ZETA: a parameter-free statistical test for neuronal responsiveness

Authors

Jorrit S. Montijn[1], Koen Seignette[1], Marcus H. Howlett[1], J. Leonie Cazemier[1], Maarten Kamermans[1], Christiaan N. Levelt[1], J. Alexander Heimel[1]

Affiliations

1 Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, 1105 BA, the Netherlands

Contact Information

Correspondence and requests for materials should be addressed to J.S.M. (jsmontijn@gmail.com) or J.A.H. (heimel@nin.knaw.nl).

Abstract

Neurophysiological studies depend on a reliable quantification of whether and when a neuron responds to stimulation. Simple methods to determine responsiveness require arbitrary parameter choices, such as binning size, while more advanced model-based methods require fitting and hyperparameter tuning. These parameter choices can change the results, which invites bad statistical practice and reduces the replicability. New recording techniques that yield increasingly large numbers of cells would benefit from a test for cell-inclusion that requires no manual curation. Here, we present the parameter-free ZETA-test, which outperforms t-tests, ANOVAs, and renewal-process-based methods by including more cells at a similar false-positive rate. We show that our procedure works across brain regions and recording techniques, including calcium imaging and Neuropixels data. Furthermore, in illustration of the method, we show in mouse visual cortex that 1) visuomotor-mismatch and spatial location are encoded by different neuronal subpopulations; and 2) optogenetic stimulation of VIP cells leads to early inhibition and subsequent disinhibition.

Keywords

Neural data analysis, Statistics, Neurons, Responsiveness, Response Latency, Electrophysiology, Calcium Imaging, Open Source, Visual Cortex, VIP disinhibition

Introduction

Many neuroscience studies rely on the analysis and visualization of neuronal spiking signals. Classical studies used manual curation during experiments to select cells for analysis (Hubel & Wiesel, 1959; Mountcastle, 1957), but this method cannot provide a statistically unbiased sample. Moreover, such manual curation is unsuitable for state-of-the-art large-scale recording techniques, such as Neuropixels and high-density multi-electrode arrays (Bartolo et al., 2020; Jun et al., 2017; Semedo et al., 2019; Steinmetz et al., 2019).

Despite the widespread application of neuronal responsiveness analyses, neuroscience currently lacks a standard practice for determining whether a neuron is responsive to an experimental stimulus or treatment (Mesa et al., 2021). Common approaches, such as comparing a neuron's average spike rate during the presence and absence of a stimulus, can only detect mean-rate modulated cells (Mazurek et al., 2014; Ringach et al., 2002). On the other hand, approaches such as computing a peri-stimulus time histogram (PSTH) and applying an ANOVA, require the *a priori* selection of an arbitrary binning window size (Palm et al., 1988). Choosing the wrong bin size reduces the test's sensitivity, while optimizing from a range of window sizes creates a multiple-comparison problem This lowers the approach's statistical power when corrections are applied, or can even lead to (unintentional) "p-hacking" if the results are not corrected (Head et al., 2015). Finally, while (point-process) model-based approaches can circumvent many of the above problems (Kass et al., 2014), they still require the *a priori* selection, or tuning, of hyperparameters specific to the statistical properties of classes of cells, or even individual neurons. Many model-based approaches are therefore not well suited to an unsupervised analysis of large-scale data recorded with state-of-the-art techniques.

To solve these problems, we developed a method that detects whether a cell is responsive to stimulation in a statistically robust way and avoids binning and parameter selection altogether. This method, which we call ZETA (Zenith of Event-based Time-locked

Anomalies), either outperformed or matched that of t-tests, ANOVAs and point-process-based methods in all conditions tested. Building upon this framework, we also present a procedure to visualize instantaneous spiking rates without the need for binning, and show how this can be used to estimate peak-activity latencies with sub-millisecond accuracy. We apply these methods to transient-detected two-photon calcium imaging data from the visual cortex of mice traversing a virtual linear track and find that visuomotor mismatch signals and spatial location are encoded by different V1 neuronal subpopulations. Finally, we apply our approach to Allen Brain Institute Neuropixels data and show that optogenetic stimulation of VIP-expressing cells in mouse visual cortex has a separable early inhibitory and late disinhibitory effect on the local neural circuit. We anticipate that the ZETA-test will be a useful resource for a wide range of applications across various disciplines.

Results

ZETA: Zenith of Event-based Time-locked Anomalies

A common procedure in pre-processing neural data is removing cells that are not responsive to an experimental stimulus. Many experimenters determine the "stimulus responsiveness" of a cell by comparing its average spiking rate during the presentation and absence of a stimulus (see fig. 1A-C for an example V1 cell). This procedure will therefore remove neurons that show no response, but has the risk of also removing neurons that show a strong, but complex time-locked response to stimuli. To remedy this shortcoming, we developed a binning-free method for determining whether a neuron shows any time-locked modulation. We call this statistical test ZETA for Zenith of Event-based Time-locked Anomalies (fig. 1D-F). It represents whether a neuron's spike train could be observed by chance, if it were not responding to an experimenter's event of interest: e.g., the presentation of a visual stimulus, the onset of optogenetic stimulation, or a self-generated variable, such as an animal's location on a track.

ZETA is calculated on a single cell by performing the following steps. First, we align all spikes to stimulus onsets, i.e. building a raster plot (fig. 1A). Pooling all spikes across trials, we obtain a single vector of spike times relative to stimulus onset, and calculate the cumulative distribution as a function of time (fig. 1D). The deviation of this curve from a linear baseline represents whether the neuron has a higher or lower spiking density relative to a non-modulated spiking rate (fig. 1E, blue curve). We compare this pattern to the likelihood of observing it by chance by running multiple bootstraps by jittering stimulus-onset times to generate a null hypothesis distribution (fig. 1E, grey curves). After scaling the experimentally observed curve to the variation in the null hypothesis distribution, we use it to obtain a p-value corresponding to the Zenith of Event-based Time-locked Anomalies. Low ZETA-test p-values indicate that the neuron's firing pattern is statistically unlikely to be observed if the neuron is not modulated by the event of interest.

The ZETA-test bears similarities to a mean-subtracted Kolmogorov-Smirnov test applied to a renewal process model (see methods). In the method section "ZETA and renewal-process models" we show how the ZETA-test gains robustness to violations of the assumptions underlying renewal processes and outperforms alternative approaches. In short, the ZETA-test's main difference from other approaches used to infer a neuron's stimulus responsiveness, is that our test makes no *a priori* assumptions about the underlying distribution of temporal modulations and is binning-free. It can therefore detect both long-timescale changes in mean firing rate, as well as short-timescale stimulus-locked bursts or lapses of activity at any point in time relative to stimulus onset.

Benchmarking the ZETA-test

To investigate whether the ZETA-test includes more cells recorded in mouse visual cortex in response to a drifting grating, while still retaining a 5% false-positive rate at a significance level

of $\alpha = 0.05$, we used a benchmarking test comparing onset-jittered and non-jittered data (fig. 2). In the non-jittered case, we compared the inclusion rate of the ZETA-test, as described above, to that of a mean-rate t-test. For the t-test, we calculated the average spiking rate of a cell during stimulation (0 s – 1 s after onset) and after stimulation (1 s – 1.5 s), and performed a paired t-test over trial repetitions (fig. 1C). This showed that cells included with a t-test were almost exclusively a subset of the cells detected with ZETA (see fig. 2B for all V1 cells recorded with Neuropixels). In other words, if a cell is detected as being visually responsive with a t-test, it is almost guaranteed to also be detected by the ZETA-test. In addition to these cells, the ZETA-test also includes cells that were not registered by a t-test. Although many varieties exist, fig. 2C shows an example cell that is detected by both t-tests and ZETA (top, sustained change in firing rate), and an example only detected by ZETA (bottom, balanced on/off peaks).

Next, we ran the same tests again, but now on data where we randomly jittered the stimulus onset times between $-\tau$ and $+\tau$, where τ is median onset-to-onset duration. This procedure preserves the properties of a neuron's spike train, but removes locking of the responses to the stimulus (fig. 2D). If the ZETA-test simply always gave low p-values, then this would result in a high false-positive rate as many neurons would still be included. In contrast, the false-positive rate of the ZETA-test was generally low, and consistent with that expected for a significance level of $\alpha=0.05$ (fig. 2E-F).

Sensitivity of ZETA is superior to mean-rate t-tests

We performed this benchmark for single-cell activity obtained from n=12 combinations of various visual regions (V1, AM, PM, LGN, SC, LP, NOT, APN, Retina) using multiple techniques (Neuropixels, n=8; Neuronexus, n=2; GCaMP6, n=1; pMEA, n=1), in response to light flashes (retina, n=1) or drifting gratings (all others, n=11). Under all conditions, the inclusion rate using ZETA-tests was higher than using t-tests: at a significance level of $\alpha=0.05$,

the inclusion rate for the ZETA-tests was 79% and for mean-rate t-tests was 64%; t-test of ZETA vs mean-rate t-test inclusion rates: n=12 data sets, $p=2.8 \times 10^{-7}$ (fig. 1G, fig. 3, fig. 3 – Figure Supplement 1). This means that the ZETA-test includes 42% of the cells that were not included by a t-test. A significance level of $\alpha=0.05$ is rather arbitrary, so we also performed a receiver operating characteristic (ROC) analysis, where we investigated the number of inclusions as a function of the number of false positives (fig. 3). The ROC's summary statistic is the area under the curve (AUC); 1 being a perfect discriminator. Again, we found that the ZETA-test showed a higher statistical sensitivity (ZETA-test, mean AUC=0.914) than a mean-rate t-test (t-test, mean AUC=0.843), and that this difference was statistically significant (paired t-test, n = 12, $p = 4.9 \times 10^{-6}$)

We also benchmarked various versions of tests derived from the theoretical framework of renewal and Poisson process models. None of these models reached the statistical power and computational efficiency of the ZETA-test (Figure 3 - Supplements 2,3), but they do provide an attractive mathematical connection to a more widely studied class of models. We have therefore provided more information on the mathematical relationship between this class of models and the ZETA-test in the method section.

The percentage of visually responsive cells detected by the ZETA-test is higher than typically reported. For example, using gratings that only differed by orientation, we found 93.3% of all V1 cells to be visually-modulated. Even more striking was that the lower bound of the binomial 95%-confidence interval was at 89.7%. This lower-bound is higher than the responsiveness previously reported in many studies, including our own (Montijn, Meijer, et al., 2016; Niell & Stryker, 2008; Shuler & Bear, 2006; Steinmetz et al., 2019), exemplifying the advantage of the ZETA-test.

The data used so far for benchmarking contains more stimulus repetitions than commonly used in neuroscience. This raises the possibility that the ZETA-test is only

advantageous when large numbers of trials are used. To test this, we randomly subsampled the number of trials included in the analysis, and repeated our benchmark. As the results for V1 in Figure 3 - Supplement 4A show, the ZETA-test consistently included more than the t-test, regardless of the number of trials.

We hypothesized that the t-test's worse performance might result from pooling the responses to different orientations into one group. Therefore, we repeated the t-test's benchmark after first splitting the trials into 24 groups corresponding to the 24 directions we presented. A neuron was included if the spiking rate during stimulus presentation was significantly different from its pre-stimulus baseline rate in any group, after applying a Bonferroni correction. However, this procedure reduced the t-tests' performance (Figure 3 - Supplement 4A, middle panel).

We noticed that the t-test's false positive rate was rather low after Bonferroni corrections. To test whether we over-corrected the t-test, we removed the multiple-comparison correction (Figure 3 - Supplement 4A, right-hand panel). In this case, the t-test false positive rate increased to >50%, while its inclusion rate (89.9%) remained lower than that of the ZETA-test (95.0%). Finally, we investigated whether the t-test's performance was hampered by including the immediate off-response after stimulus offset in the baseline period. We reran the above analyses, but now limited the baseline to the 300 ms preceding the stimulus onset. As can be seen in Figure 3 - Supplement 4B, this did not improve the t-test's performance. In summary, the t-test is at its most sensitive when using the full 500-ms epoch in-between stimulus presentations as baseline period and when pooling data across all orientations.

The sensitivity of ZETA is superior to an ANOVA with an optimal bin width

Mean-rate t-tests are common in neuroscientific analysis, but it could be argued that this is somewhat of a strawman to use as baseline performance. An alternative is to construct a peri-

stimulus time histogram (PSTH) and run a one-way ANOVA across bins to test a neuron's responsiveness to a particular stimulus. However, because this requires picking a bin width, this can lead to arbitrary choices based on the experimenter's visual inspection of the data, which might increase false positive rates. A better solution is to use one of the various methods to estimate optimal widths for binning (Freedman & Diaconis, 1981; Scott, 2009; Shimazaki & Shinomoto, 2007). We therefore calculated the optimal bin width using the Shimazaki & Shinomoto method, which was specifically designed for building a PSTH, and repeated the benchmark described above, but now testing the responsiveness of neurons using an ANOVA (see Methods). This ANOVA procedure performed markedly better than a mean-rate t-test (fig. 1G, fig. 3). However, we found that it still showed an inclusion rate (at α=0.05) that was lower than using the ZETA-test (ANOVA mean inclusion rate = 71%; ZETA-test inclusion = 79%, paired t-test, n=12, p=0.0014). Importantly, this difference could not be explained by different levels of false positives, as an ROC analysis also showed a superior statistical sensitivity for the ZETA-test: mean ANOVA-AUC=0.880, mean ZETA-AUC=0.914, paired t-test, n=12, $p=7.7 \times 10^{-4}$.

Taken together, the results of comparing the ZETA-test to t-tests, ANOVAs, and renewal-process based tests show that the binless ZETA-test has a statistical sensitivity superior to all alternative tests, regardless of number of trial repetitions, brain region where the data were recorded, or specifics of the data preparation.

ZETA-test in the absence of short peaks of activity

Having established that the ZETA-test performs well in real neural data, we looked for conditions under which the ZETA-test fails. We know that the t-test has access to information that the ZETA-test does not: the spike times used by the ZETA-test are flattened over trials, while the t-test uses the variability across trials. Therefore, when the variability of mean activity

across trials is low, but the variability of spike times within a trial is high, the ZETA-test could perform worse than a t-test.

To test this hypothesis, we simulated Poisson-spiking artificial neurons (see Methods) where we changed two variables: 1) we varied the difference in spiking rate (dHz) between stimulus and baseline (both 1 s) from 0 to 1 Hz with a background rate of 1 Hz, and 2) we varied the period over which the neuron was active during stimulus presentation (T_r) from 0.5 to 1.0 s while keeping the total spike count constant. In effect, neurons with $T_r < 1.0$ show a tri-phasic response, first increasing their spiking rate, then showing a cessation of activity, and finally returning to baseline. As before, we compared the tests' performance using an ROC analysis. The ROC's area under the curve (AUC) is conveniently similar to a Wilcoxon-Mann-Whitney statistic and can be used to directly determine which of two procedures is more sensitive (Calders & Jaroszewicz, 2007). As expected (Fig. 4A), the t-test's AUC depends only on dHz and fails when dHz=0, while the ZETA-test also discriminates well when there is no difference in spike counts, but there is a consistent temporal discontinuity (dHz=0, T_r<1.0). Interestingly, while the t-test performs better than the ZETA-test when T_r=1.0, the ZETA-test also still performs reasonably well (Fig. 4B).

While this scenario is important to consider from a theoretical perspective, pure Poisson-spiking neurons probably do not exist in the brain. We therefore proceeded with a (somewhat) more biologically plausible simulation of bursting cells, where their bursting probability is orientation-tuned (Fig. 4C, see Methods and Table 1). These neurons show no consistent peaks or troughs of activity (Fig. 4D,E). However, the highly variable spike counts this bursting produces result in the ZETA-test outperforming the t-test (AUC, ZETA=0.941, t-test=0.902, z-test, $p = 4.1 \times 10^{-11}$). To conclude, even in hypothetical scenarios that we specifically constructed to investigate the limits of the ZETA-test, it performs close to the t-test

(fig. 4B4). Importantly, in the case of strongly bursting cells (Fig. 4F), the ZETA-test clearly outperforms the t-test.

Neuronal responsiveness to natural movies

Next, we asked how the performance of the ZETA-test compares to that of an ANOVA, in a case where there is no *a priori* knowledge regarding the neuronal response profile, but where the stimulus itself provides a natural timescale that may be used for binning neuronal responses. We therefore determined the responsiveness of neurons to natural movies, using either the ZETA-test, or a one-way ANOVA across bins, repeated for different bin sizes (i.e., timescales). Single-cell data were recorded using Neuropixels in 7 visual brain areas of 3 mice, while the animals were presented with repetitions of 20 s long natural movies (fig. 5A).

To ensure that the ANOVA approach could detect short bursts of activity, as well as long timescale whole-scene modulations of firing rates, we chose a wide range of bin sizes. We picked a single movie frame duration (0.0167s) as the centre point, and used bins from $1/512^{th}$ up to 512 movie frames, spaced equidistantly on a base-2 logarithmic scale in 18 steps. For each area, we pooled all cells, and calculated the area-level inclusion rate with either the timescale-free ZETA-test or Bonferroni-corrected ANOVAs at different timescales (fig. 5B). Bin sizes of 33 – 267 ms did not differ significantly from the ZETA-test's inclusion rate (FDR-corrected paired t-tests, $p>0.05$, $n=7$ areas), while the ZETA-test's inclusion rate was higher than with an ANOVA for all short (< 33 ms) and long (> 267 ms) bin durations ($p<0.05$).

The above approaches give some insight into which bin sizes best capture the dominant temporal components in neuronal responses in our data, but a more powerful approach might be to classify a neuron as "included" whenever any of the 19 ANOVAs reached significance (i.e., $p < \alpha$). Repeating this procedure for various significance levels α on the interval (0,1) produces an ROC curve (fig. 5C). Using this approach, we found there was no significant

difference in performance between a set of ANOVAs and the ZETA-test (z-test, p=0.554): ANOVAs AUC=0.798 ± 0.010, ZETA-test AUC=0.792 ± 0.011 (mean ± sd). Overall, the above results show that, under these conditions, the binless and timescale-free ZETA-test performs as well as an aggregate set of ANOVAs binned at various timescales.

Instantaneous firing rates (IFRs) for visualization and onset latency detection

The above paragraphs have shown that the ZETA-test is a sensitive statistical tool to detect whether neurons respond to a stimulus. However, it cannot be used to determine when exactly the strongest response of a neuron occurs. Therefore, we also developed a method that determines the instantaneous firing rate (IFR) using the temporal deviations upon which ZETA is based. Like the ZETA-test, it avoids the bin size selection issue of peri-stimulus time histograms (PSTHs). Another advantage of this IFR is that its temporal resolution is limited only by the neuron's spike density. Moreover, unlike model-based methods, such as the multiplicative inhomogeneous Markov interval (MIMI) model (Kass & Ventura, 2001), it requires no fitting and is orders of magnitude faster (see Methods, Figure 3 – Supplement 2). It is therefore a useful tool for determining spike train features with high precision, such as a neuron's onset latency.

