore likely to have been an evolutionary constraint on the vesicle code itself
580 (Sterling and Laughlin, 2015).

581 To investigate the potential effects of MVR on the efficiency of information
582 transmission we normalized the measurements of total information summarized in Fig. 6A
583 and B as bits per vesicle and bits per spike, as shown in Fig. 7A and B. Considering the
584 amount of information per vesicle, the relative advantages of hybrid and rate codes can be
585 seen to follow a clear pattern: the more reliably inputs responded (i.e the higher k/N) the
586 broader the range of cell time-constants over which hybrid-coded inputs improved the
587 efficiency with which vesicles transferred information through spikes in the target neuron
588 (Fig. 7A right). The rate-code was only slightly advantageous in neurons acting as the
589 best temporal integrators receiving the least reliably responding inputs. The highest
590 vesicle efficiency using either synaptic code was ~0.044 bits per vesicle.

591

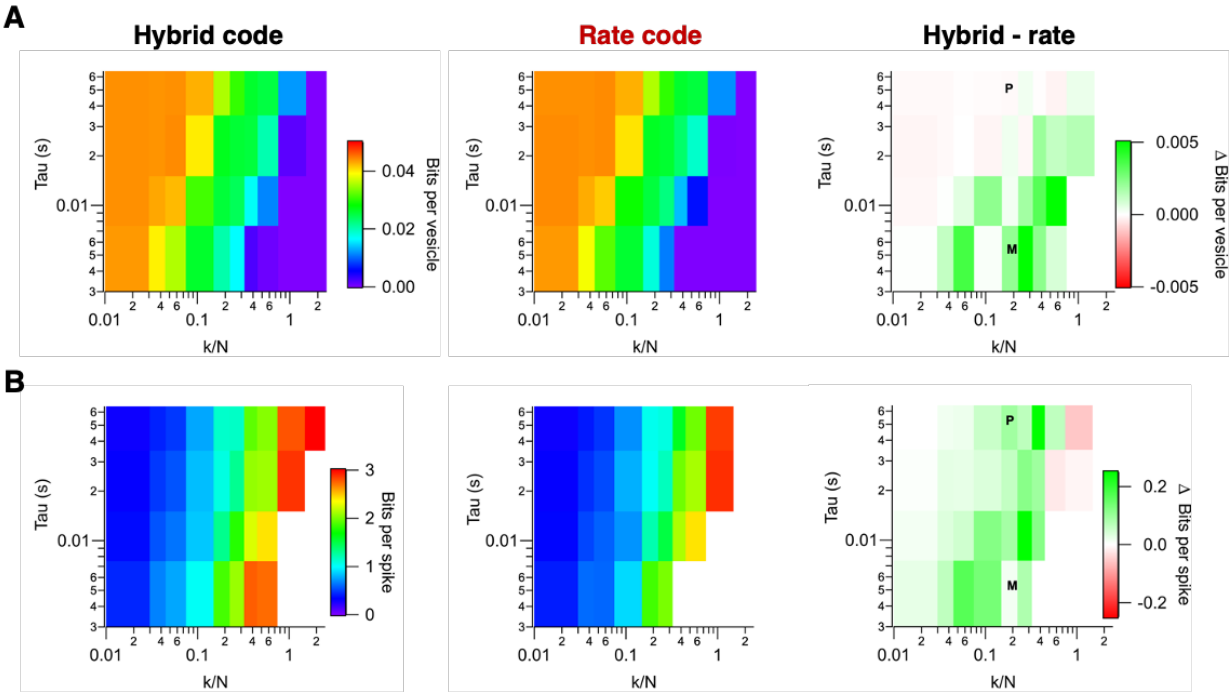


Figure 7. Hybrid coding increased the efficiency of information transmission over a range of conditions