Fig. 6A,B show two example V1 neurons with a relatively high (fig. 6A) and low (fig. 6B) firing rate. Here we define "onset" as the time the half-maximal response of the peak is first crossed, a metric that is heavily influenced by the chosen bin size when using PSTH-based analyses. Moreover, this bin-width-dependent estimation makes PSTH-based comparisons across cell classes problematic, as the choice of optimal bin width depends on spiking properties such as the firing rate and peak firing duration (Freedman & Diaconis, 1981; Scott, 2009; Shimazaki & Shinomoto, 2007), which are heterogeneous across neuronal cell types (fig. 6A5, B5). Hence, the main advantage of our method is that it avoids having to tailor the bin size to

each neuron individually; allowing for better comparisons across varying cell types and brain regions.

To test performance on real data, we compared the estimated onset latency using our IFR to what we obtain from a PSTH analysis with varying bin sizes. Across all V1 neurons (fig. 6C, n=119 neurons) as well as all neurons recorded in visual areas (fig. 6D, n=1403 neurons), our metric showed the strongest agreement in latency estimation with PSTH bin sizes between 1 – 10 ms.

As no ground truth is known for the real latencies of experimentally recorded neurons, we performed a benchmark test with artificially generated spike trains. Like before, we used Poisson neurons with a constant background spiking rate, where each neuron was assigned a background rate ranging from 0.5 to 32 Hz (fig. 6E-G). We used 100 trials, each 2 seconds long, and created a peak response on top of this baseline rate by adding a single spike in 50% of all trials (i.e., 50 spikes in total). We also varied the peak-response width by jittering the time each spike was added according to a normal distribution, σ ranging from 1 – 10 ms. We found that the error in the peak estimate was independent of background spiking rate (fig. 6H), indicating robust performance even when the spikes contributing to the peak were only 0.78% of the total (i.e., for 32 Hz). The mean error was also independent of the jitter, while the standard deviation of the error estimate grew in a theoretically optimal fashion as $O(\sigma)$ (fig. 6I).

We next tested the estimator accuracy when using a binning-based PSTH approach (bin width 1 – 58 ms) and found it depends on both the bin width and the width of the peak response (fig. 6J). This means that accurate latency determination using PSTHs requires the use of multiple bin sizes. More importantly, when comparing the estimation error between these bin-based methods and our IFR, we found that the binning-less IFR-based latencies were consistently as accurate as, or more accurate than, the best possible bin-width for any given peak-width (fig. 6K). Finally, the IFR-latency accuracy also exceeded the MIMI-model fit

based method. In other words, the IFR-based accuracy supersedes the PSTH-based (and MIMI-based) accuracy without the need to hand-pick the optimal binning width per neuron for a PSTH-method, nor tune hyperparameters such as knot number, location and regularization strength for the MIMI-method.

Visuomotor mismatch and spatial location are mediated by different neuronal subpopulations

Having established that the ZETA-test and IFR are statistically robust and have clear advantages over mean-rate approaches and PSTHs, we applied these tools to a GCaMP6 data set. Many theories, such as predictive coding (Friston, 2005; Gregory et al., 1980; Rao & Ballard, 1999), biologically realistic error backpropagation (Ooyen & Roelfsema, 2003; Whittington & Bogacz, 2019), and canonical cortical microcircuit operation (Bastos et al., 2012; Douglas et al., 1989), define a "top-down" signal representing an expectation, error or surprise signal as distinct from a bottom-up sensory drive. In mouse V1, such top-down visuomotor mismatch signals have been reported previously (Attinger et al., 2017; Keller et al., 2012; Leinweber et al., 2017; Saleem et al., 2013). However, whether individual V1 neurons can be classified into different groups based on their encoding of top-down visuomotor mismatch or bottom-up sensory-driven spatial location signals has not been studied.

To examine this issue, and to provide an example of how one could use ZETA in a neurophysiology study, we used neuronal calcium data recorded in L2/3 V1 of 4 mice running on a virtual-reality linear track (fig. 7A). In 87% of all corridor runs (N= 622/713 trials), the track was rendered normally, and the mice received visual feedback matching their running speed. In the remaining 13% of runs (N=91), rendering was halted at a random location for 500 ms before resuming (fig. 7B). After performing calcium-transient detection to obtain putative spike times (Montijn, Olcese, et al., 2016), we calculated ZETA-scores for all neurons in three

different ways. We aligned the spikes to mismatch-onsets, to trial starts, or converted the spike times into locations on the track, and aligned these spike locations to the start. For each recording (n=7), we calculated the Pearson correlation for each pair of these three ZETA-scores (fig. 7C-F). Across recordings, time- and location-modulation were positively correlated (mean r=0.23, one-sample t-test, n=7 recordings, p=0.016); time- and mismatch-modulation were not significant (r=0.13, p=0.28); and location- and mismatch-modulation were negatively correlated (r=-0.22, p=0.04).

We next used a kernel-density estimate (KDE) to directly test whether the joint encoding of two features within single neurons was different from chance (fig. 7G-I; Supplemental Figure 2). Indeed, we found that it was less likely that a neuron showed high modulation values for both spatial location and visuomotor mismatch than if the two features were encoded independently (z=-4.3, p=2.0 x 10^{-5}). We also found time and location to be more likely to be encoded by the same neurons (z=3.8, p=1.1 x 10^{-4}), while time- and mismatch-modulation show no effect (z=-0.06, p=0.95). This suggests a functional specialization for many neurons to encode either visuomotor mismatch signals or spatial location, but not both. While it is possible that location- and mismatch-encoding are also mediated by specific genetic subtypes of interneurons (Attinger et al., 2017), our analysis of putative pyramidal cells demonstrates that principal cells also show encoding specialization.

Optogenetic stimulation of VIP cells disinhibits visual cortex

Finally, we applied the IFR and ZETA-test to data recorded at the Allen Brain Institute (Siegle et al., 2019). In this case, we investigate whether optogenetic stimulation of VIP cells in visual cortex disinhibits the local circuit, as has been shown previously for auditory cortex and mPFC (Pi et al., 2013). The analysis of optogenetic stimulation in visually-responsive areas is complicated by the fact that mice can see the blue light used for optogenetic stimulation. In

other words, if a neuron is active after a laser pulse, it could be caused by direct stimulation, indirect circuit disinhibition, or simply be a sensory-driven response. Using the methods described in this paper to overcome these issues, we show that optogenetic stimulation of VIP-expressing cells in mouse visual cortex causes short-latency inhibition and longer-latency disinhibition in separate neuronal subpopulations (fig. 8).

Data used for this analysis were recorded from 5 Vip-IRES-Cre; Ai32 mice at the Allen Brain Institute. Cells were only included for analysis if the clustering quality was sufficient, and they were recorded in a visual cortex area (1706 cells total in AL, AM, PM, L, V1, RL, MMP). We performed a ZETA-test to discard cells that were not modulated within the interval (-0.5, +0.5 s) after optogenetic stimulation (N=1144 cells remaining). Instantaneous firing rate peak- and trough-latencies were computed and cells were discarded if their peak occurred earlier than +1 ms after optogenetic stimulation onset. The remaining cells (N=909) were classified as VIP (N=13) if their peak latencies occurred within the 10ms duration of the optogenetic stimulation, as Inhibited (N=59) or Activated (N=137) if they displayed respectively decreased or increased firing within 20ms from the stimulus offset, or Other (N=700). We limited our classification window to the first 30ms after the onset of the optogenetic stimulus as it takes about 30-50ms for retinal light responses to emerge in the visual cortex. Hence, the majority of 'Other' cells likely show sensory-driven responses.

Based on these classifications, we compared the single-cell latencies of Inhibited and Activated neurons, restricting our analysis to peaks within the 10-30ms post-optogenetic stimulation window. The latency of Inhibited cells was significantly shorter after optogenetic stimulation than it was for Activated cells (median latency Inh = 16.5 ms, Act = 23.6, Mann-Whitney U-test, $p=5.1 \times 10^{-7}$). These results show a VIP-mediated disinhibition mechanism operates in vivo in visual cortex, confirm slice connectivity studies (Pfeffer et al., 2013), and are in general agreement with results from the auditory cortex and mPFC (Pi et al., 2013).

Discussion

We developed the ZETA-test, a statistical method for determining whether neuronal spiking responses are modulated by the occurrence of events, such as the onset of sensory stimulation. The ZETA-test is widely applicable: we have shown that it can accurately detect neuronal responsiveness in a wide range of settings, across various brain regions, stimuli, and recording techniques. In most cases, the ZETA-test showed markedly improved statistical sensitivity compared to established and powerful statistical techniques, such as t-tests and ANOVAs. For example, the ZETA-test detected a visual response in 42% of the cells that were not included by a t-test and in 28% of the cells not included by a PSTH-based ANOVA. In addition to its improved statistical performance, the ZETA-test avoids arbitrary parameter choices, as it does not require the selection of a temporal bin size. This makes the ZETA-test even easier to use than established methods, as it can be applied directly to raw spike times and stimulus onsets, and the lack of a parameter selection naturally lends itself to the bulk-analysis of large numbers of cells.

Secondly, we developed an instantaneous firing rate for analysis and visualization of neuronal firing patterns. Similar to the ZETA-test, it has two main advantages over alternative common approaches, such as PSTHs: 1) it does not require binning, and therefore removes the need to tune binning widths for individual cells or cell types; and 2) its temporal resolution is only limited by the spike density, which allows more accurate determination of spike train events such as peak response latencies. When benchmarking its performance for latency-estimation, it outperformed PSTH-based approaches without the need to fine-tune bin sizes for each neuron individually (fig. 6K). Finally, compared to methods that rely on model fitting, such as the multiplicative inhomogeneous Markov interval (MIMI) model, the ZETA-test is

considerably faster, not subject to overfitting, and does not require tuning of hyperparameters (Figure 3 – Supplement 2).

We investigated the performance of the ZETA-test in a variety of cases and observed that the ZETA-test sometimes performed as well as other techniques and sometimes better. So when, and why, exactly does the ZETA-test provide extra statistical power? We have shown in the methods that the ZETA-test shares mathematical properties with a Kolmogorov-Smirnov test of a neuron's spike train against permuted onset-jitter bootstraps. Various components of the ZETA-test were chosen to relax the assumptions made by other tests with the aim to gain relative invariance to the specifics of a neuron's spike train (Figure 3 – Supplement 3). While the ZETA-test is slightly worse than a t-test for purely Poisson-distributed spike counts, it outperforms the t-test in every other case we tested, including those where the interspike interval distribution is highly peaked, such as in bursting cells (Figures 1 – 4). Finally, while more sophisticated model-based approaches might be able to attain better performance than the ZETA-test, this requires that their hyperparameters be individually tuned to a cell's firing statistics (e.g., MIMI-based methods; Figure 3 – Supplement 2). Therefore, the ZETA-test may often be the preferred choice as it required no fitting, hyperparameter tuning, and shows a statistical sensitivity superior to model-based approaches in all cases we tested.

An alternative to using the ZETA-test could be to perform a set of ANOVAs as in fig. 5. While this would certainly be an improvement over doing a paired t-test, multiple ANOVAs are more complex to implement and the procedure needs multiple-comparison correction (Head et al., 2015). More importantly, it still requires choosing an arbitrary set of bin sizes, as simply taking the optimal bin size still leads to an underperformance relative to the ZETA-test (fig. 1G). Moreover, the ZETA-test has the advantage of not having to choose any parameter at all. As we have shown in Figure 8, using the ZETA-test and IFR allowed us to significantly simplify

various analysis steps when investigating the heterogeneous latency effects of optogenetic stimulation of VIP neurons.

However, like all statistical tools, the ZETA-test also has its limitations. As we designed it to be a generalist test applicable to any spike train, the ZETA-test might not perform as well as models manually fitted to describe a particular cell's response. However, we believe in many cases it will provide a superior alternative to other responsiveness tests, and response-latency determinations. The ZETA-test shows great merit especially as an unbiased parameter-free inclusion criterion, which is something that currently often varies between studies and can adversely affect the replicability of scientific results (Mesa et al., 2021). Another, and fairly obvious, limitation is that the ZETA-test only compares point-events (e.g., spike times) to other point-events (e.g., stimulus onsets). Consequentially, it is a powerful tool for state-of-the-art electrophysiological techniques, like Neuropixels recordings, but does not apply to calcium imaging dF/F0 traces. However, as we show in fig. 7, the ZETA-test performs well when applied to putative spiking events extracted from calcium imaging data. Therefore, the ZETA-test would still be useful in calcium imaging data sets where the temporal response profile of neurons to the experimental treatment is unknown. Moreover, optical recording techniques continue to improve, allowing better single-spike extractions with calcium imaging (Packer et al., 2015) and genetically encoded voltage indicators (Knöpfel & Song, 2019). As such, in all likelihood near-future neural data will remain spike-based regardless of the underlying recording technique used.

The main aim of this paper is to present the IFR and ZETA-test and show how they improve upon commonly used analysis tools, such as PSTHs, t-tests, Markov-model based approaches, and ANOVAs. However, we also show some results that by themselves are scientifically noteworthy. For example, our results indicate that almost all V1 neurons (>93%) are responsive to drifting gratings, even if the spatial frequency (SF) and temporal frequency

(TF) parameters are not optimal to drive these cells (fig. 3). This calls into question the idea that V1 cells are generally narrowly tuned and can only be driven by specific stimulus feature parameters (Xing et al., 2004). Our analysis indicates that neuronal responsiveness rarely drops to 0 for a combination of stimulus features, as otherwise we would not have found such a large number of responsive V1 cells using a feature-sparse stimulus (24 directions, 1 SF: 0.05cpd, 1 TF: 2Hz).

This suggests that the response of V1 neurons to features such as SF and TF may drop off rather slowly with distance to their preferred stimulus properties. Conceptually, this would mean that the V1 neural code may in fact be more "dense" than proposed by some (Ohiorhenuan et al., 2010; Olshausen & Field, 1997; Vinje & Gallant, 2000). That said, our results show that only a relatively small group of neurons is strongly driven by any one stimulus: sparse coding may therefore still operate in V1, but it acts on top of a dense, but weak code.

Two other noteworthy findings are that VIP cell activation drives disinhibition in visual cortex (fig. 8) and that the strength of visuomotor signals and modulation by spatial location are negatively correlated (fig. 7). This suggests that bottom-up and top-down processing in visual cortex may be mediated by distinct neuronal subnetworks. While this functional segregation has previously been shown based on a macroanatomical (e.g., laminar) analysis (Kok et al., 2016; Markov et al., 2014; Poort et al., 2012; Self et al., 2019), our results suggest that this functional segregation also holds at a microanatomical level for neurons located within a single recording plane (fig. 7).

In conclusion, the IFR and ZETA-test are simpler, more statistically powerful, and less error-prone tools than bin-based PSTHs, t-tests, Markov-model based approaches, and ANOVAs widely used in neuroscience today. Statistically underpowered studies are still common in neuroscience, which makes the development of statistically sensitive tools especially important (Button et al., 2013). Moreover, when power analyses are used to

determine the necessary sample size, then increased statistical sensitivity, such as with the ZETA-test, can reduce the number of required experimental animals. To facilitate the adoption of the ZETA-test by the neuroscientific community, we provide easy-to-use and well-documented open-source implementations online in MATLAB (https://github.com/JorritMontijn/ZETA) and Python (https://github.com/JorritMontijn/zetapy).

Methods

ZETA

Well documented and easy-to-use Matlab and python code performing the procedures described in the following paragraphs can be found here: https://github.com/JorritMontijn/ZETA and https://github.com/JorritMontijn/zetapy.

We developed a timescale-free, binning-less statistical test for determining whether a neuron shows a time-locked modulation of spiking activity. It is derived from a metric that represents the reliability, as number of standard deviations away from chance, that the temporal density of spikes is non-random across trial repetitions. This metric, we call ZETA (ζ), can be computed on a vector of $i = [1 \ldots N]$ spike times $\boldsymbol{x}$, and a vector of $k = [1 \ldots q]$ event times (e.g., stimulus onsets), $\boldsymbol{w}$, using the following steps.

First, we make a vector $\boldsymbol{v}$ of the spike times in $\boldsymbol{x}$ relative to the most recent stimulus onset, as when making a raster plot of spike times:

$$v_i = x_i - w_k \tag{1}$$

Where

$$w_k < x_i \leq w_{k+1} \tag{2}$$

Next, we remove all spike times that are larger than a cut-off value τ, for example the trial duration, and add two artificial spikes at t=0 and t=τ to ensure coverage of the full epoch. We

sort the n spike times in $\boldsymbol{v}$ such that $v_i < v_{i+1}$, and calculate the fractional position g_i, ranging from $1/n$ to 1, of each spike time in $\boldsymbol{v}$:

$$g_i = i/n \tag{3}$$

Therefore, another interpretation is that $\boldsymbol{g}$ represents a neuron's cumulative density function sampled at the spike times in $\boldsymbol{v}$. In order to quantify whether this distribution is different from our null hypothesis – i.e., that the neuron's firing rate is not modulated with respect to the stimulus onset – we compare this vector to a linear baseline density vector $\boldsymbol{b}$. If a neuron's spiking rate is constant, the cumulative density function is linear over time, and therefore the expected fractional position of spike i at time v_i converges to the spike time divided by the trial duration τ as the number of events q increases:

$$\lim_{q \to \infty} b_i = v_i/\tau \tag{4}$$

The difference δ_i between g_i and b_i therefore gives a neuron's deviation from a temporally non-modulated spiking rate at time point v_i:

$$\delta_i = g_i - b_i \tag{5}$$

As we show in the Supplementary Methods, this variable is not time-invariant. Therefore, we create $\boldsymbol{d}$, a mean-normalized version of $\boldsymbol{\delta}$:

$$d_i = \delta_i - \bar{\delta} \tag{6}$$

Where

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \tag{7}$$

We then define the Zenith of Event-based Time-locked Anomalies (ZETA, or ζ) as the most extreme value, i.e. the maximum of the absolute values:

$$\zeta \equiv \max(|\boldsymbol{d}|) \tag{8}$$

Null hypothesis for ZETA

Having calculated ZETA from the temporal deviation vector ***d***, we wish to quantify its statistical significance. First, we scale it such that its value is interpretable as a z-score. We therefore construct a null hypothesis distribution by repeating the above procedure *P* times with jittered event-times ***w'***, where we move each event time by a random sample drawn from the interval [-τ, τ]. This way, we calculate the chance of observing randomly high values in ***d*** without having to make assumptions about the underlying distribution of ***d***. However, a naïve approach would lead to difficulties here, as jittering ***w*** also changes the corresponding values of ***v***; and any jittered vector ***d'*** we obtain would be sampled at different times than the original vector ***d***. Therefore, we instead linearly interpolate the values of jittered fractional position vector ***g'*** at the original spike times of ***v***. First, we construct a vector ***f*** of fractional spiking positions analogously to ***g***, but based on jittered event-times ***w'***:

$$f_i = i/n' \qquad 9$$

Note that we cannot simply take ***g***, as the total number of spikes *n'* in this jittered version is likely different from the original number of spikes *n*, because we only consider the spike times in the interval [0, τ] after the (jittered) event times. Next, we interpolate the values of ***f*** at sample times ***v'*** to the original sample times ***v***:

$$g_i' = (1 - w)f_{j-1} + wf_j \qquad 10$$

Where

$$w = \frac{v_i' - v_j}{v_k - v_j} \qquad 11$$

With

$$v_{j-1} \leq v_i' \leq v_j \qquad 12$$

We repeat this process *P* times; where for each jitter iteration *j*, we calculate ***δ'(j)***:

$$\boldsymbol{\delta}'(j) = \boldsymbol{g}'(j) - \boldsymbol{b} \qquad 13$$

Note that $\boldsymbol{b}$ is invariant with respect to the jitter iteration, as it is simply the *n*-element linear vector from *1/n* to *n*. As before, we mean-normalize $\boldsymbol{\delta}'(j)$ to obtain a temporal deviation vector $\boldsymbol{d}'(j)$.

$$\boldsymbol{d}'(j) = \boldsymbol{\delta}'(j) - \bar{\delta}'(j) \qquad 14$$

Now we can define a null-hypothesis ZETA sample *j* as:

$$\zeta'(j) \equiv \max(|\boldsymbol{d}'(j)|) \qquad 15$$

Statistical significance of ZETA

Having constructed a way to generate samples from a null-hypothesis distribution, we are left with the task of using it in calculating the statistical significance of ZETA. If we had infinite samples, we could directly calculate the percentile of the empirical ζ from the null-distribution. However, as this is computationally intractable, we will approximate the true distribution from a finite number of null-hypothesis samples. From extreme value theory we know that the distribution of maximum values is known as a Gumbel distribution (Gumbel, 1941). Its cumulative density is given by:

$$F(x;\, m, \beta) = e^{-e^{-(x-m)/\beta}} \qquad 16$$

Here, *x* is the sample maximum (i.e., ζ), *m* is the mode, and *β* is the scale parameter. Therefore, we need to find *m* and *β*, which can be derived from the estimated sample mean and variance over jittered ZETAs of $\boldsymbol{\zeta}'$. The mean $\bar{x}$ and variance *v* are given by (Gumbel, 1954):

$$\bar{x} = m + \beta\gamma \qquad 17$$

$$v = \frac{\pi^2\beta^2}{6} \qquad 18$$

Here, γ is the Euler–Mascheroni constant (γ ≈ 0.577), *m* is the mode, and *β* is the scale parameter. Using $v = \mathrm{Var}(\boldsymbol{\zeta}')$, and equation 18, we can write the scale parameter *β* as:

$$\beta = \frac{\sqrt{6 \cdot \mathrm{Var}(\boldsymbol{\zeta}')}}{\pi} \qquad 19$$

Then using equation 17 and $\bar{x} = \overline{\zeta'}$, the mode can be computed from β and the mean:

$$m = \overline{\zeta'} - \beta\gamma \qquad 20$$

Now we can define the p-value by reading out the cumulative Gumbel distribution at ζ:

$$p = 1 - F(\zeta;\ m, \beta) \qquad 21$$

Finally, we can use p to obtain a corrected ZETA ζ_c that is interpretable as a z-score:

$$\zeta_c = \Phi^{-1}\left(1 - \frac{p}{2}\right) \qquad 22$$

Note that when we refer to ZETA or ζ in the rest of the manuscript, we mean the corrected version and its p-value as defined above.

Computing an optimal bin size

For the analyses where we used an optimal binning width to compare the performance of the ZETA-test and a bin-wise ANOVA, we computed the optimal bin width using the procedure described by (Shimazaki & Shinomoto, 2007). Their method describes a loss function that can be computed for a given bin-width. To find the optimal bin width, we used a simple iterative 10-point grid search until a local minimum was found. The code used for finding the optimal bin size is available online at https://github.com/JorritMontijn/GeneralAnalysis in the function *opthist.m*.

The multiplicative inhomogeneous Markov interval (MIMI) model

A classic model for neuronal firing rates are inhomogeneous Poisson processes, where a time-dependent function *f(t)* can be used to describe how the mean firing rate λ of a cell varies with time *t* after some experimental intervention, such as the onset of a visual stimulus (Kass et al., 2014):

$$\lambda(t) = f(t) \tag{23}$$

Spike-times $\boldsymbol{v}$ can then be generated by sampling from an exponential distribution with an inter-spike interval equal to $1/\lambda$:

$$v_{i+1} = v_i + \mathrm{Exp}\left(\frac{1}{\lambda}\right) \tag{24}$$

While this framework is attractive in its simplicity, it cannot capture several important properties of spiking dynamics, such as refractory periods or burst firing. Even when one is only interested in the question whether a particular neuron responds to a visual stimulus, bursting cells might produce apparent bumps in a peri-stimulus time histogram (PSTH). This could lead to the possibly erroneous conclusion that a cell is stimulus-modulated, simply because their spiking patterns are non-Poisson by nature. This problem is remedied by a class of models that combine a Poisson process with a renewal process, which describes the likelihood of a spike conditioned on the time since the last spike. These processes are called multiplicative inhomogeneous Markov interval (MIMI) processes (Kass & Ventura, 2001):

$$\lambda\big(t, t - s_*(t)\big) = \lambda_1(t) \cdot \lambda_2\big(t - s_*(t)\big) \tag{25}$$

Here, $s_*(t)$ is the time of the last spike, the λ_1 term refers to the inhomogeneous Poisson process described above, while the latter λ_2 term captures the inter-spike-interval dependent spiking probability. While this model can be extended to include interaction terms, n-back spike dependencies, and bias constants per trial, prior work has shown these additions do not appreciably improve the fitting quality (Kass & Ventura, 2001). As an additional baseline model, we therefore also compared the performance of the ZETA-test to that of a method based on the MIMI-model (eq. 25).

MIMI-model fit evaluation as a statistical test for responsiveness

To use the MIMI-model framework as a statistical test, we binned spikes in 1-ms bins. We used cubic splines with 16 B-form coefficients spread uniformly over the trial's 1.5 s duration for

the inhomogeneous Poisson component, and 16 B-form coefficients spread uniformly over a time horizon of 500 ms for the renewal-process component (Kass et al., 2014). We then simultaneously fitted these 32 coefficients to produce the closest match to the neuron's spike train by running a least-squares curve fitting algorithm. Specifically, we minimized the error between the stimulus-locked trace reconstituted from the B-form splines and the real average spiking rate per bin of the PSTH. The fitting procedure used 1-ms binning, but the resulting model can be resolved at theoretically infinitesimal time steps. We ran a couple of fits with different numbers of coefficients ranging from 8+8 to 32+32, but this did not strongly impact either the fitting quality or computation time (Figure 3 - Figure supplement 2).

Theoretically, we could now follow the same procedure as with the ZETA-test by generating a null-hypothesis distribution from multiple iterations of jittered onset times. However, the MIMI fitting procedure's computational time cost meant this was not a realistic solution: even a single MIMI-model fit took 557 times as long to run as the ZETA-test, which already included 100 random jitter iterations (Figure 3 - Figure supplement 2, panel H). This means that using 100 random jitters for the MIMI-model test would take 55,700 times as long as the ZETA-test. We therefore took an alternative approach by calculating pairwise d' values for all bin pairs, and transformed the highest d' (i.e., the distance in standard deviations between the two most dissimilar bins) into a p-value. This is obviously a suboptimal approach, as this will lead to a high number of false positives. However, our goal here is not to develop a full-fledged MIMI-based test, but rather to verify what the statistical sensitivity of such an approach could be. It is possible to investigate the statistical sensitivity as the low p-value bias exists for both real data and shuffle controls, so instead of comparing the inclusion level at an alpha of 0.05, we performed an ROC analysis over all cells in V1.

The ROC analysis is insensitive to the absolute level of significance values, but instead provides insight in the discriminability of real inclusions from false positives (Figure 3 - Figure

supplement 2, panel G). To keep the computational time tractable, we subsampled the data to include only V1 cells, and only their response to the first 480 drifting grating trials. We compared the ZETA-test, t-test and MIMI-model method as described above. The ZETA-test gave an area under the curve (AUC) of 0.996, the t-test 0.900, and the MIMI-method 0.749. However, we noticed that the MIMI-method appeared to fail mostly for cells with low firing rates, so we also added a hypothetical curve where we only included neurons with >1000 spikes during the 480-trial long epoch (MIMI-1k). This significantly boosted the MIMI-method's discriminability to an AUC similar to the t-test's at 0.898. These results suggest it will require significant work to develop MIMI-based test that can compete with a t-test, and that even if we were successful, it might not exceed the ZETA-test's performance. Moreover, as the MIMI-model requires parameters to be iteratively fitted to experimental data, it is multiple orders of magnitude slower than the ZETA-test and t-test. To conclude, full MIMI-model based methods do not seem to be suited for unsupervised, large-scale use as neuronal responsiveness tests.

Decomposition of ZETA

While using full multiplicative inhomogeneous Markov interval models of the form of eq. 25 is not a viable option when creating a responsiveness test, we used it to as a starting point to further explore which properties of the ZETA allow it to function so well. As already noted above, the main problem for developing a robust responsiveness test is finding a suitable null-hypothesis distribution to test against. As a first naïve baseline, we built a simple test that checks whether a cell's firing rate, binned with width τ and averaged across trials, differs from a homogeneous Poisson process with rate λ. Under this null-hypothesis, the number of spikes per bin X is therefore distributed as:

$$X \sim Pois(\lambda\tau) \qquad 26$$

The null-hypothesis random variable H_0 for rates averaged over T trials follows:

$$H_0 = 1/T \sum_{i=1}^{T} X_i \quad (27)$$

Whether a neuron's observed number of spikes per bin, averaged over trials, differs from this null distribution can then be tested using a standard Kolmogorov-Smirnov test.

We benchmarked this approach with V1 cells, using 1-ms bins, and performed an ROC analysis of real inclusions versus false positives using two different ways to construct neuronal responses that are unmodulated by the visual stimulation: 1) we shuffled the inter-spike intervals to construct a new set of spike times (Figure 3 – Figure Supplement 3A), and 2) we jittered the stimulus onset times by ±6 seconds (Figure 3 – Figure Supplement 3B). We found that the Poisson-based test performed well if the H0 spiking statistics matched the shuffle-control (Poiss-KS, AUC=0.976), but failed due to high numbers of false positives when we instead jittered the stimulus onset times (Poiss-KS, AUC=0.611). By construction, the fluctuations in the rate are not linked to the stimulus onset times anymore, but the distribution of the spike times over the trial period are not consistent with a Poisson neuron.

This may seem a trivial result, as the null-hypothesis distribution in this latter case does not match the null (Poisson) distribution used by the statistical test. This is a critical issue, however: in the case of real experimental data sets, there is no known ground truth, so a robust statistical test for neuronal responsiveness must be able to handle a wide variety of intrinsic spiking behaviors. Clearly, the Poisson test fails this requirement, as it is only able to distinguish i.i.d. Poisson-distributed spiking from anything that is not exactly i.i.d. Poisson-distributed spiking. For comparison, both the ZETA-test and t-test show robust behaviour that is insensitive to the specifics of the shuffle-control we use. The ZETA-test gives an AUC of 0.983 using ISI-shuffles and 0.984 using onset jittering; and the mean-rate t-test gives AUCs of 0.899 and 0.902 respectively. In the following section we investigate which properties of the ZETA-test allow it to perform so much better than the Poisson test.

The first aspect we investigated is the assumption of the homogeneous Poisson-distributed spiking when a neuron is unmodulated by visual stimulation. We know that neurons can be intrinsically bursting and have refractory periods, so even the most purely sensory-driven cell in V1 is likely to not fire i.i.d. Poisson when no visual stimulus is present. A more versatile and possibly more accurate H0 might be a renewal process. We therefore constructed an inter-spike-interval (ISI)-based test, where we first calculated a neuron's inter-spike intervals $\boldsymbol{dt}$ from its spike time vector $\boldsymbol{t}$:

$$dt_i = t_i - t_{i-1} \tag{28}$$

We then randomly permuted the ISIs, creating a shuffled ISI vector $\boldsymbol{dt}^s$ and using it to construct a null-hypothesis vector of spike times $\boldsymbol{t}^0$:

$$t_i^0 = t_{i-1}^0 + dt_i^s \tag{29}$$

In effect, this null-hypothesis vector is a different random sample of the same renewal process that would generate the real neuron's spiking times, under the simplifying assumption that the spike times are generated by a renewal process. We constructed mean firing rates $\boldsymbol{x}^0$ by binning $\boldsymbol{t}^0$ using 1-ms bins. We repeated this 100 times, and used a two-sample K-S test between the real binned spike count vector $\boldsymbol{x}$ and shuffle-control spike count matrix $\boldsymbol{X}^0$, as we also did for the Poisson test described above. Unfortunately, this SISI-KS (Shuffled Inter-Spike-Interval Kolmogorov-Smirnov) test performed very similar to the Poisson test, resulting in an AUC of 0.985 for the ISI-shuffle control, and an AUC of 0.610 for the jitter-control.

This similarity might be explained by the ability of the KS test to distinguish with high sensitivity between the spiking distributions obtained from shuffling ISIs and jittering stimulus onsets, as these are not identical. If we wish to construct a more robust test, we must therefore use a procedure that is less sensitive to the full shape of the H0 distribution, and only takes into account the likelihood of observing extreme deviations from the average firing rate. This will make the test less sensitive to real stimulus-induced activity, but also less sensitive to errors in

the specific shape of the null hypothesis distribution we use to estimate the neuron's natural variability. We achieved this by using the Gumbel-distribution of maximum absolute deviations in our random ISI-shuffle samples to calculate a p-value of the maximum absolute deviation in the real, unshuffled, PSTH. The procedure works as intended: this SISI-G test gives an AUC of 0.844 for the ISI-shuffle controls, and an AUC of 0.821 for the jitter controls.

We noticed that the SISI-G test suffered from high variance in the PSTH when using 1-ms bins, especially for cells with few spikes. One option would therefore be to increase the bin width, but this would come at the expense of temporal resolution and ability to detect short peaks of activity. We therefore opted instead to calculate the maximum absolute deviation of the cumulative sum of spikes counts, similar to the ZETA-test's, but in this case over the discrete 1-ms spike count vector $\boldsymbol{x}^0$. For the SISI-∫G test, we defined the normalized cumulative spike count vector T0 (similar to eq. 29) as:

$$T_j^0 = \sum_{i=1}^{j} x_i^0 - \bar{x}^0 \qquad 30$$

Moreover, as this would create fixed points with 0 variance at j=1 and j=n, we also mean-subtracted the T0 vector itself:

$$\boldsymbol{T}^{0,c} = \boldsymbol{T}^0 - \bar{T}^0 \qquad 31$$

In essence, this test is a 1-ms binned and ISI-shuffle based version of the binless onset-jittering ZETA-test. Benchmarking this test, we found that it performed close to, but slightly less well than, the ZETA-test. This SISI-∫G test gave an AUC of 0.968 for the ISI-shuffle controls and an AUC of 0.974 for the jitter controls.

Finally, we created an alternative, also binless, version of the ZETA-test where we created the null distributions by shuffling the inter-spike intervals rather than jittering the stimulus onsets. The alternative ZETA-ISI test performed at a level indistinguishable from ZETA; the ZETA-ISI gave an AUC of 0.986 for the ISI-shuffle controls and an AUC of 0.982

for the jitter controls. To conclude; the (alternative) ZETA-test strongly outperforms other tests, mainly for two reasons: 1) using the Gumbel distribution to calculate a cell's significance based on the most extreme stimulus-locked spiking deviation rather than a KS test allows the ZETA-test to be relatively invariant to the full, and a priori unknown, spike time distribution of a neuron, and 2) using an integral-based approach has a timescale-free smoothing effect that reduces spurious peaks that can occur in the firing rate domain. Finally, the binless ZETA-test shows a small, but significant, improvement over its 1-ms binned cousin in terms of both statistical power and computational efficiency (Figure 3 – Figure Supplement 3C). While the random null distribution (ISI shuffling or onset jittering) did not seem to have a large impact on the ZETA-test's performance for this data set of predominantly regular-spiking V1 neurons, the following section shows that this distinction becomes more important when one tests the responsiveness of bursting cells.

Simulated bursting cells: onset jittering versus inter-spike interval shuffling

Stimulus onset jittering and inter-spike interval shuffling produce different distributions, unless a neuron's probability of spiking only depends on the time since the last spike. The previous section showed that, when using our V1 data set, performance has already saturated too much to show a difference between the ISI-shuffle ZETA and onset-jitter ZETA tests. To better differentiate their performance, we therefore generated a population of simulated bursting cells. While bursting cells are rare in visual cortex, they are abundant in many brain regions, such as the subiculum and others (Cooper, 2002; Mattia et al., 1993). To test the performance on bursting cells, we generated artificial spike trains, using the parameters for burst properties from (Chen et al., 2009) as listed in Table 1.

All cells were assigned a background single-spike firing rate that was on average 1 Hz; on top of these single spikes we generated bursts with varying length and inter-spike intervals.

The inter-burst intervals were 16s on average during inter-trial intervals and non-preferred stimuli; and 0.8 s during the neuron's preferred stimulus. Each neuron had a randomly assigned preferred orientation and tuning width following a von Mises distribution. We presented 20 repetitions of 24 orientations (15 degree steps) for a total of 480 trials. We generated responses for 1000 independently parameterized neurons (see Figure 3 - Supplement 3D-E). To create stimulus non-responsive neurons, we set the inter-burst intervals to be approximately 6.8 s both within and outside stimulus presentation; we chose this value to ensure the spiking rate distributions for stimulus-responsive and unresponsive neurons were approximately equal (25th-75th population percentiles of firing rates; stimulus-responsive: 6.7 – 19.7 Hz; unresponsive: 6.9 – 20.7 Hz).

Running the same benchmark as before on this artificial data set, we found that the onset-jitter ZETA-test indeed outperformed both the ISI-shuffle ZETA-test and mean-rate t-test (Figure 3 - Supplement 3F). The ZETA-test gave an AUC of 0.941, the ISI-shuffle ZETA-test an AUC of 0.845, and the mean-rate t-test an AUC of 0.902. This difference in performance can be attributed to the fact that jittering stimulus onsets keeps the properties of individual bursts intact, while ISI-shuffling changes these properties, leading to more variable burst spike trains. While shuffling of ISIs or stimulus onsets both produce spike trains that are unmodulated by stimulus presence, we have shown that they are not equivalent.