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 hat we believe is the

612 most fundamental insight if this study – the potential of MVR to increase the efficiency with
613 which vesicles are used to transmit information.
614
615 **The first synapse in hearing**
616 MVR is also a prominent property of the ribbon synapse of auditory hair cells, where the
617 fusion of multiple vesicles can be synchronized on timescales as short as 100 μs (Fuchs
618 and Glowatzki, 2015; Glowatzki and Fuchs, 2002). This presents a special case because
619 there is no convergence onto the target, each Type I afferent fiber receiving sensing
620 glutamate released from just one ribbon synapse. This simplified situation allowed us to
621 use a modelling framework employing marked Poisson Processes (see Methods).
622 Experiments indicate that spikes in the afferent fiber can be triggered by just a few
623 vesicles and in some cases perhaps as a few as one (Grabner and Moser, 2018;
624 Rutherford et al., 2012). We therefore considered two broad situations:
625 i) k = 1. *Any* vesicular event, regardless of quantal content, is capable of generating a
626 spike.

627 ii) k > 1. Some (but not all) multivesicular events are capable of generating a spike from
628 rest.
629 The transmission of information was strongly dependent on both the number of
630 vesicles required to generate a spike, k, and the membrane time constant but again the
631 nature of this dependence differed for synapses employing a pure rate code compared to
632 those also employing MVR (Fig. 8A). Rate coding produced the maximum mutual
633 information in the simulations when k = 1, where the spike sequence in the afferent
634 reproduced the vesicle sequence leaving the synapse (Fig. 7A, red). In this situation a
635 multivesicular event is no more effective than a univesicular event so the hybrid code
636 produces fewer spikes per vesicle and reduces the efficiency of information transmission
637 (Fig. 8A, black).
638