Multi-scale derivatives of ZETA for latency detection

The ZETA-test indicates whether a neuron shows reliable deviations in spiking rate with respect to a particular series of events. The time of this maximum deviation, however, is not necessarily when the neuron shows its strongest firing rate modulation, but rather when the cumulative distribution reaches peak statistical significance. Therefore, in order to use the ZETA procedure to calculate the time of peaks in modulations (e.g., onset latencies), we should take the

derivative of the temporal deviation vector $\boldsymbol{d}$ underlying ZETA. A naïve approach with a simple spike-to-spike derivative unfortunately yields a curve with many spurious peaks. One solution would be to calculate the derivative over a larger time interval, but this comes at the expense of temporal resolution. Moreover, many different cell types exist with different dominant time constants. To balance temporal resolution and robustness, we therefore developed a multi-scale derivative procedure. First, we define a vector $\boldsymbol{t}$ of S timescales at which to compute derivatives. By default, we define the timescales to lie on a logarithmic scale with base 1.5, as this gave a reasonable trade-off between computational speed and accuracy. Base values closer to 1 will give more accurate results at the cost of computational speed. For base b and a trial duration of τ:

$$\boldsymbol{t} = b^{\boldsymbol{p}} \tag{32}$$

Where

$$\boldsymbol{p} = \{x \mid x \in \mathbb{Z}, \log_b 10^{-3} < x < \log_b \frac{\tau}{10}\} \tag{33}$$

The derivative $\dot{d}$ at spike i can then be defined for d_i and timescale t_k as:

$$\dot{d}_{i,k} = \frac{d_b - d_a}{v_b - v_a} \tag{34}$$

Where

$$a = \arg\max v_a \{v_a \mid v_a \in \boldsymbol{v}, v_a < v_i - \frac{t_k}{2}\} \tag{35}$$

$$b = \arg\min v_b \{v_b \mid v_b \in \boldsymbol{v}, v_b > v_i + \frac{t_k}{2}\}$$

Here, $\boldsymbol{v}$ are spike times, following the definition above. To avoid undefined edges, we set a and b to 1 and n respectively, iff $v_i \pm t_k/2$ falls outside the interval $[0, \tau]$. Taking the mean over all S timescales, we obtain an average of multi-scale derivatives $\boldsymbol{m}$,

$$m_i = \frac{1}{S}\sum_{k=1}^{S} \dot{d}_{i,k} \tag{36}$$

which has two important properties. First, long-timescale derivatives tend to 0, so there is a bias of $\boldsymbol{m}$ to more strongly follow shorter timescales. Secondly, random noise at the shortest timescales averages out over multiple short-timescale derivatives. Therefore, these two properties combined lead $\boldsymbol{m}$ to reflect the shortest timescales at which a real signal starts to emerge from random noise.

Calculation of high-resolution instantaneous firing rates

Another interesting property of the mean multi-scale derivative $\boldsymbol{m}$ is that it scales with the actual firing rate. In other words, it produces a time-locked neural activation curve, similar to a peri-stimulus time histogram (PSTH). If we properly rescale $\boldsymbol{m}$, we can therefore create an instantaneous firing rate metric with a temporal resolution that is only limited by the spike density.

Remember that the temporal deviation vector $\boldsymbol{\delta}$ itself is scaled to lie between -1 and +1, as its value depends on the difference between the fractional position of a spike (from 0 to 1) and the linear interval from $x,y=[0,0]$ to $[\tau,1]$. The theoretical lower limit of the multi-scale derivative is therefore $-1/\tau$. This can be illustrated as follows. Imagine a hypothetical neuron where all spikes are fired in an arbitrarily short interval close the start of each trial. This means that $\boldsymbol{\delta}$ rises from 0 to 1 in a short interval and from then decays linearly from 1 to 0 over an interval of τ. As $\dot{d}$ is defined between two spikes (including window edges 0 and τ), this means that the lowest possible value in this extreme case is the point between the last spike n and τ, i.e.:

$$\dot{d}_{\min} = \frac{d_\tau - d_n}{v_\tau - v_n} = \frac{0-1}{\tau - 0} = -1/\tau \qquad 37$$

The lowest possible firing rate is 0 Hz, which therefore corresponds to $-1/\tau$. An upper bound for $\dot{d}$ does not exist, as an arbitrarily short interval with a finite number of n spikes would lead to arbitrarily high $\dot{d}$:

$$\dot{d}_{\max} = \frac{d_j - d_i}{v_i - v_j} = \frac{1/n}{t - (t + 1/\infty)} = \infty \qquad 38$$

This is a desirable property, as the maximum instantaneous firing rate of a neuron is not theoretically constrained. Finally, the average firing rate of our metric should correspond to the real average firing rate in Hz ($n/\tau q$), where q is the number of events as defined above. We therefore define our instantaneous firing rate metric r as:

$$\boldsymbol{r} = \frac{n}{\tau \cdot q} \cdot \left(\frac{\boldsymbol{m} + 1/\tau}{\overline{m} + 1/\tau}\right) \qquad 39$$

Here, $\overline{m}$ is the weighted average of $\boldsymbol{m}$ by the inter-spike interval, such that the averaging occurs in the time domain and not the spike-number domain:

$$\overline{m} = \frac{1}{\tau}\sum_{i=2}^{n} \frac{m_{i-1} + m_i}{2}(v_{i-1} - v_i) \qquad 40$$

Considering the definitions above, we can therefore state that the maximum firing rate in $\boldsymbol{r}$ occurs where

$$r_i = \frac{1}{S}\sum_{k=1}^{S} \frac{b_k - a_k}{v_{b_k} - v_{a_k}} \qquad 41$$

is at its maximum, with a_k the index of the largest spike time smaller than $v_i - \frac{t_k}{2}$ and b_k the index of the smallest spike time larger than $v_i + \frac{t_k}{2}$; with t_k being the logarithmically distributed time ranges. In the limit of a high number of spikes, we find that $v_{a_k} \approx v_i - \frac{t_k}{2}$ and $v_{b_k} \approx v_i + \frac{t_k}{2}$, and therefore

$$r_i = \frac{1}{S}\sum_{k=1}^{S} \frac{\#\text{spikes in}\left(v_i - \frac{t_k}{2}, v_i + \frac{t_k}{2}\right)}{t_k} \qquad 42$$

which is an average of the instantaneous spike rates at v_i at timescales t_k.

Mean-rate artificial Poisson neurons

We tested whether ZETA required short bursts of activity to work by generating artificial spike trains that only varied in mean-rate between a 1-second stimulus presentation and a 1-second inter-stimulus interval, and did not show onset peak responses. We created spike trains for 100 neurons with an orientation preference θ randomly sampled from a uniform distribution on the interval (0,π). The shape of the tuning curve was defined as the sum of two von Mises distributions centered at preferred orientation θ and θ+π with a concentration parameter of κ=5+ε, where ε was randomly sampled from a uniform distribution on the interval (0,5). The von Mises probability density function with mean θ and concentration parameter κ is given by:

$$f(x \mid \theta, \kappa) = \frac{e^{\kappa \cos(x-\theta)}}{2\pi I_0(\kappa)} \qquad 43$$

Here, where $I_0(\kappa)$ is the modified Bessel function of order 0. The baseline mean spiking rate μ_{base} was defined by randomly sampling from an exponential distribution with a mean of λ_{base} = 5 Hz. μ_{base} defined the trough of the neuron's tuning curve (i.e., the activity at θ+π/2 and θ-π/2) as well as the activity of the neuron when no stimulus was present. The firing rate for the preferred stimulus μ_{stim} was determined by similarly sampling from an exponential distribution with a mean of $\lambda_{stim} = \mu_{base}$ + 20 Hz. We generated spiking activity for 160 trials (20 repetitions of 8 stimulus orientations: θ_{stim}= [0, 45, …, 315]). The average baseline firing across all n=10000 artificial neurons was therefore 5 Hz and the average firing rate during the preferred stimulus was 25 Hz. Spike times were generated for each trial-epoch (stimulus/baseline) independently by consecutively drawing inter-spike intervals from a Poisson distribution with λ = 1/μ.

Artificial Poisson neurons for peak-latency benchmarking

We also addressed the question whether our instantaneous firing rate metric was sufficiently robust to allow accurate peak-time detection over a range of background firing rates and a range

of peak widths. The procedure here was similar as above, with the exception that the firing rate during baseline and stimulus periods was identical: $\mu_{base} = \mu_{stim}$. We tested 13 base rates (0.5 – 32 Hz) and 19 jitter widths (1 – 10 ms in steps of 0.5). For each combination of base rate and jitter width, we generated 100 neurons and 100 trials per neuron. Peaks were added to the background activity by adding a single spike to half of all trials. Spike times were chosen by random sampling from a normal distribution with the standard deviation equal to the above jitter width and centered at 100 ± 10 ms after stimulus onset.

Figure panels 6J,K used slightly different parameters. Instead, we used 10 jitter widths (1 – 10 ms in steps of 1), simulated only a base rate of 32 Hz, used 160 trials for 1000 neurons, and compared the peak-latency detection using the ZETA-IFR with binning windows ranging from 1.00 ms to 57.67 ms; a logarithmic scale of base 1.5 with the exponent ranging from 0 to 10 in steps of 1.

Acquisition and preprocessing of laminar probe data (Neuronexus and Neuropixels)

We performed silicon probe recordings in six C57BL/6 mice, 2–7 months of age. Mice were housed in a 12 h/12 h dark/light cycle with ad libitum access to food and water. All experiments were approved by the animal ethics committee of the Royal Netherlands Academy of Arts and Sciences, in compliance with all relevant ethical regulations.

Mice were habituated for 1-4 weeks before being implanted with a cranial bar used for head-fixation. Mice were anesthetized with isoflurane (3% induction, 1-1.5% maintenance in 50% O2) and injected subcutaneously with an analgesic and anti-inflammatory compound (Metacam, 2 mg/kg). The eyes were protected from drying by Cavasan eye ointment. They were moved to a stereotact with a thermal mat to keep their core temperature at 37 degrees Celsius, and the fur on their heads was removed. Once anesthesia was sufficiently deep, as indicated by the absence of a toe-pinch reflex, we applied lidocaine locally on the skin of the

head and sterilized it with 70%-ethanol or betadine. The skin was removed, the skull cleaned, and a small metal rod was fixed to the skull anterior of bregma with the use of blue-light curing dental cement. If necessary, we sutured the skin, and let the mice recover for 2-7 days.

After the mice recovered as indicated by a return to their pre-operative weight, mice were habituated to sitting head-fixed in the electrophysiology rig for 3-10 days. Once habituated, they underwent a craniotomy surgery, following the same preparatory steps as described above. Before performing the craniotomy (1.5-3 mm in diameter), we first constructed a small ring of dental cement so the brain could be bathed in saline during recordings to avoid tissue desiccation. Once the craniotomy was complete, this ring was filled with sterile silicone, and the animals were left to recover for at least 16 hours. Over the next 1-7 days, we performed repeated-insertion recordings using either NeuroNexus or Neuropixels silicon probes. For a subset of animals, we dipped the probe into DiI on the last day of recording, and perfused the animal to perform post-hoc tracing of the electrode recording locations. For the remaining recordings, the probe position was determined using anatomical landmarks (i.e., bregma and lambda).

NeuroNexus recordings were performed using either a Tucker-Davis Technologies digitizer and custom-written MATLAB code as described previously (Ahmadlou & Heimel, 2015). Neuropixels recordings were performed using a National Instruments I/O PXIe-6341 module and SpikeGLX (https://github.com/billkarsh/SpikeGLX). Visual stimulation was performed as described previously (Montijn, Meijer, et al., 2016), and synchronized with high accuracy (<1 ms) using photodiode signals that recorded visual stimulus onsets. Spikes were sorted post-hoc using Kilosort2 (https://github.com/MouseLand/Kilosort2) and only clusters of sufficient quality, as defined by Kilosort2's default threshold, were included for further analysis. High-quality clusters (i.e., putative neurons) were assigned a brain region using the AllenCCF MATLAB toolbox (https://github.com/cortex-lab/allenCCF), which automatically

calculates a neuron's anatomical position based on penetration location, angle, and depth of the silicon probe. Abbreviations for brain areas mostly follow the Allen Brain Atlas area codes. V1: primary visual cortex. AM: anteromedial visual cortex. PM: posteromedial visual cortex. LGN: lateral geniculate nucleus. LP: lateral posterior nucleus. NOT: nucleus of the optic tract. APN: anterior pretectal nucleus. SC: superior colliculus. All code used in laminar probe data acquisition and pre-processing is available online (https://github.com/JorritMontijn/Acquipix).

Visual stimulus parameters

Visual stimuli during Neuronexus and Neuropixels recordings were shown at 60 Hz on a 51 by 29 cm Dell screen at 17 cm distance from the animal's eyes, using Psychtoolbox 3 in Matlab. Drifting gratings were displayed within a 120 visual-degree diameter window with 2 visual-degree cosine edge that faded smoothly into a neutral-grey background. Drifting gratings were shown in 24 directions: [0, 15, … 345] degrees at a spatial frequency of 0.05 cycles per degree and a temporal frequency of 1 cycle per second. Natural movies were 20 s long and consisted of 4 distinct scenes taken from the BBC nature documentary Earthflight (Montijn, Meijer, et al., 2016).

Acquisition and preprocessing of retinal multi-electrode array data

Adult (>1year) zebrafish (*Danio rerio*), were dark-adapted for at least 1 h. Then under IR illumination fish were euthanized by rapid immersion in ice-cold water, the eyes removed, the retina isolated and placed photoreceptor side up on a perforated 60 electrode array (60pMEA200/30iR-Ti using a MEA2100 system: Multichannel systems, Reutlingen, Germany) in a recording chamber mounted on an Nikon Optiphot-2 upright microscope and viewed under IR with an Olympus 2x objective and video camera (Abus TVCC 20530). Room

temperature Ames' medium (Sigma-Aldrich) gassed with a mixture of O2 and CO2 at pH of 7.6 continuously superfused the MEA recording chamber.

Extracellular multiunit GC activity was recorded at 25 kHz in MC rack (Multichannel systems, Reutlingen, Germany), zero-phase bandpass filtered (250–6,250 Hz) with a fourth-order Butterworth filter in Matlab (MathWorks, Natick, MA, USA), and sorted into single-unit activity with "offline spike sorter" (Plexon, Dallas, TX, USA). Spikes were detected using an amplitude threshold > $4\sigma_n$ where σ_n is an estimation of the background noise

$$\sigma_n = median\left(\frac{|x|}{0.6745}\right) \quad 44$$

with x being the bandpass-filtered signal (Quiroga et al., 2004). The detected spikes were manually sorted into single units based on the first two principal components versus time.

The light stimulus consisted of a 500 ms full field light flash, preceded and followed by a 500- and 1000- ms period of darkness, generated using Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner et al., 2007), and repeated either 50 or 100 times. Stimuli were projected onto the retina from the photoreceptor side by a DLP projector (Light Crafter 4500, Wintech, Carlsbad, CA, USA) using a custom-built 2x water immersion objective. Only white light stimuli were used. The "dark" light intensity was 6 µW/m2, and the maximal "light" intensity was 176.2 µW/m2.

Acquisition and preprocessing of calcium imaging data

Drifting grating responses were recorded as previously described (Montijn, Meijer, et al., 2016) and putative spike times were extracted using an exponential fitting algorithm (Montijn, Olcese, et al., 2016). All code used in pre-processing of drifting grating calcium imaging data is available online (https://github.com/JorritMontijn/Preprocessing_Toolbox).

Virtual corridor experiments were performed on adult male RCFL-tdTOM x PV FlpO x VIP2r-cre (n = 1) and VIP2r-cre mice (n = 3). During water restriction, the weight of the

animals was carefully monitored and kept stable through powdered milk intake (10% diluted in drinking water) in the corridor or additionally fed solid drink. Mice were kept on a reversed day/night schedule (12h/12h) and experiments were performed in their active dark phase.

For virus injection and chronic window implantation procedures, animals were anesthetized with isoflurane at 5% for induction and 1.0-1.5% for maintenance. Metacam (1 mg/kg) was administered subcutaneously for systemic analgesia and dexamethasone (8 mg/kg) was administered to prevent brain swelling. The eyes were covered throughout the surgeries with Cavasan ointment. For virus injections, three small holes were drilled around the center of the right V1 (2.9 mm lateral from midline, 0.5 mm anterior from lambda). Animals were injected with saline-diluted AAV2/1.hSyn.GCaMP6f.WPRE.SV40 (Addgene) at a depth of 250 and 550 µm (final concentration, 10^{12} viral particles/ml; 36.8 nl/depth/location). Mice were allowed to recover for a minimum of two weeks in their home cage before window implantations. For chronic window implantations, a custom-made head bar was positioned above V1 and fixed on the skull with dental cement (Kerr). A circular 3 mm craniotomy was made on the area of injection while the dura was kept intact. A double glass window (3 + 4 mm diameter) was placed inside the craniotomy and fixed on the skull with dental cement (Kerr). Animals were allowed to recover for a minimum of one week before training began.

Before imaging, mice were trained to being head fixed and run through a virtual corridor on a custom-made treadmill (Fig. 7A). Running speed of the mice was measured with a rotary encoder and processed using an Arduino and Matlab. Absolute running speed was used to render the virtual corridor in real-time. The left half of the virtual corridor was rendered on a gamma-corrected monitor (Dell) placed under an angle of 45° with a mirror to create the perception of a symmetrical corridor. The virtual corridor was written in Matlab using OpenGL and Psychophysics Toolbox 3 and contained a 100 cm black and white Gaussian noise texture with overlying visual stimuli. We used three vertical gratings and three checkerboard stimuli

positioned 11 cm apart, at locations 22-77 cm (Fig. 7B). After completing a run in the corridor, mice were immediately shown a luminance-matched gray screen followed by a 0.5 s auditory cue (8 kHz) 1 second later. Two seconds after this cue, the mice received a 5-10 µl milk reward. Mice were trained for ~10 sessions until they completed up to ~150 trials during imaging sessions.

Imaging was performed with a two-photon microscope (Neurolabware) equipped with a Ti-sapphire laser (Mai-Tai 'Deepsee', Spectraphysics; wavelength, 920 nm) and a 16x, 0.8 NA water immersion objective (Nikon). The microscope was controlled by Scanbox (Neurolabware) running on Matlab. One one mouse we performed dual-plane imaging at 15.5 Hz/plane using an electrically tunable lens (OptoTune). The other three mice were imaged in a single plane at 15.5 Hz. During the pre-processing stage, we discarded all interneurons and only included putative pyramidal cells for further analysis.

Analysis of virtual corridor joint-feature encoding

For each neuron, we calculated ZETA-scores after aligning the putative spike times to trial start ('time-aligned'), converting spike times to locations and aligning them to the beginning of the corridor ('location-aligned'), or aligning the spike times to a visuomotor mismatch event, where rendering of the virtual corridor was paused for 500 ms ('mismatch-aligned') (fig. 7C). We call the resulting values time-modulation, location-modulation and mismatch-modulation respectively. In the case of the spatial "location-aligned" analysis, we furthermore discarded the first and last 10% of the track to avoid including the start and end box locations. We investigated whether there was a relationship between the modulation values that neurons showed for these three different features in two ways. First, we simply calculated the Pearson correlation at the level of a single recording between time/location, time/mismatch, and location/mismatch modulations (fig. 7D-F). A one-sample t-test on these correlation values

showed there was a significant, positive correlation for time/location and a significant, negative correlation for location/mismatch. However, this analysis does not directly answer the question whether the joint encoding of two features at the level of single neurons is more or less likely than we would expect by chance.

We therefore z-scored the modulation values per recording and pooled all neurons. We then used a kernel-density estimator (KDE) to construct a probability density estimate for the distribution of modulation values for time, location, and mismatch (Figure 7 – Figure Supplement 1). Combining these distributions for a pair of features, we obtain a null-hypothesis distribution of what the joint-feature-encoding distribution would look like if the modulation-scores are independent of each other. Using the same bandwidths as for the single-feature KDE-distributions, we smoothed the real data so we could directly compare which regions are over- or under-represented in the real data, producing the heat maps in fig. 7G-I. To quantify this over/underrepresentation, we counted the number of neurons in the upper-right quadrant, where neurons lie that have high modulation scores for both features. We compared the real count to the distribution obtained from 10,000 random samples taken from the joint-feature KDE-derived null-hypothesis distribution, where there is no correlation between the two feature modulation scores. As shown in fig. 7G-I and fig. 7 – Supplement 1, time and location-modulation scores were more often both high in single neurons than expected from chance. Moreover, there were fewer neurons that showed a joint encoding of both spatial location and visuomotor mismatch than expected from chance.

VIP-optogenetics analysis of Allen Brain Institute Ecephys data

Detailed information on experimental and data acquisition procedures can be found online at the Allen Brain Institute website: https://portal.brain-map.org/explore/circuits/visual-coding-neuropixels. We used data from 5 Vip-IRES-Cre; Ai32 mice that underwent laser-based

optogenetic stimulation. We pre-selected 1706 clusters that were recorded in visually-responsive cortex (AL, AM, PM, L, V1, RL, or MMP) and were of sufficient quality, specifically: KiloSort2 tagged the cluster as “good”, ISI violations were under 0.5, amplitude thresholds under 0.1, and presence ratios over 0.9. The next step was to cull this population to only cells that showed modulated spiking with respect to the onset of optogenetic stimulation. We therefore included N=1144 cells that showed $p<0.05$ with a ZETA-test on the window between (-0.5, +0.5 s) after optogenetic stimulation. We calculated the peak-response latency using our IFR method and discarded cells with peak responses earlier than 1 ms after optogenetic stimulation onset (N=909). The remaining cells were classified as VIP if their IFR peak latency was < 10 ms (N=13), as Inhibited if their firing rate was significantly lower during 10-30 ms after optogenetic stimulation than during the 50 ms preceding optogenetic onsets (N=59), as Activated if their rate was significantly higher during 10-30 ms after optogenetic onset than during the pre-stimulus baseline (N=137), and otherwise as Other (N=700). The large size of this latter group could be explained by many cells being visually-responsive to the blue laser light. We chose a 10-30 ms window to compare the IFR peak/trough latencies of Activated and Inhibited cells, as 10 ms was the time of laser offset, and visual responses start to emerge in visual cortex after about 30-50 ms.

Acknowledgments

JSM was funded by a grant awarded to JAH by the Stichting Vrienden van het Herseninstituut. Brain images in figures 3-5 were generated using Brain Explorer 2 (Lau et al., 2008). We thank the Allen Brain Institute for their openly accessible data sets. We also thank the engineers of the mechatronics workshop at the NIN.

Author contributions

JSM designed ZETA, performed the neuropixels recordings, and performed the analyses. KS and JSM performed the GCaMP6 recordings. JSM and JLC performed the neuronexus recordings. MHH performed the retinal recordings. JSM and JAH refined ZETA, performed the mathematical derivations, and wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

Declaration of Interests

The authors declare no conflicts of interest exist.

References

Ahmadlou, M., & Heimel, J. A. (2015). Preference for concentric orientations in the mouse superior colliculus. *Nature Communications*, *6*, ncomms7773. https://doi.org/10.1038/ncomms7773

Attinger, A., Wang, B., & Keller, G. B. (2017). Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. *Cell*, *169*(7), 1291-1302.e14. https://doi.org/10.1016/j.cell.2017.05.023

Bartolo, R., Saunders, R. C., Mitz, A. R., & Averbeck, B. B. (2020). Information-Limiting Correlations in Large Neural Populations. *Journal of Neuroscience*, *40*(8), 1668–1678. https://doi.org/10.1523/JNEUROSCI.2072-19.2019

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Calders, T., & Jaroszewicz, S. (2007). Efficient AUC Optimization for Classification. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Knowledge Discovery in Databases: PKDD 2007* (pp. 42–53). Springer Berlin Heidelberg.

Chen, L., Deng, Y., Luo, W., Wang, Z., & Zeng, S. (2009). Detection of bursts in neuronal spike trains by the mean inter-spike interval method. *Progress in Natural Science*, *19*(2), 229–235. https://doi.org/10.1016/j.pnsc.2008.05.027

Cooper, D. C. (2002). The significance of action potential bursting in the brain reward circuit. *Neurochemistry International*, *41*(5), 333–340. https://doi.org/10.1016/s0197-0186(02)00068-2

Douglas, R. J., Martin, K. A., & Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Computation*, *1*(4), 480–488.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator:L2 theory. *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, *57*(4), 453–476. https://doi.org/10.1007/BF01025868

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Gregory, R. L., Longuet-Higgins, H. C., & Sutherland, N. S. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *290*(1038), 181–197. https://doi.org/10.1098/rstb.1980.0090

Gumbel, E. J. (1941). The Return Period of Flood Flows. *The Annals of Mathematical Statistics*, *12*(2), 163–190. JSTOR.

Gumbel, E. J. (1954). *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. U.S. Government Printing Office.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, *13*(3). https://doi.org/10.1371/journal.pbio.1002106

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., … Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236. https://doi.org/10.1038/nature24636

Kass, R. E., Eden, U., & Brown, E. (2014). *Analysis of Neural Data*. Springer-Verlag. https://doi.org/10.1007/978-1-4614-9602-1

Kass, R. E., & Ventura, V. (2001). A spike-train probability model. *Neural Computation*, *13*(8), 1713–1720. https://doi.org/10.1162/08997660152469314

Keller, G. B., Bonhoeffer, T., & Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, *74*(5), 809–815. https://doi.org/10.1016/j.neuron.2012.03.040

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Knöpfel, T., & Song, C. (2019). Optical voltage imaging in neurons: Moving from technology development to practical tool. *Nature Reviews Neuroscience*, *20*(12), 719–727. https://doi.org/10.1038/s41583-019-0231-4

Kok, P., Bains, L. J., van Mourik, T., Norris, D. G., & de Lange, F. P. (2016). Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. *Current Biology*, *26*(3), 371–376. https://doi.org/10.1016/j.cub.2015.12.038

Lau, C., Ng, L., Thompson, C., Pathak, S., Kuan, L., Jones, A., & Hawrylycz, M. (2008). Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, *9*(1), 153. https://doi.org/10.1186/1471-2105-9-153

Leinweber, M., Ward, D. R., Sobczak, J. M., Attinger, A., & Keller, G. B. (2017). A Sensorimotor Circuit in Mouse Cortex for Visual Flow Predictions. *Neuron*, *95*(6), 1420-1432.e5. https://doi.org/10.1016/j.neuron.2017.08.036

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., & Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, *522*(1), 225–259. https://doi.org/10.1002/cne.23458

Mattia, D., Hwa, G. G., & Avoli, M. (1993). Membrane properties of rat subicular neurons in vitro. *Journal of Neurophysiology*, *70*(3), 1244–1248. https://doi.org/10.1152/jn.1993.70.3.1244

Mazurek, M., Kager, M., & Van Hooser, S. D. (2014). Robust quantification of orientation selectivity and direction selectivity. *Frontiers in Neural Circuits*, *8*. https://doi.org/10.3389/fncir.2014.00092

Mesa, N., Waters, J., & Vries, S. E. J. de. (2021). The Effect of Inclusion Criteria on the Functional Properties Reported in Mouse Visual Cortex. *ENeuro*, *8*(1). https://doi.org/10.1523/ENEURO.0188-20.2021

Montijn, J. S., Meijer, G. T., Lansink, C. S., & Pennartz, C. M. A. (2016). Population-Level Neural Codes Are Robust to Single-Neuron Variability from a Multidimensional Coding Perspective. *Cell Reports*, *16*(9), 2486–2498. https://doi.org/10.1016/j.celrep.2016.07.065

Montijn, J. S., Olcese, U., & Pennartz, C. M. A. (2016). Visual Stimulus Detection Correlates with the Consistency of Temporal Sequences within Stereotyped Events of V1 Neuronal Population Activity. *The Journal of Neuroscience*, *36*(33), 8624–8640. https://doi.org/10.1523/JNEUROSCI.0853-16.2016

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology*, *20*(4), 408–434. https://doi.org/10.1152/jn.1957.20.4.408

Niell, C. M., & Stryker, M. P. (2008). Highly Selective Receptive Fields in Mouse Visual Cortex. *Journal of Neuroscience*, *28*(30), 7520–7536. https://doi.org/10.1523/JNEUROSCI.0623-08.2008

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., & Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, *466*(7306), 617–621. https://doi.org/10.1038/nature09178

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by VI? *Vision Research*, *37*(23), 3311–3326.

Ooyen, A., & Roelfsema, P. (2003). A biologically plausible implementation of error-backpropagation for classification tasks. In *Artificial Neural Networks and Neural Information Processing Supplementary Proceedings ICANN/ICONIP 2003*.

Packer, A. M., Russell, L. E., Dalgleish, H. W. P., & Häusser, M. (2015). Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature Methods*, *12*(2), 140–146. https://doi.org/10.1038/nmeth.3217

Palm, G., Aertsen, A. M. H. J., & Gerstein, G. L. (1988). On the significance of correlations among neuronal spike trains. *Biological Cybernetics*, *59*(1), 1–11. https://doi.org/10.1007/BF00336885

Pfeffer, C. K., Xue, M., He, M., Huang, Z. J., & Scanziani, M. (2013). Inhibition of inhibition in visual cortex: The logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, *16*(8), 1068–1076. https://doi.org/10.1038/nn.3446

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., & Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, *503*(7477), 521–524. https://doi.org/10.1038/nature12676

Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., & Roelfsema, P. R. (2012). The Role of Attention in Figure-Ground Segregation in Areas V1 and V4 of the Visual Cortex. *Neuron*, *75*(1), 143–156. https://doi.org/10.1016/j.neuron.2012.04.032

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*(8), 1661–1687. https://doi.org/10.1162/089976604774201631

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation Selectivity in Macaque V1: Diversity and Laminar Dependence. *The Journal of Neuroscience*, *22*(13), 5639–5651.

Saleem, A. B., Ayaz, A., Jeffery, K. J., Harris, K. D., & Carandini, M. (2013). Integration of visual motion and locomotion in mouse visual cortex. *Nature Neuroscience*, *16*(12), 1864–1869. https://doi.org/10.1038/nn.3567

Scott, D. W. (2009). Sturges' rule. *WIREs Computational Statistics*, *1*(3), 303–306. https://doi.org/10.1002/wics.35

Self, M. W., van Kerkoerle, T., Goebel, R., & Roelfsema, P. R. (2019). Benchmarking laminar fMRI: Neuronal spiking and synaptic activity during top-down and bottom-up processing in the different layers of cortex. *NeuroImage*, *197*, 806–817. https://doi.org/10.1016/j.neuroimage.2017.06.045

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. *Neuron*, *102*(1), 249-259.e4. https://doi.org/10.1016/j.neuron.2019.01.026

Shimazaki, H., & Shinomoto, S. (2007). A Method for Selecting the Bin Size of a Time Histogram. *Neural Computation*, *19*(6), 1503–1527. https://doi.org/10.1162/neco.2007.19.6.1503

Shuler, M. G., & Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science (New York, N.Y.)*, *311*(5767), 1606–1609. https://doi.org/10.1126/science.1123513

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., Groblewski, P. A., Ahmed, R., Arkhipov, A., Bernard, A., Billeh, Y. N., Brown, D., Buice, M. A., Cain, N., Caldejon, S., … Koch, C. (2019). A survey of spiking activity reveals a functional hierarchy of mouse corticothalamic visual areas. *BioRxiv*, 805010. https://doi.org/10.1101/805010

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., & Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, *576*(7786), 266–273. https://doi.org/10.1038/s41586-019-1787-x

Vinje, W. E., & Gallant, J. L. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, *287*(5456), 1273–1276. https://doi.org/10.1126/science.287.5456.1273

Whittington, J. C. R., & Bogacz, R. (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

Xing, D., Ringach, D. L., Shapley, R., & Hawken, M. J. (2004). Correlation of local and global orientation and spatial frequency tuning in macaque V1. *The Journal of Physiology*, *557*(3), 923–933. https://doi.org/10.1113/jphysiol.2004.062026

Figures, legends and tables

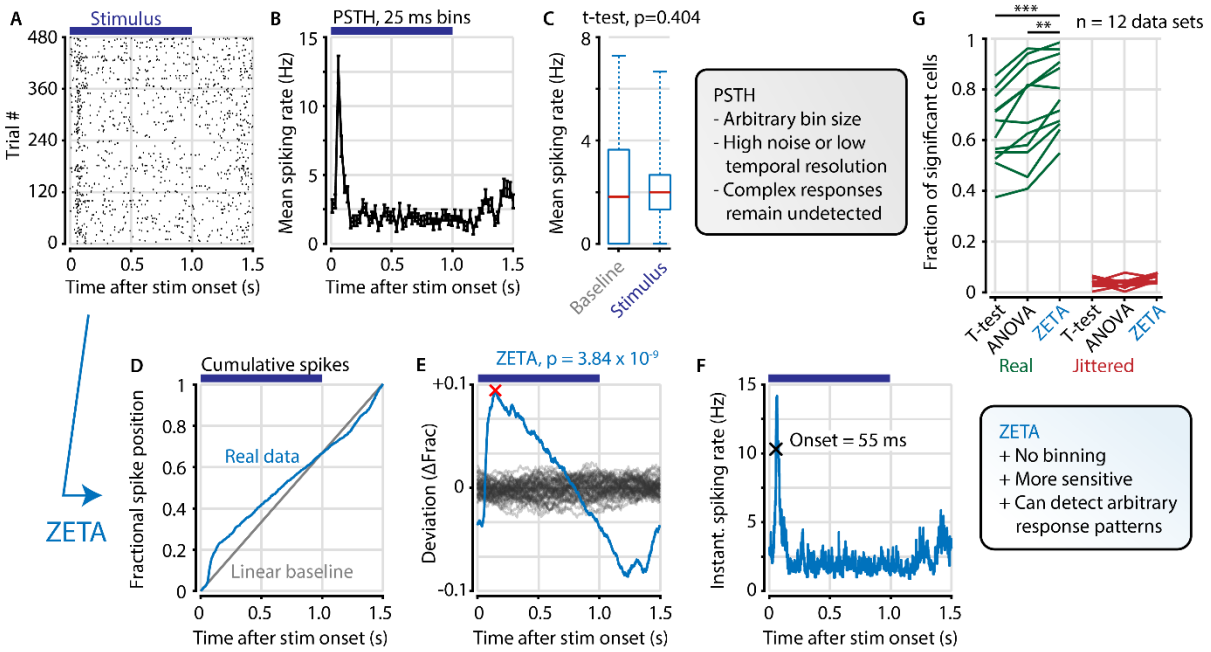


Figure 1. The ZETA-test detects an example visually responsive neuron in V1. A,B) Raster plot (A) and PSTH (B) for a neuron that shows an onset peak and a reduced sustained spiking in response to a visual stimulus (purple bars). C) A common approach for computing a neuron's responsiveness is to perform a t-test on the average activity per trial during stimulus presence (0 – 1s) and absence (1.0 – 1.5s), which fails to detect this neuron's response (paired t-test, p=0.404, n=480 trials). D) ZETA avoids binning, by using the spike times to construct a fractional position for each spike (blue) and compares this with a null-distribution of a stationary rate (grey). E) The difference between the real and null curves gives a deviation from expectation (blue), where the most extreme value is defined as the Zenith of Event-based Time-locked Anomalies (ZETA, red cross). To compute its statistical significance (here, $p=3.84 \times 10^{-9}$), we compare ZETA to the variability over repeats of the procedure with jittered onsets (grey curves). F) A ZETA-derived instantaneous spiking rate allows a reliable estimation of response onset. G) At a significance threshold of $\alpha=0.05$, ZETA detects more stimulus-responsive cells than both a mean-rate t-test (***: paired t-test, $p = 2.8 \times 10^{-7}$, n=12 data sets) and an ANOVA over bins using an optimal bin width (**: p=0.0014).