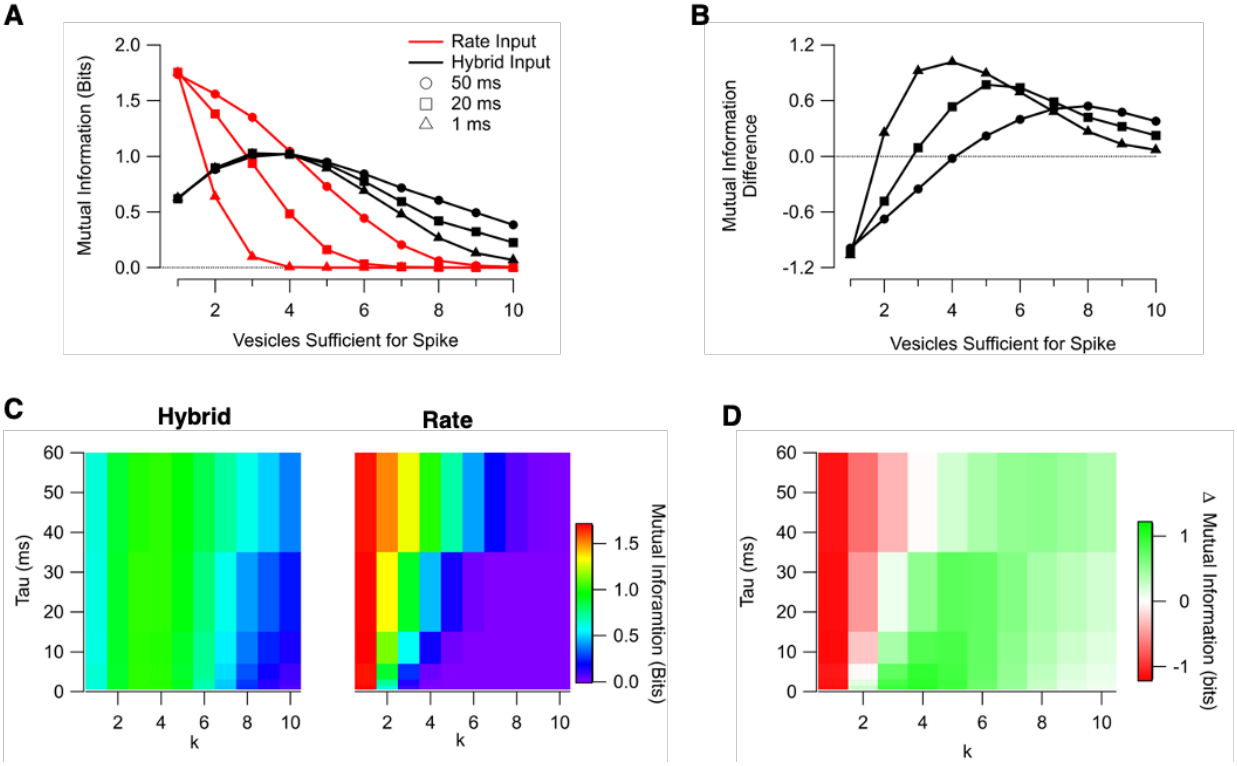


Figure 8. Spike Count Information for Amplitude (black) and Rate (red) inputs as a function of τ_I and k where N = 1. A. Mutual information for the rate input (dash) and amplitude input (solid) as a function of the vesicles sufficient to generate a spike. Values shown for τ_I of 50, 20, 10, and 1 ms.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

656 vesicles by MVR compensates for situations where the integration time of the post-
657 synaptic neuron is short and limits summation post-synaptically.

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

666

667 **Discussion**

668 The question motivating this study was "why MVR?" and the results indicate that the
669 answer is "because it can increase the efficiency with vesicles are used to transmit
670 information".

671 A number of studies have used leaky integrate-and-fire models to investigate the
672 relationship between vesicle release, spikes and information, usually treating synapses as
673 Poisson machines in which vesicles are released independently (de la Rocha et al., 2002;
674 Goldman, 2004; London et al., 2010; Zador, 1998). But this assumption does not hold at
675 many synapses exhibiting MVR (Auger et al., 1998; Glowatzki and Fuchs, 2002; Hays et
676 al., 2021; Singer et al., 2004) and here we have shown that over a range of conditions
677 mimicking transmission of visual signals to retinal ganglion cells of different sizes,
678 presynaptic integration of vesicles by MVR increased the efficiency of spike generation
679 compared to a simple rate-code in which all vesicles were released independently (Figs.
680 4, 6-7). In the special case of a single synaptic input, as occurs where auditory hair cells
681 drive afferent fibers, MVR increased the efficiency of information transfer whenever spike
682 generation required depolarization greater than that caused by a single vesicle (Fig. 8).
683 We identified two aspects of the process driving spikes that interact to determine the

684 relative efficiency of a pure rate-code compared to a hybrid code: the time-constant of the
685 target neuron and the reliability of the synaptic inputs. MVR tended to be advantageous
686 when i) the neurons time-constant was shorter, causing less post-synaptic summation of
687 synaptic potentials at physiological rates of vesicle release, and ii) individual synaptic
688 inputs were more reliable so that a lower degree of convergence was required to
689 depolarize the target to threshold for a given stimulus strength.
690

691 **Drawbacks of a pure rate code**

692 It is clear that neurons represent sensory stimuli, at least in part, using a simple rate-code,
693 where information is contained in the number of spikes over a given time-window (Rieke
694 and Warland, 1999). But it has long been recognized that a potential drawback of a
695 simple rate-code is the time-scale on which it can transmit information: the more variable
696 a neural response to a stimulus the longer the observation time required to detect it to a
697 given degree of certainty (Gerstner et al., 1997; Stein et al., 2005). At the ribbon synapse
698 of bipolar cells, for instance, the maximum sensitivity to changes in contrast occurs
699 around average release rate R = 20 vesicles s^{-1} and a 10% increase in contrast increases
700 release by a factor of a = 1.8 (James et al., 2019). If release events are Poisson
701 distributed, signal detection theory can be used to quantify the time-window over which
702 the response would need to be observed to differentiate it from ongoing release: a SNR of
703 2 would require counting an extra 14 vesicles over an observation time (Δt) of ~0.9 s
704 (equations 7-9 in Methods). But vision operates more rapidly. Our visual system can
705 process a scene in less than 150 ms (Thorpe et al., 1996) and neurons in the cortex of
706 monkeys can respond to images changing at intervals as short as 14 ms (Keysers et al.,
707 2001).

708 One feature of the retinal circuit that will shorten the time-scale of information
709 transmission using a rate-code is convergence of bipolar cell synapses onto RGCs. N
710 equivalent Poisson inputs allows a given SNR to be achieved with a time-window that
711 varies as $1/\sqrt{N}$ (Faisal et al., 2008; Gerstner et al., 1997; Rusakov et al., 2020).