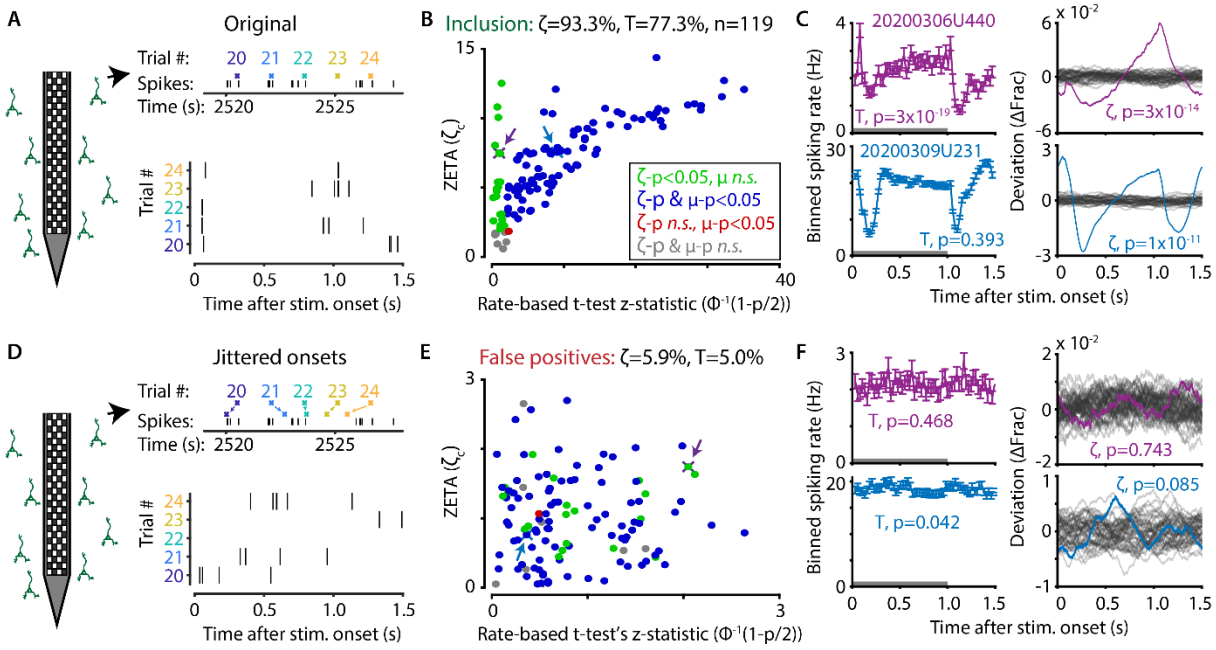


Figure 2. The ZETA-test outperforms the t-test for V1 neurons (n=119). A) Spikes were recorded with neuropixels in mouse V1 and aligned with the onset of square-wave drifting gratings. B) Applying the ZETA-test, we found that 93.3% of all V1 neurons showed significant firing rate modulations that were time-locked to stimulus onset. Using a rate-based t-test between stimulation (0 - 1s) and no stimulation (1 – 1.5s) epochs across trials registered 77.3% of neurons to be visually responsive. Neurons detected by only ZETA are green, by both methods are blue, by only a mean-rate t-test are red, and by neither are grey. Arrows indicate the neurons shown in C. C) Two example neurons that are 1) detected by a rate-based t-test as well as ZETA (top) and 2) not detected by a rate-based t-test, but only by ZETA (bottom). D) We investigated the false-positive rate of both approaches by jittering the onsets of visual stimuli; this preserved the temporal structure of the spiking response, but destroys the time-locked modulations in activity. E-F) same cells and analyses as in B-C but for jittered onsets. Red indicates neurons included by ZETA, but not a t-test; green indicates neurons included by a t-test, but not ZETA. As expected, the percentage of false positives (i.e., neurons with ζ_c/z-statistic > 2.0) was around 5% (α=0.05) for both approaches. Note the change in axis magnitude from B to E.

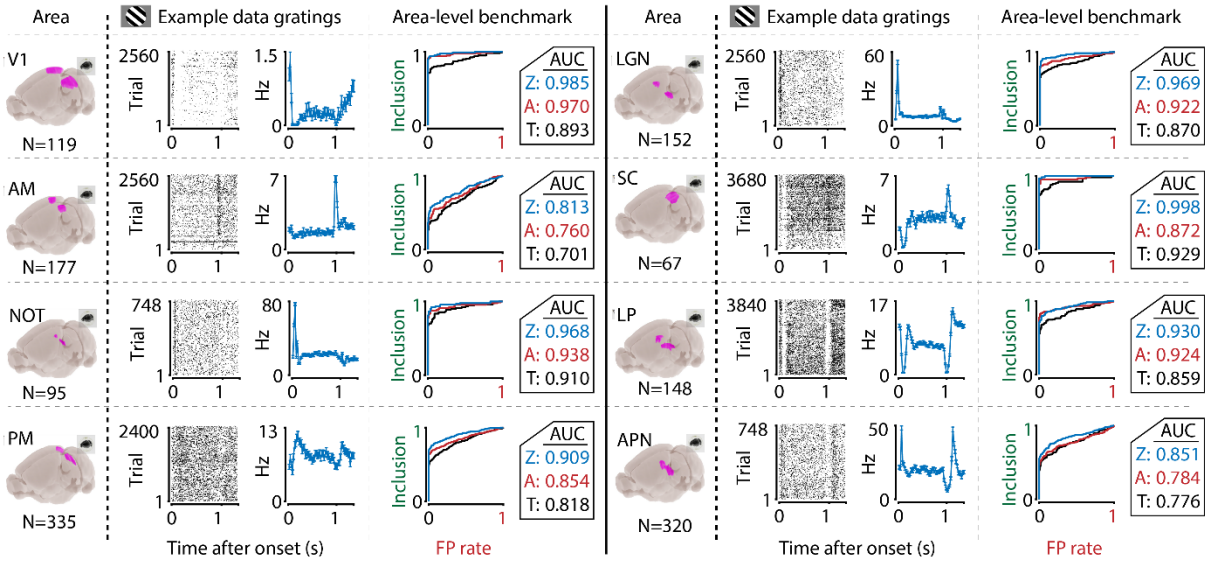


Figure 3. The ZETA-test's superior sensitivity is independent of brain area. We recorded neuronal responses to drifting gratings (1s with 500ms blank ITIs) from various visual brain areas: V1, SC, AM, LP, NOT, APN, PM, and LGN. For each area, we show an example neuron's raster plot (left) and binned responses (right). All cells depicted here were significant using ZETA. Area-level benchmark summaries show ROC analyses of the inclusion rate (y-axis) and false-positive (FP) rate (x-axis) for ZETA (Z, blue), bin-wise ANOVA (A, red), and rate-based t-tests (T, black). In all cases, the AUC of the ZETA-test exceeded both the ANOVA's AUC and t-test's AUC.

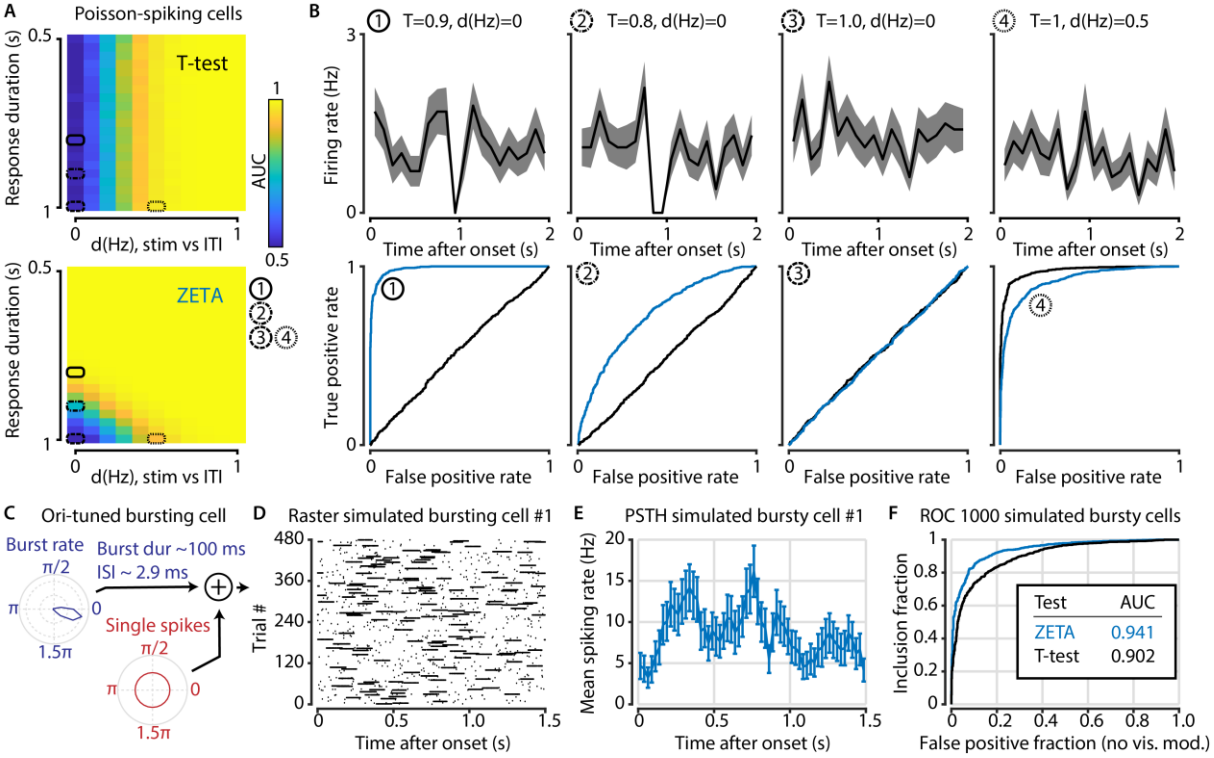


Figure 4. The t-test exceeds the ZETA-test's performance only in the hypothetical case of purely Poisson-distributed spike counts and only for a small range of spiking rate differences. A) Top: the ability of the t-test to differentiate simulated stimulus (non)modulated cells depends only the total spike count during stimulus (0 – 1 s) and inter-trial interval (ITI, 1 – 2 s) periods (x axis; firing rate difference *d(Hz)*), but not on the duration of the cell's response when keeping the spike count constant (y axis; response duration *T*). Bottom: in contrast, the ZETA-test can differentiate stimulus modulation using either variable. B) Top: example PSTHs of single cells corresponding to the markers 1-4 in (A). Bottom: ROC curves for the combination of variables marked as 1-4 in (A). The ZETA-test always exceeds the t-test's performance, except for a limited range where the response duration is 1 second. C) A more biologically plausible test case is bursting neurons that have an elevated probability of bursting during stimuli. D-E) This simulation produces no "onset" peaks of activity that the ZETA-test can exploit. F) However, despite the lack of clear peaks of activity, the ZETA-test exceeds the t-test's ability to detect stimulus-modulated bursting cells (ZETA AUC=0.941, t-test AUC=0.902).

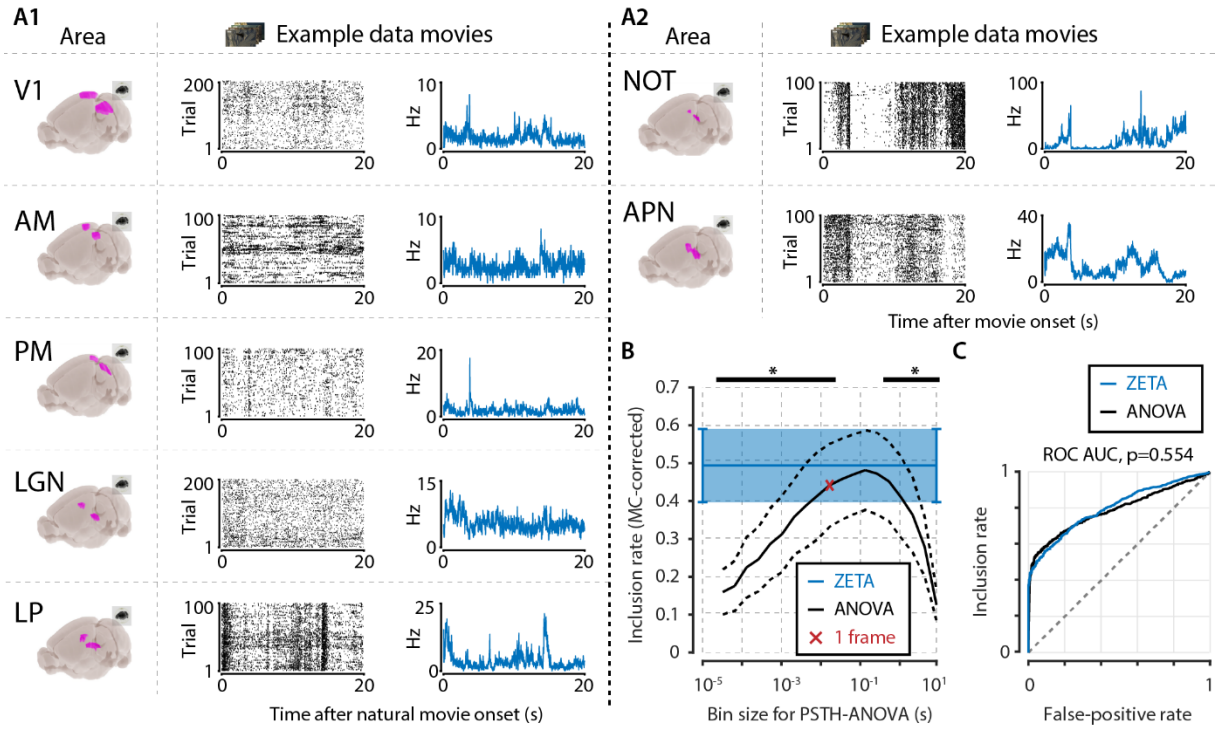


Figure 5. Determining neuronal responsiveness to natural movies. A) We recorded neuronal responses to natural movies (4 scenes that repeated every 20 seconds) from various visual brain areas: V1, AM, PM, LGN, LP, NOT, and APN. For each area, we show an example neuron's raster plot (left) and binned responses (right). All cells depicted here were significant using ZETA. B) We determined a neuron's responsiveness using a 1-way ANOVA over all binned responses (i.e., PSTH-level ANOVA) for various bin sizes and MC-corrected (black) and using the ZETA-test (blue). Note that the ZETA-test is timescale-free and plotted at all x-values only for easy comparison with the ANOVAs. Curves show mean ± SEM over brain regions (n=7). ZETA shows an inclusion rate similar to the most optimal bin sizes (0.0333s – 0.2667s), and significantly higher than bin sizes of 0.0167s or shorter, as well as 0.533s or longer (*, FDR-corrected paired t-tests, $p<0.05$). C) An ROC analysis on all n=977 cells showed that the ZETA-test and a combined set of ANOVAs were similarly sensitive (ZETA-test AUC: 0.792 ± 0.011; ANOVAs AUC: 0.798 ± 0.010; z-test, p=0.554).

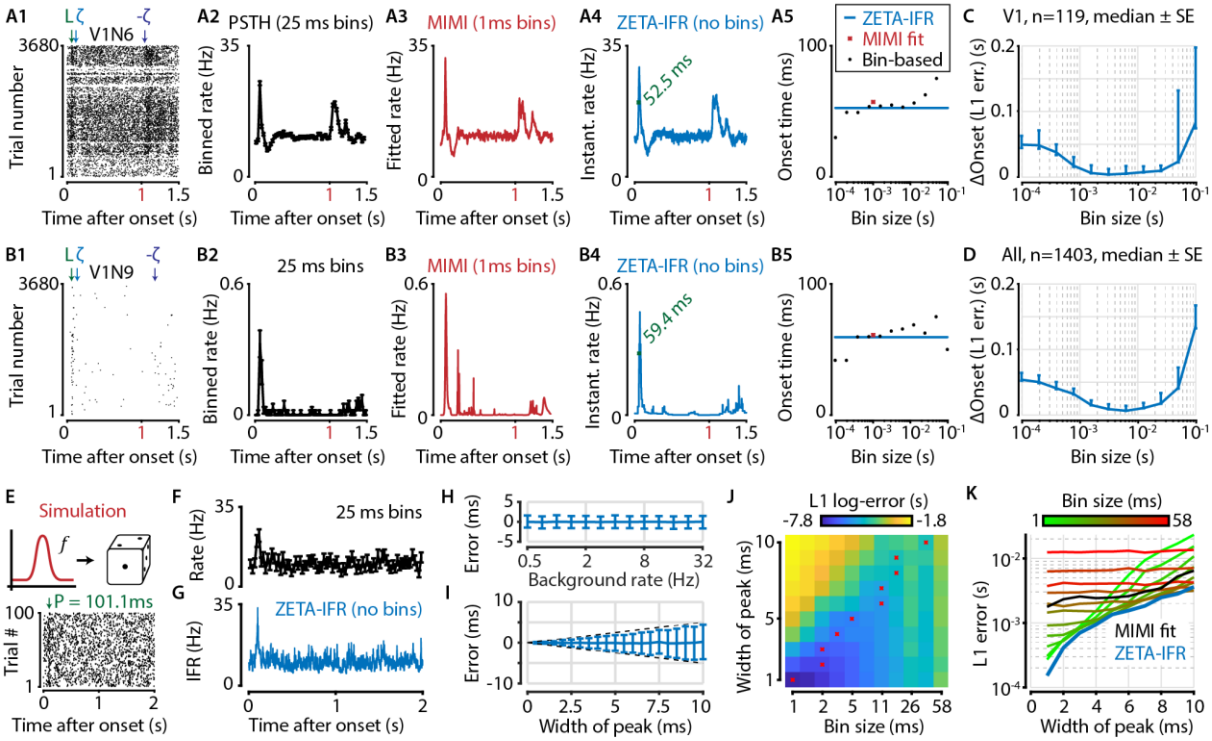


Figure 6. A ZETA-derived measure for instantaneous firing rates (IFR) avoids binning and allows more accurate latency determination than with PSTHs. A,B) Responses of two example V1 cells to drifting gratings. From left to right: 1) Raster plots showing the estimated onset latency ('L', green), and times of ZETA (blue) and –ZETA (purple). 2) Spiking rates using 25 ms bins. 3) Spiking rates using multiplicative inhomogeneous Markov interval (MIMI) model-based fits. 4) Binning-free instantaneous spiking rates provide a much higher temporal resolution. Using the first crossing of half-peak firing rates, we determined the onset latencies of these cells to be 52.5 and 59.4 ms. 5) Estimated onset times using our method for instantaneous firing rates (blue), MIMI model-based fits (red), or PSTHs with bin widths from 0.1 to 100 ms (black). Onset latency estimates depend on the chosen bin size, and the optimal size varies across cells. C-D) The median (±SE) difference in onsets estimated by our instantaneous firing rates compared to that of various sized bins for V1 cells (C, n=119) and for cells from all brain regions (D, n=1403). Both C and D show the onsets estimated by the two methods were most similar for bin sizes between 1 – 10 ms. E-K) Benchmarking of peak detection using artificial Poisson neurons that show a transient peak. E) Example Poisson cell

for a background rate of 10Hz and a peak-width of 10 ms. With the true peak at 100 ms, the estimation error here was 1.1 ms. F) Binning the cell's spiking response in 25-ms bins reduces the peak height and temporal precision. G) Our instantaneous spiking rate preserves a sharper peak response and allows for a temporally accurate latency estimation. H) The detection of peak latencies is insensitive to realistic levels of a stationary Poisson background firing rate (13 base rates, 0.5 – 32 Hz). I) The mean error is unbiased, and the standard deviation in the onset peak latency estimate scales linearly with the width of the peak. Dotted lines show the real peak width. Graphs in H-I show mean ± sd. J) The error in peak latency estimation depends on both the bin width and the width of the neuron's peak response. Red crosses indicate the bin size with the lowest error for a given peak width. K) Plotting the latency estimation error shows that different bin sizes (red-green) are optimal for different peak-widths. The accuracy of the latency obtained from MIMI-based fits (black) is less sensitive to the peak width, but never performs as well as well as the most optimal bin size. The error based on our binning-less IFR (blue) is as at least as low as the most optimal bin size, for any peak width.