712 Continuing with the example above and considering a time-window of 100 ms, the output
713 of 50 synapses is would provide a SNR of 2 in a time-window of 100 ms, but at the cost of
714 $N*(1-a)*R*\Delta t = 80$ extra vesicles. Although faster, this still seems costly.
715
716 **Encoding discrete visual events**
717 The discussion of potential problems with a pure rate code is closely linked to the idea
718 that the firing rate of a neuron or release rate of a synapse encodes the sensory variable
719 continuously. In fact, if a varying stimulus is applied, most RGCs fire in bursts from rest,
720 driven by fusillades of vesicles released from bipolar cells (Berry et al., 1997b; Freed,
721 2005; Keat et al., 2001). The information that RGCs convey is therefore represented not
722 just by the spike count but also by the precise timing of spikes (Gutig et al., 2013; Meister
723 et al., 1995; Victor, 2000) and it has been shown that such a temporal coding strategy can
724 significantly improve the speed of information transmission (Butts et al., 2007; Gollisch
725 and Meister, 2008). Focusing on an individual active zone using iGluSnFr reveals that
726 these also signal visual events in bursts, including MVR events of varying amplitude (Fig.
727 1A and C).
728 In considering different synaptic coding strategies it might therefore be more helpful to
729 focus on the representation of temporally discrete events rather than the value of a
730 continuous variable such as contrast. It has been suggested that the onset of a burst in
731 an RGC provides information about when a visual event occurs, while the number of
732 spikes within the burst represent properties such as its magnitude (Keat et al., 2001). The
733 temporal jitter in thr first spike within a burst can be as low as a few milliseconds (Berry et
734 al., 1997a; Freed, 2005) and it is notable that larger MVR events from bipolar cells
735 achieve a similar temporal precision (Fig. 1E and F).
736 This comparison of the vesicle code in bipolar cells and the spike code in post-
737 synaptic RGCs highlights one of the potential advantages of representing a stimulus using
738 synaptic symbols of different amplitude: an unexpected symbol immediately imparts new
739 information and larger MVR events that are more likely to trigger a spike are indeed rarer

740 (Fig. 1C) (James et al., 2019). Our analysis of mutual information was more concerned
741 with the "what" of the stimulus (contrast), and spike count carried more of the total
742 information than spike time for either of the synaptic coding strategies (Fig. 7). This
743 feature of the results might, however, represent limitations of the LIF model, in which we
744 only considered excitation. The precise timing of spikes in RGCs is also dependent on
745 inhibitory inputs, especially feedforward inhibition that increases the temporal precision of
746 spikes and, crucially, also causes spikes to be fired in bursts (Johnston and Lagnado,
747 2015; Murphy-Baum and Taylor, 2018). Network models incorporating a wider range of
748 biophysical parameters will help us understand how excitation and inhibition interact to
749 determine the temporal precision of the spike code leaving the retina.
750
751 **Encoding auditory events**
752 The spike code carrying auditory information in cochlear afferents is determined only by
753 excitatory inputs from hair cells and is therefore likely to reflect the temporal precision of
754 ribbon synapses more directly than the spike code leaving the retina. The synaptic output
755 from hair cells has been measured *in vitro* with electrophysiological precision where it is
756 found that larger MVR events are more tightly phase-locked to a sinusoidal voltage-clamp
757 stimulus (Li et al., 2014). Notably, MVR events in hair cells can be as large as ~20 vesicle
758 equivalents when triggered using electrophysiology (Glowatzki and Fuchs, 2002) but it is
759 unclear if these can also be triggered by sound. Here we used the statistics of MVR
760 measured in bipolar cells *in vivo*, where the largest events evoked by the highest
761 contrasts are composed of 9-11 vesicles (James et al., 2019). A better understanding of
762 how the ribbon synapse of hair cells drive the spke code in auditory fibers will require
763 experimental measurements of the statistics of vesicle release in response to sounds of
764 different intensities and frequencies in hair cells with different characteristic frequencies.
765
766

767 **Future directions**

768 To put the statistics of vesicle release into a functional context we also need an
769 understanding of the spike code that it drives. A key next step will be to obtain direct
770 experimental measurements of how effectively MVR events of different size trigger spikes.
771 This will likely vary according to the time-constant of the RGC and the number and
772 strength of it's inputs, which in turn will vary according to the functional class it belongs
773 and its location in the retina. In the peripheral retina, for instance, alpha ganglion cells
774 have membrane time-constants less than 5 ms and on average ten quanta are released
775 from thousands of synapses for each spike generated (Freed, 2005). More
776 morphologically compact neurons, such as midget ganglion cells in the fovea of primates,
777 only receive inputs from a single midget bipolar cell and are likely to have even shorter
778 time-constants (Sinha et al., 2017). The simpler organization of the first stage in hearing,
779 where spikes in afferents are driven by a single synapse, provides a beautiful context in
780 which to investigate how the vesicle code determines the spike code transmitting sensory
781 information.

782

783

784 **References**

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

787 Allen, C., and Stevens, C.F. (1994). An evaluation of causes for unreliability of synaptic
788 transmission. Proceedings of the National Academy of Sciences *91*, 10380-10383.

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

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

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

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

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

799 Borst, A., and Theunissen, F.E. (1999). Information theory and neural coding. Nat Neurosci *2*, 947-
800 957.

Branco, T., and Staras, K. (2009). The probability of neurotransmitter release: variability and feedback control at single synapses. Nat Rev Neurosci *10*, 373-383.

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.

Carleton, A., Accolla, R., and Simon, S.A. (2010). Coding in the mammalian gustatory system. Trends Neurosci *33*, 326-334.

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

de la Rocha, J., Nevado, A., and Parga, N. (2002). Information transmission by stochastic synapses with short-term depression: neural coding and optimization. Neurocomputing *44*, 85-90.

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.

Goldman, M.S. (2004). Enhancement of information transmission efficiency by synaptic failures. Neural computation *16*, 1137-1162.

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.

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.

Johnson, D.H. (1996). Point process models of single-neuron discharges. J Comput Neurosci *3*, 275-299.

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

Kaplan, E., and Shapley, R.M. (1986). The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. Proc Natl Acad Sci U S A *83*, 2755-2757.

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.

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.

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.

London, M., Roth, A., Beeren, L., Hausser, M., and Latham, P.E. (2010). Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. Nature *466*, 123-127.

Manwani, A., and Koch, C. (2001). Detecting and estimating signals over noisy and unreliable synapses: information-theoretic analysis. Neural Comput *13*, 1-33.

Moreno-Bote, R. (2014). Poisson-like spiking in circuits with probabilistic synapses. PLoS computational biology *10*, e1003522-e1003522.

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.

Neves, G., and Lagnado, L. (1999). The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells. J Physiol *515 (Pt 1)*, 181-202.

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.

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.

Quian Quiroga, R., and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci *10*, 173-185.

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.

Schwartz, O., Pillow, J.W., Rust, N.C., and Simoncelli, E.P. (2006). Spike-triggered neural characterization. J Vis *6*, 484-507.

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.

Smetters, D.K., and Zador, A. (1996). Synaptic transmission: Noisy synapses and noisy neurons. Current Biology *6*, 1217-1218.

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.