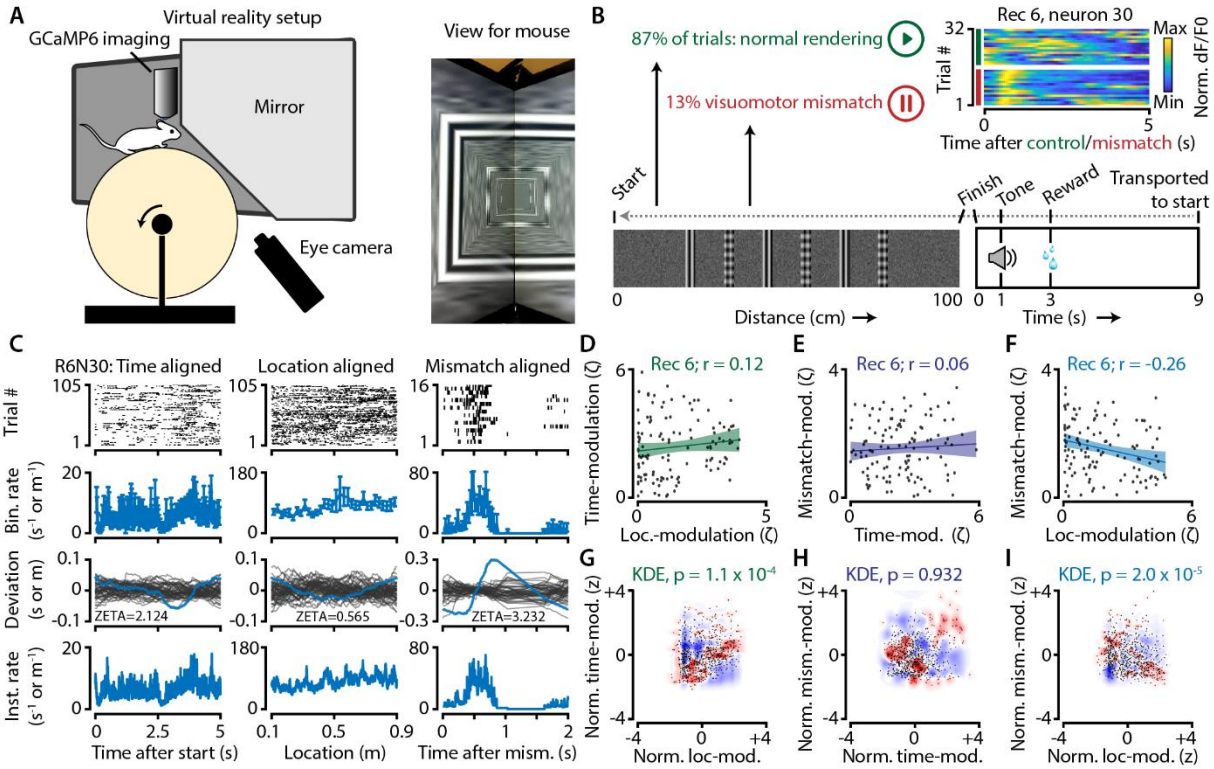


Figure 7. Neurons in V1 encode either visuomotor mismatch signals or spatial location. A) Schematic of setup showing mouse on running wheel (lhs) viewing a virtual tunnel (rhs). B) Trials consist of a 100cm linear track. One second after the mice ran to the end of the tunnel, an auditory stimulus signaled that a water reward would be delivered two seconds later. 6s after reward delivery, mice were transported to the start of the virtual tunnel. In a subset of trials, the rendering of the tunnel was paused at a random location, eliciting a visuomotor mismatch signal. Top right shows calcium imaging data for an example "mismatch neuron" during 16 control and 16 mismatch trials. C) Spiking data for example neuron obtained from exponential fits of the dF/F0 signals. Putative spikes were aligned to start (left), location of the animal on the track (middle), or mismatch onset (right). From top to bottom: raster plot of putative spike times; mean +/- SEM of firing rates over trials (n=105 trials, of which n=16 mismatch trials); spiking deviation underlying ZETA; instantaneous firing rate. D-I) Relationship between time-, location-, and mismatch-modulation. One point is one neuron. D-F) ZETA-scores for example recording 6 (N=120 neurons). G-I) Analysis using a kernel-density estimate (KDE) to test

whether joint-encoding of two features is more common than expected by chance (see figure supplement 1). G) More neurons showed joint-encoding of both spatial and temporal location than expected by chance ($p=1.1 \times 10^{-4}$). H) Joint-encoding of spatial location and mismatch was not significantly different from chance ($p=0.932$). I) Location on the virtual track and visuomotor mismatch are less likely to be encoded by the same neuron than expected from chance ($p=2.0 \times 10^{-5}$). See Figure 7 – figure supplement 1 for more details on the KDE procedure.

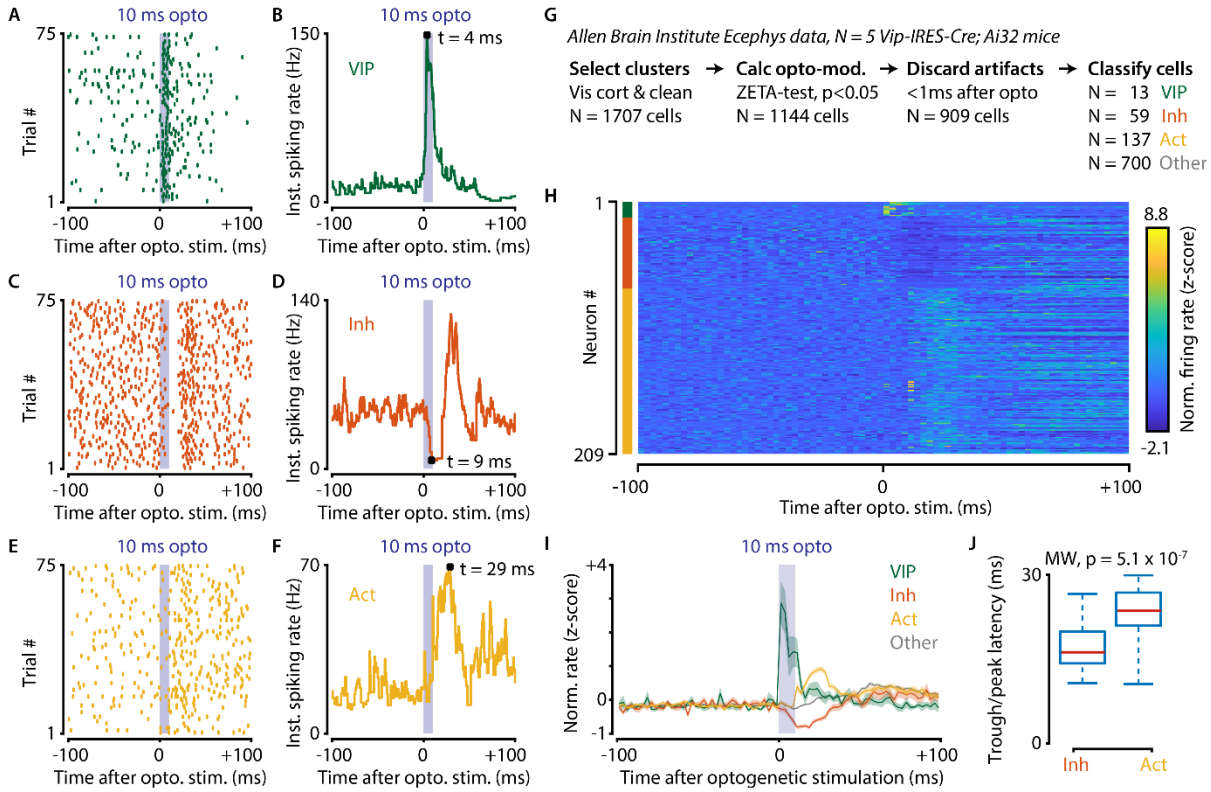
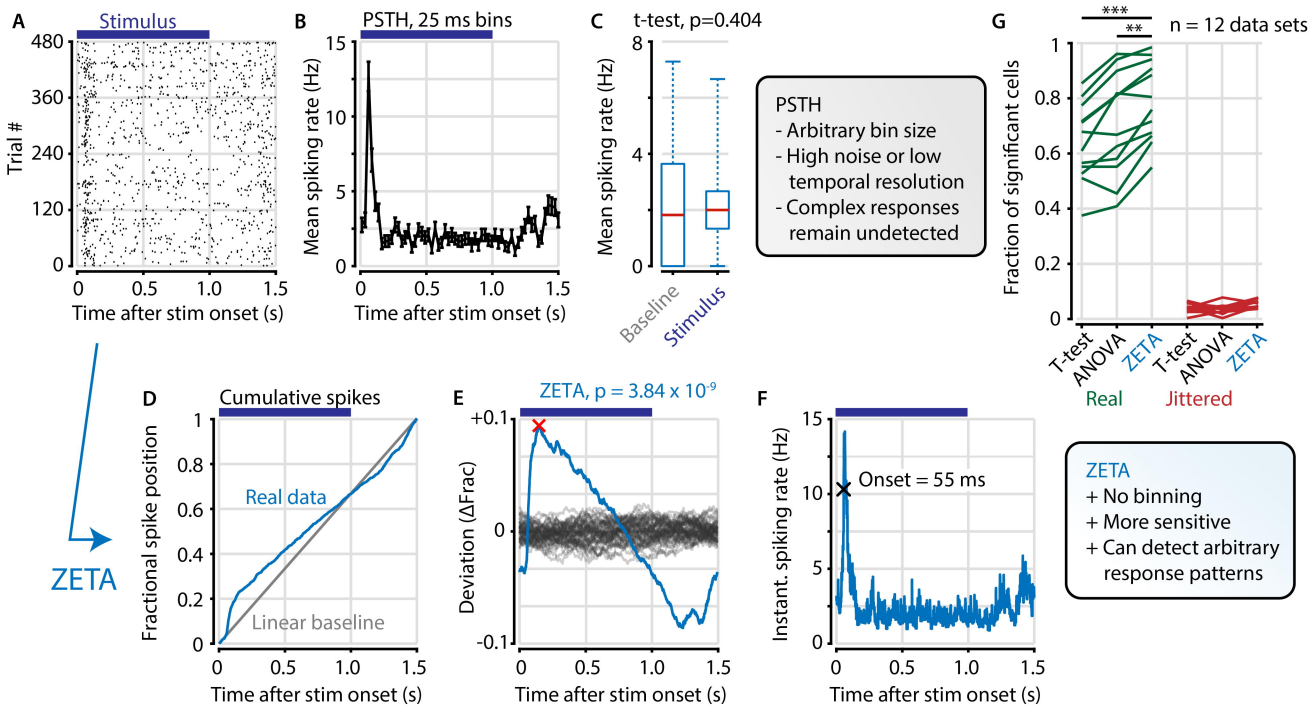
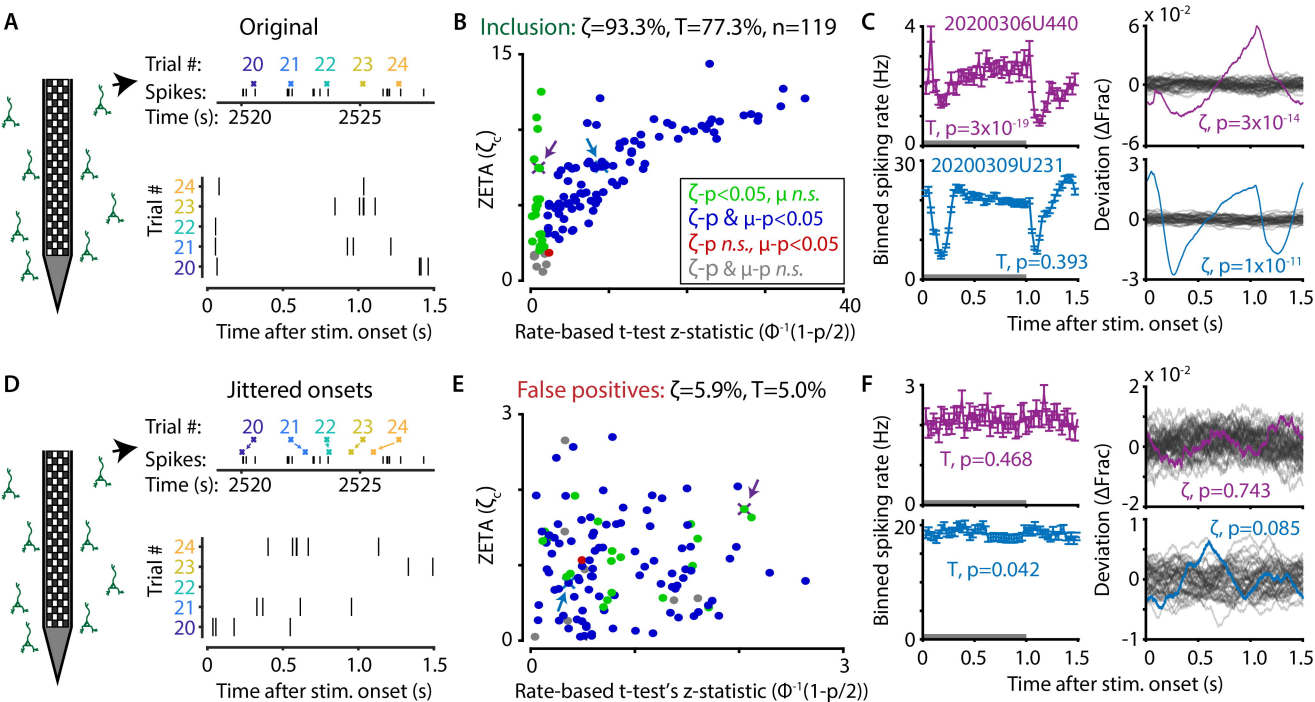


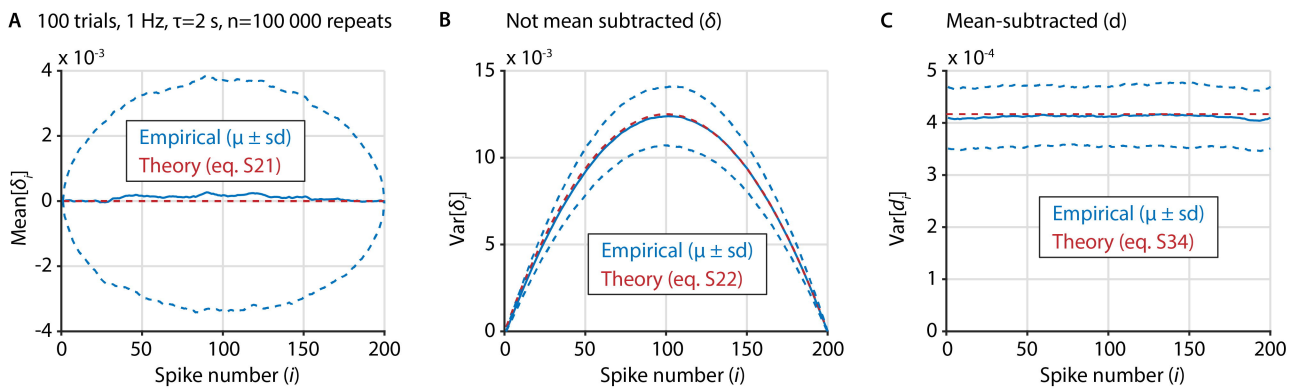
Figure 8. Optogenetic stimulation of VIP-expressing cells in mouse visual cortex causes short-latency inhibition and longer-latency disinhibition of the local neural circuit. A-F) Response of example cells classified as VIP (A,B), Inhibited (C,D) and Activated (E,F). G) Data were recorded in visual cortex from 5 Vip-Cre mice at the Allen Brain Institute. Cells were only included if the clustering quality was sufficient (1707 cells total). A ZETA-test included cells that were modulated within (-0.5, +0.5 s) after optogenetic stimulation (N=1144 cells). IFR peak- and trough-latency was computed and cells were discarded if their peak was earlier than 1 ms after optogenetic stimulation onset. Remaining cells were classified as VIP (N=13), Inhibited (N=59), Activated (N=137), or Other (N=700). H) Heat map showing normalized firing rate of VIP (top), Inhibited (middle), and Activated (bottom) cells. I) Mean ± SEM of PSTH (2.5 ms bin size) over all VIP (green), Inhibited (orange), Activated (yellow), and Other (grey) cells. J) Inhibited cells showed significantly lower mean IFR-peak latencies after optogenetic stimulation than Activated cells (Inh: 17.4 ms; Act: 23.0 ms; Mann-Whitney U-test, $p=5.1 \times 10^{-7}$).

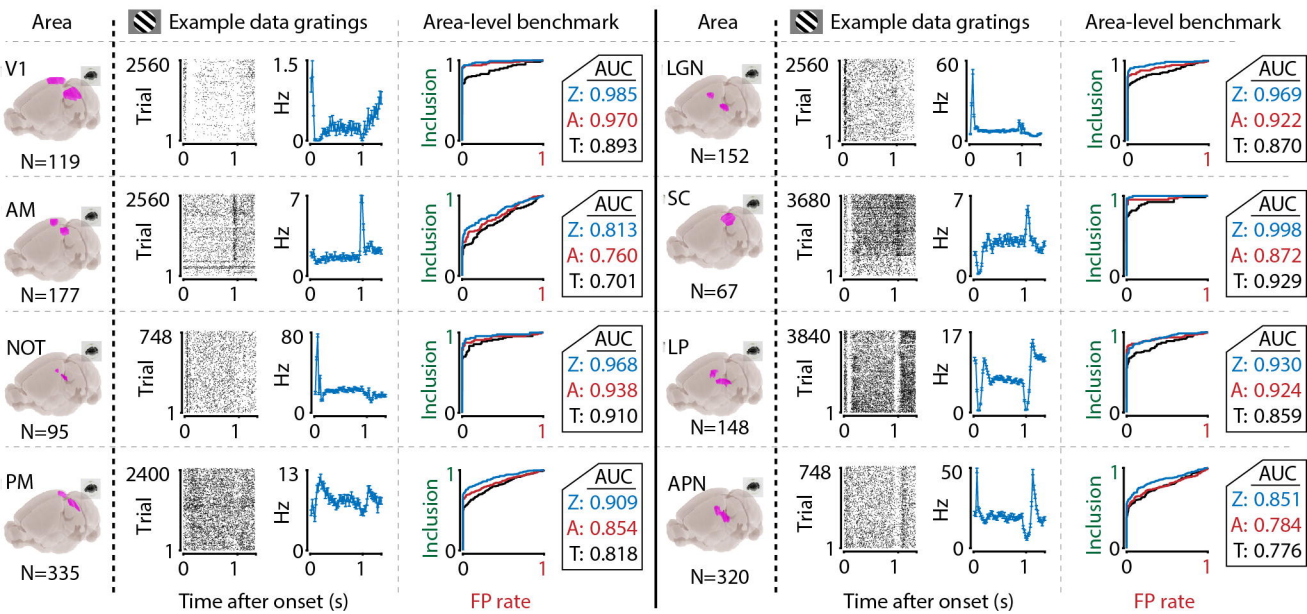
Table 1. Parameters of bursting neurons used in Figure 4. Abbreviations and mathematical symbols are as follows: ISI=Inter-spike interval; IBI=Inter-burst interval; Exp=exponential distribution; |x| = absolute of x; $\mathcal{N}$=standard normal distribution; U(x,y)=uniform distribution on interval [x,y]; $\mathcal{M}$=von Mises distribution; Γ=Gamma distribution.