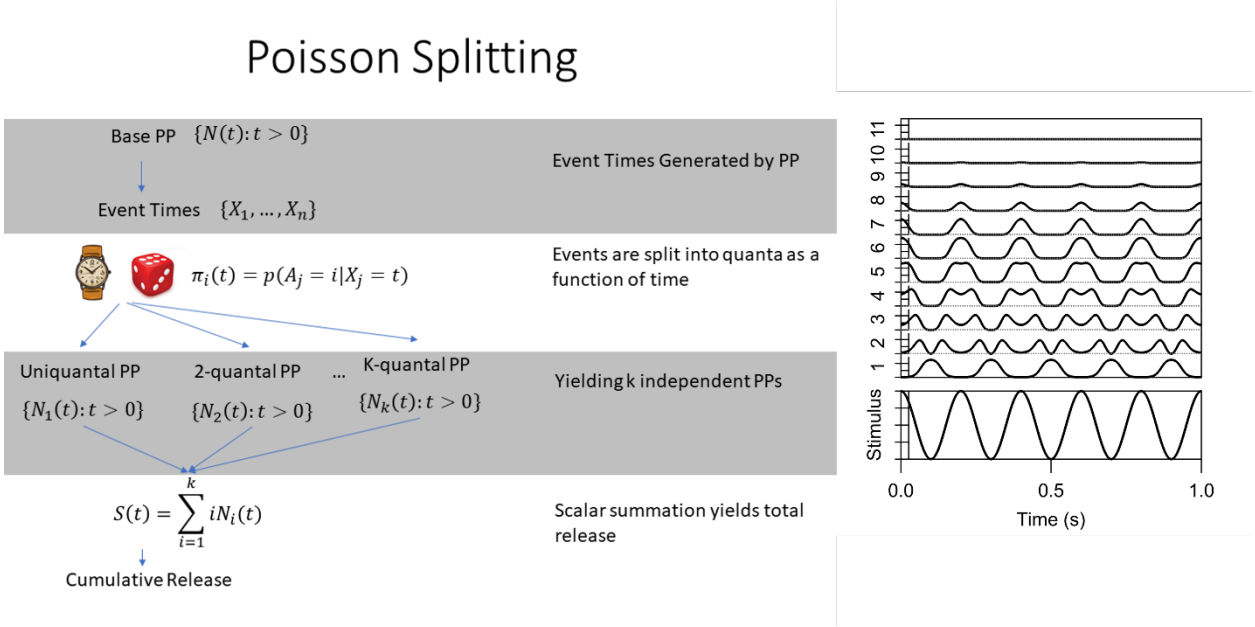
Uzzell, V.J., and Chichilnisky, E.J. (2004). Precision of spike trains in primate retinal ganglion cells. J Neurophysiol *92*, 780-789.

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

Zador, A. (1998). Impact of synaptic unreliability on the information transmitted by spiking neurons. J Neurophysiol *79*, 1219-1229.

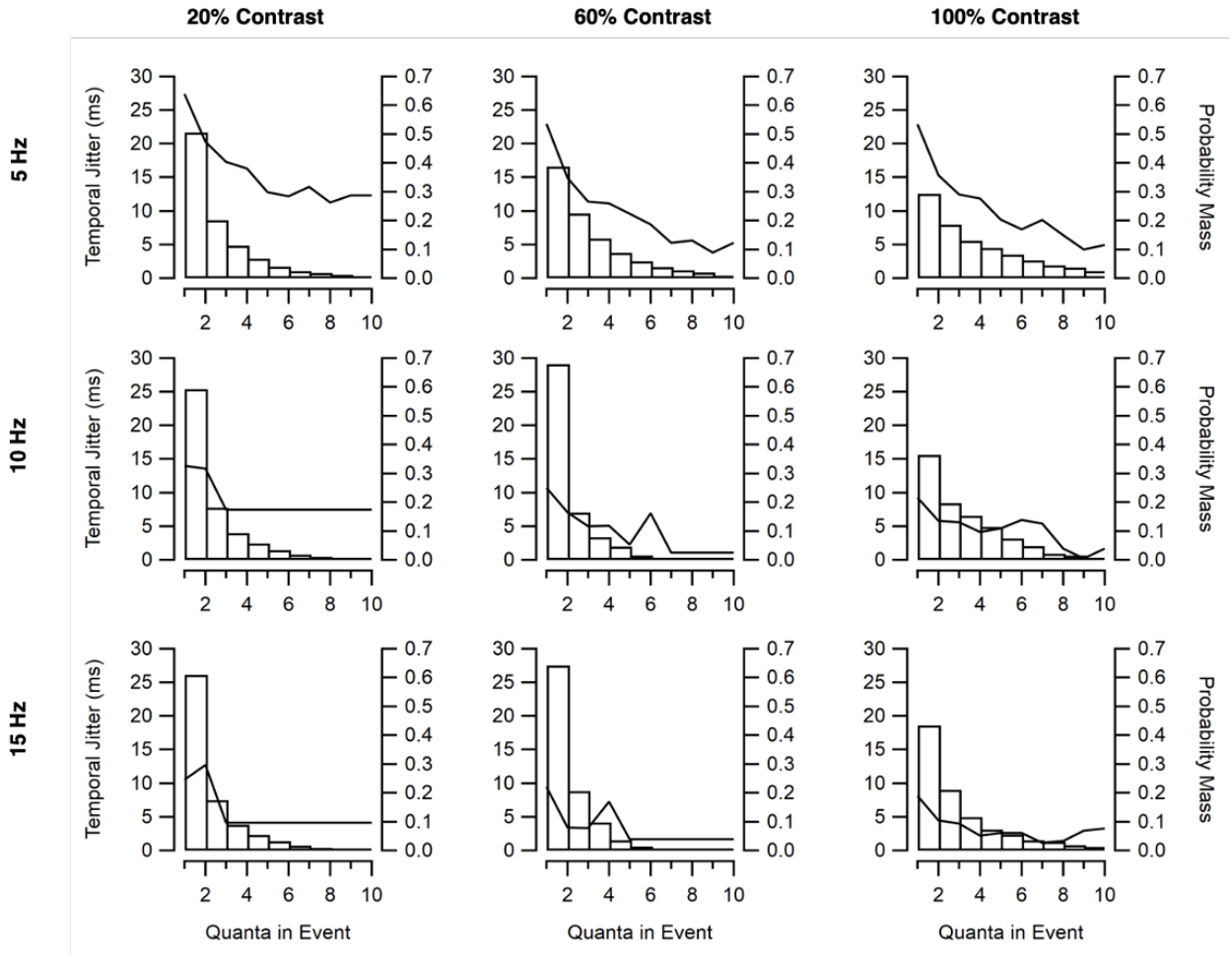
Supplementary Information

Supplementary Figure 1



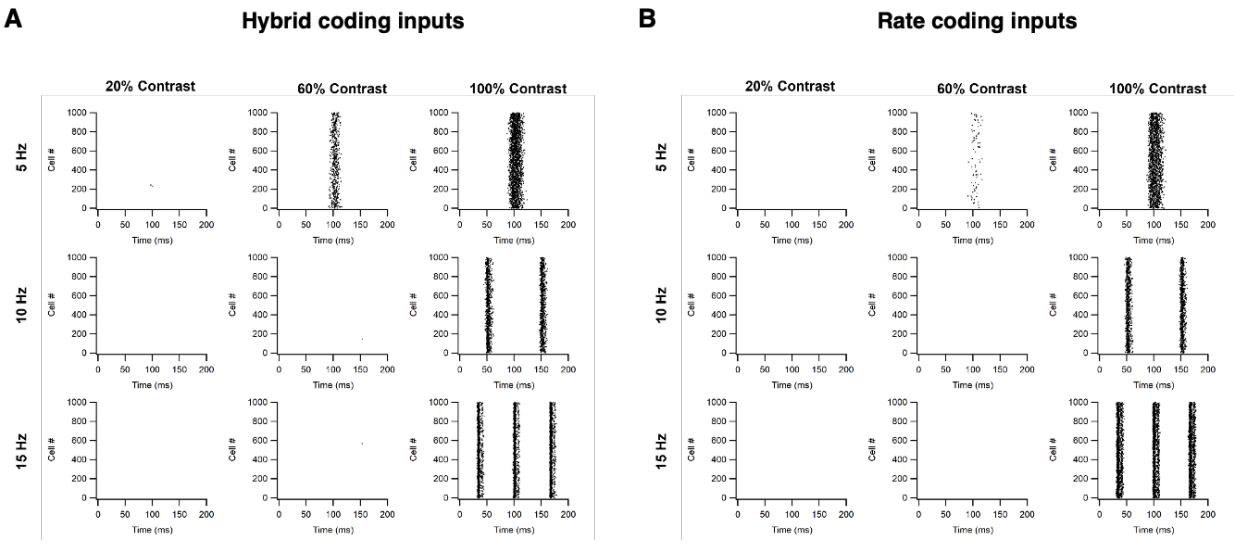
Supplementary Figure 1. Graphical description of Poisson Splitting. Left: a Poisson Process (PPs) is split dependent upon time, yielding independent PPs, where the total release can then be summed. Right: Time dependent splitting function used for the simulations.

Supplementary Figure 2



Supplementary Figure 2. Temporal Jitter functions (TJ(Q)) and Probability Mass Functions (P(Qe)) for all 9 stimuli used for the computation of mutual information. Note that increasing contrast increases shifts the distribution of quanta per event (P(Qe)) towards higher values and decreases the temporal jitter, while increasing the stimulus frequency shifts the distribution of quanta per event towards lower values and decreases temporal jitter.

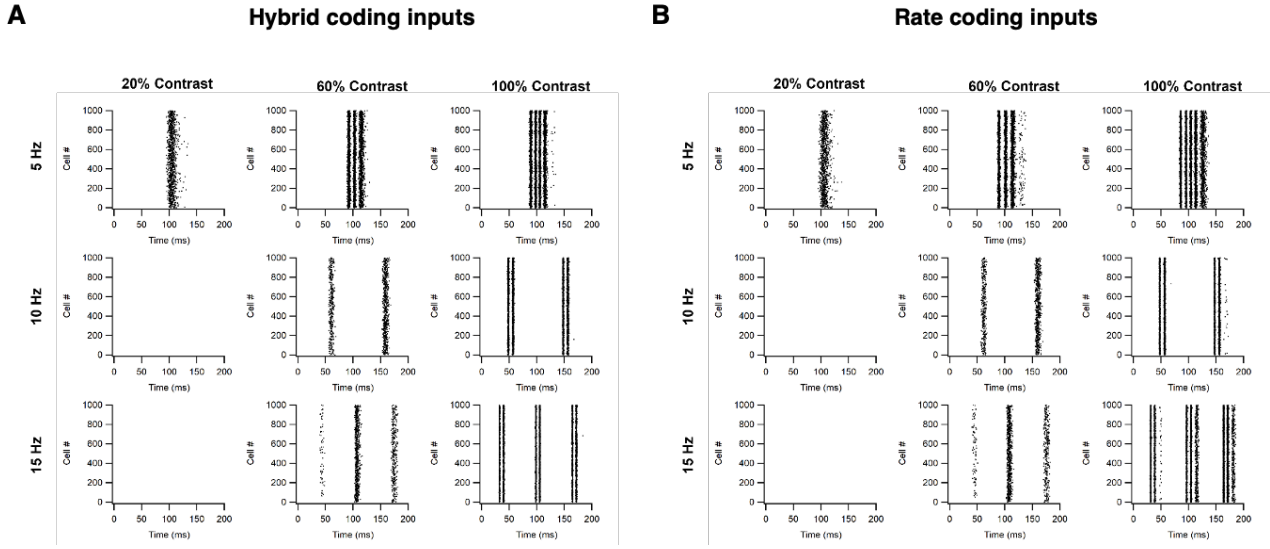
Supplementary Figure 3



Supplementary Figure 3. Responses of the example "midget" cell to the stimulus set used in information theory calculations.

A. Raster plots for hybrid coded synaptic inputs at the frequency and contrast indicated. This was a "midget" cell simulated with: inputs = 50, k = 10, τ_l = 5 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV. The number of vesicles required to generate a spike (k) was 10. These are the same parameters as the example cell featured in Fig. 3A-E. **B.** Same as **A**, but for purely rate-coded synaptic inputs in which all vesicles were released independently.

Supplementary Figure 4



Supplementary Figure 4. Responses of the example "peripheral" cell to the stimulus set used in information theory calculations.

A. Raster plots for hybrid coded synaptic inputs at the frequency and contrast indicated. This was a "peripheral" cell simulated with: inputs = 500, k = 100, τ_l = 50 ms, v_e = 0, θ = -50 mV, v_{rest} = -70 mV, k = 100. These are the same parameters as the example cell featured in Fig. 3F-J. **B.** Same as **A**, but for purely rate-coded synaptic inputs in which all vesicles were released independently.