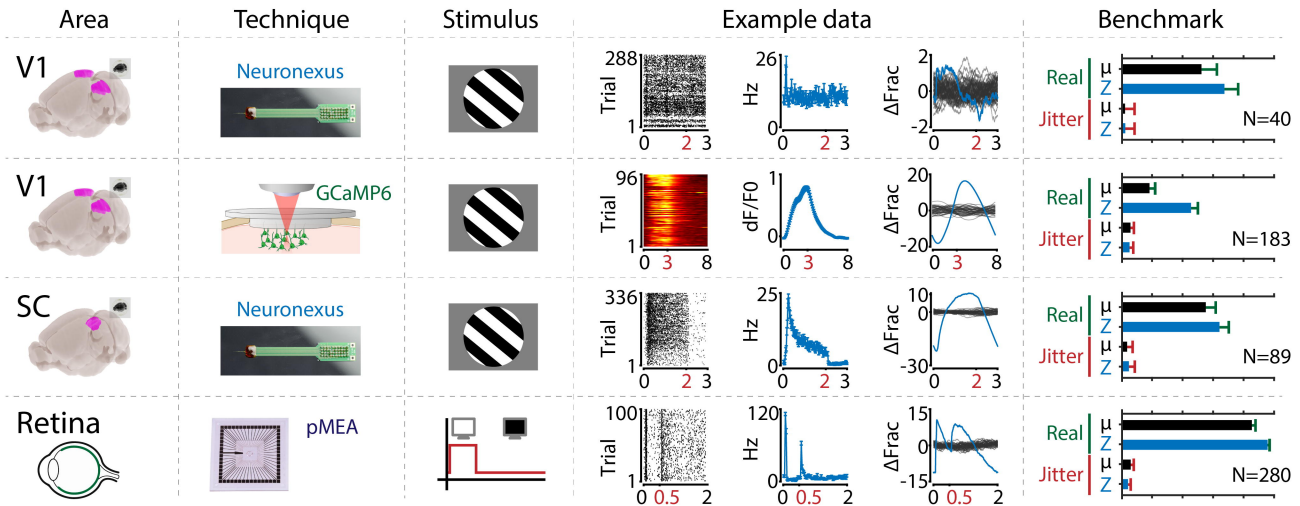
Property	Unit	Distributed as	Sampled from:
Single-spike ISI	s	$\mathrm{Exp}(1/r)$	$r \sim \mathrm{Exp}(\lambda=1)$
Baseline IBI	s	$\mathrm{Exp}(1/\lvert R_b \rvert)$	$R_b \sim \lvert\mathcal{N}\rvert/20+1/80$
Preferred orientation IBI	s	$\mathrm{Exp}(1/\lvert R_t \rvert)$	$R_t \sim \lvert\mathcal{N}\rvert+1/4$
Preferred orientation	rad	θ_p	$\theta_p \sim U(0,2\pi)$
Orientation-tuned bursting	Hz	$1/R_b + 1/R_t \cdot \mathcal{M}(\theta_p, \kappa)$	$\kappa \sim 5+U(0,5)$
Burst duration	ms	$\Gamma(2*k, \theta=0.5)$	$k \sim 90+10*\mathcal{N}$
ISI in bursts	ms	$\Gamma(2*k, \theta=0.5)$	$k \sim 0.5+\mathrm{Exp}(\lambda=2.4)$

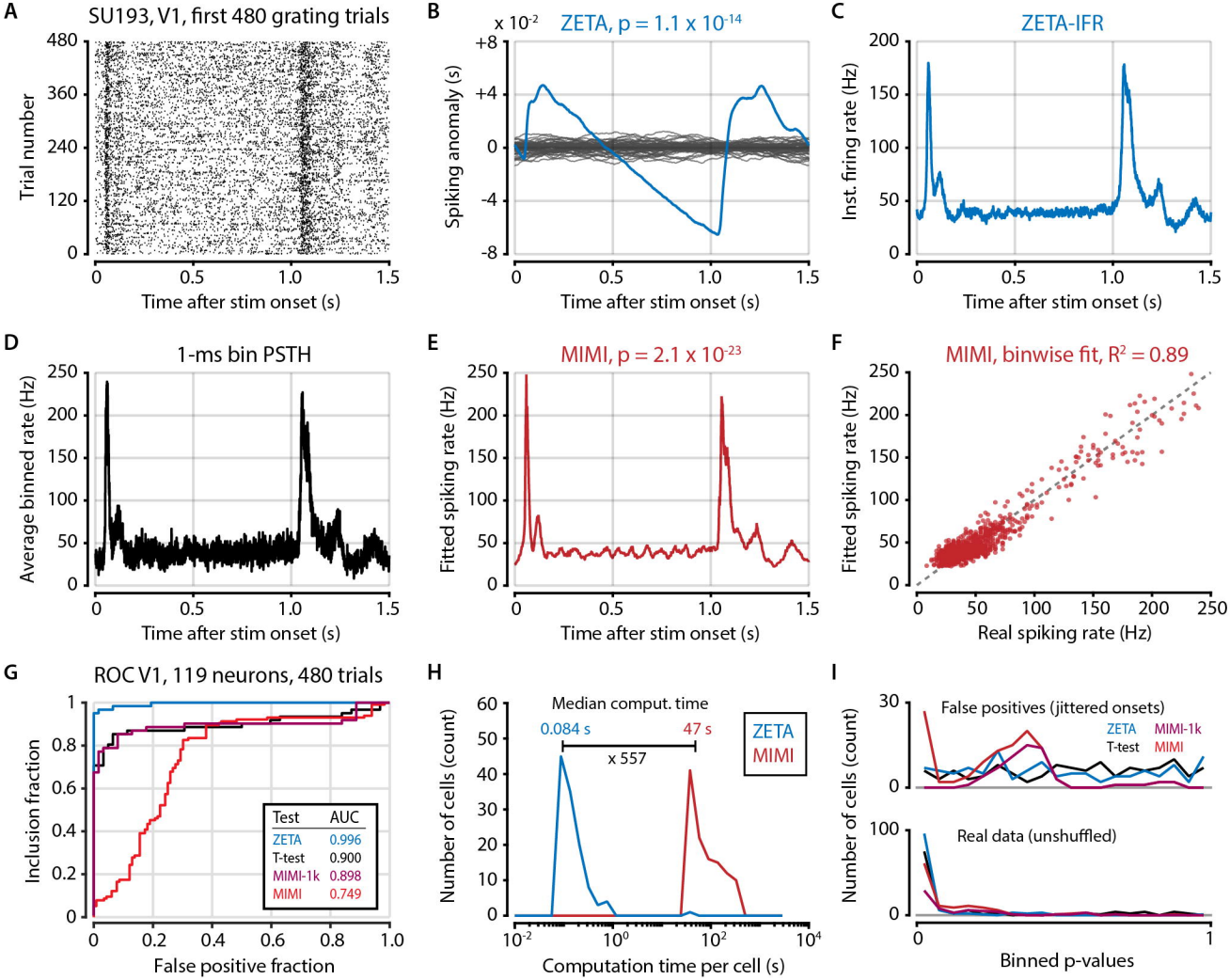

A
Stimulus
Trial #
480
360
240
120
0
0
0.5
1.0
1.5
Time after stim onset (s)
B
PSTH, 25 ms bins
Mean spiking rate (Hz)
15
10
5
0
0
0.5
1.0
1.5
Time after stim onset (s)
C
t-test, p=0.404
Mean spiking rate (Hz)
8
4
0
Baseline
Stimulus
PSTH
- Arbitrary bin size
- High noise or low temporal resolution
- Complex responses remain undetected
G
***
**
n = 12 data sets
Fraction of significant cells
1
0.8
0.6
0.4
0.2
0
T-test
ANOVA
ZETA
T-test
ANOVA
ZETA
Real
Jittered
ZETA
D
Cumulative spikes
Fractional spike position
1
0.8
0.6
0.4
0.2
0
Real data
Linear baseline
0
0.5
1.0
1.5
Time after stim onset (s)
E
ZETA, p = 3.84 x 10^-9
Deviation (ΔFrac)
+0.1
0
-0.1
0
0.5
1.0
1.5
Time after stim onset (s)
F
Instant. spiking rate (Hz)
15
10
5
0
Onset = 55 ms
0
0.5
1.0
1.5
Time after stim onset (s)
ZETA
+ No binning
+ More sensitive
+ Can detect arbitrary response patterns

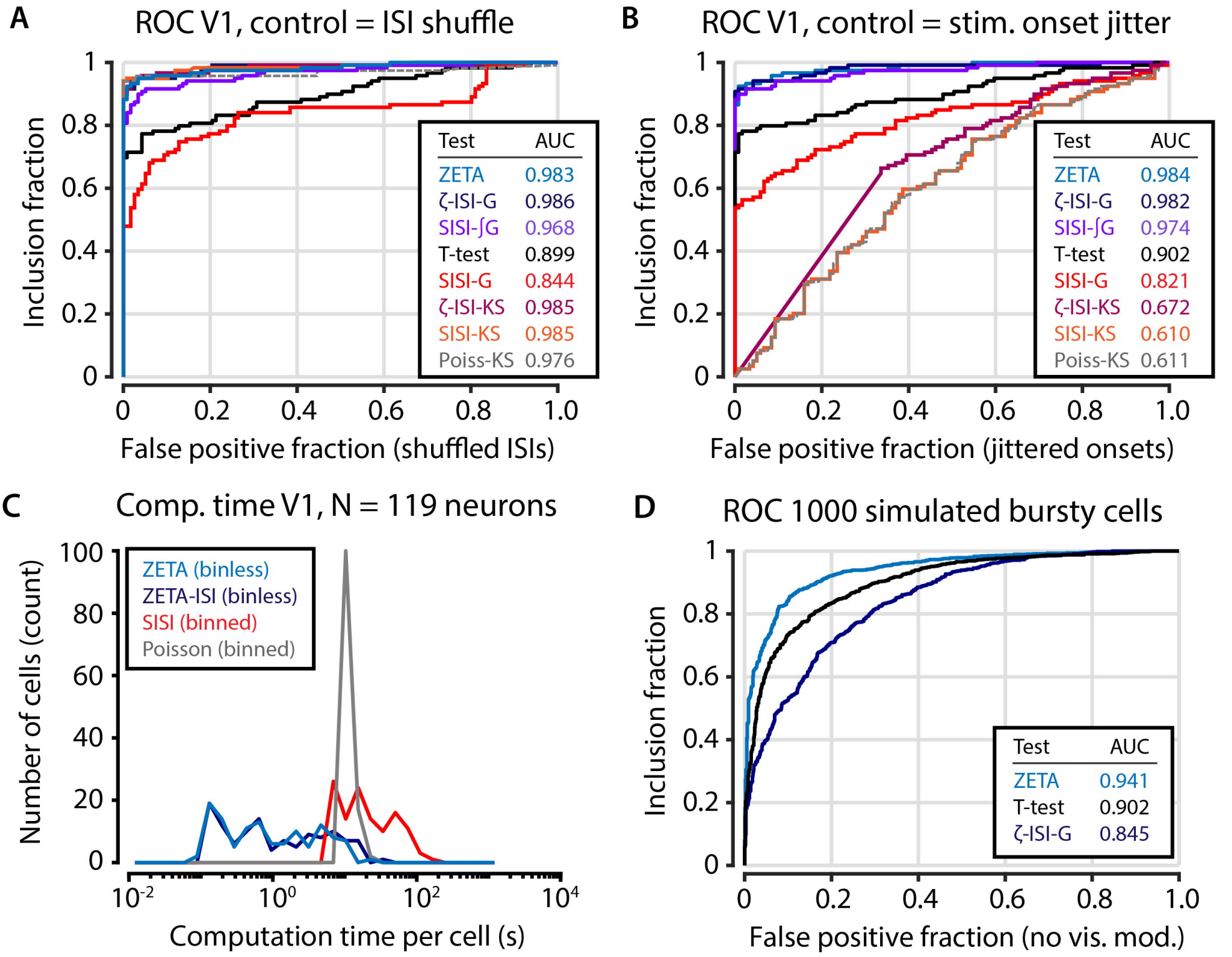

A
Original
Trial #: 20 21 22 23 24
Spikes:
Time (s): 2520 2525
Trial # 24 23 22 21 20
0 0.5 1.0 1.5
Time after stim. onset (s)
B
Inclusion: ζ=93.3%, T=77.3%, n=119
ZETA (ζc)
15
0
0
40
Rate-based t-test z-statistic (Φ-1(1-p/2))
ζ-p<0.05, μ n.s.
ζ-p & μ-p<0.05
ζ-p n.s., μ-p<0.05
ζ-p & μ-p n.s.
C
20200306U440
T, p=3x10-19
ζ, p=3x10-14
20200309U231
T, p=0.393
ζ, p=1x10-11
Binned spiking rate (Hz)
Deviation (ΔFrac)
x 10-2
Time after stim. onset (s)
D
Jittered onsets
Trial #: 20 21 22 23 24
Spikes:
Time (s): 2520 2525
Trial # 24 23 22 21 20
0 0.5 1.0 1.5
Time after stim. onset (s)
E
False positives: ζ=5.9%, T=5.0%
ZETA (ζc)
3
0
0
3
Rate-based t-test's z-statistic (Φ-1(1-p/2))
F
T, p=0.468
ζ, p=0.743
T, p=0.042
ζ, p=0.085
Binned spiking rate (Hz)
Deviation (ΔFrac)
x 10-2
Time after stim. onset (s)


A
100 trials, 1 Hz, τ=2 s, n=100 000 repeats
x 10-3
Mean[δi]
Empirical (μ ± sd)
Theory (eq. S21)
Spike number (i)
B
Not mean subtracted (δ)
x 10-3
Var[δi]
Empirical (μ ± sd)
Theory (eq. S22)
Spike number (i)
C
Mean-subtracted (d)
x 10-4
Var[di]
Empirical (μ ± sd)
Theory (eq. S34)
Spike number (i)

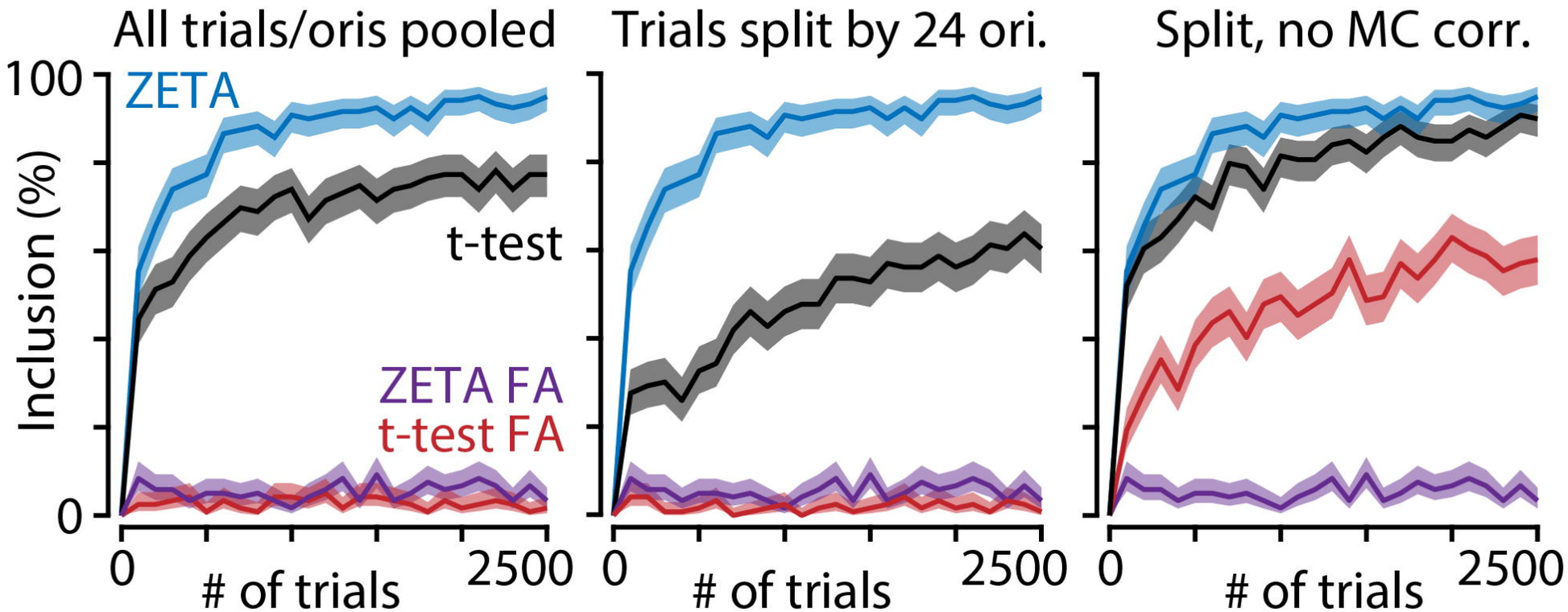

Area
Example data gratings
Area-level benchmark
V1
N=119
2560
Trial
1
0
1
1.5
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.985
A: 0.970
T: 0.893
AM
N=177
2560
Trial
1
0
1
7
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.813
A: 0.760
T: 0.701
NOT
N=95
748
Trial
1
0
1
80
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.968
A: 0.938
T: 0.910
PM
N=335
2400
Trial
1
0
1
13
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.909
A: 0.854
T: 0.818
Time after onset (s)
FP rate
Area
Example data gratings
Area-level benchmark
LGN
N=152
2560
Trial
1
0
1
60
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.969
A: 0.922
T: 0.870
SC
N=67
3680
Trial
1
0
1
7
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.998
A: 0.872
T: 0.929
LP
N=148
3840
Trial
1
0
1
17
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.930
A: 0.924
T: 0.859
APN
N=320
748
Trial
1
0
1
50
Hz
0
0
1
Inclusion
1
0
0
1
AUC
Z: 0.851
A: 0.784
T: 0.776
Time after onset (s)
FP rate


Area
Technique
Stimulus
Example data
Benchmark
V1
Neuronexus
Trial
288
1
0 2 3
Hz
26
0
0 2 3
ΔFrac
2
0
-2
0 2 3
Real
Jitter
μ
Z
N=40
V1
GCaMP6
Trial
96
1
0 3 8
dF/F0
1
0
0 3 8
ΔFrac
20
0
-20
0 3 8
Real
Jitter
μ
Z
N=183
SC
Neuronexus
Trial
336
1
0 2 3
Hz
25
0
0 2 3
ΔFrac
10
0
-30
0 2 3
Real
Jitter
μ
Z
N=89
Retina
pMEA
Trial
100
1
0 0.5 2
Hz
120
0
0 0.5 2
ΔFrac
15
0
-15
0 0.5 2
Real
Jitter
μ
Z
N=280

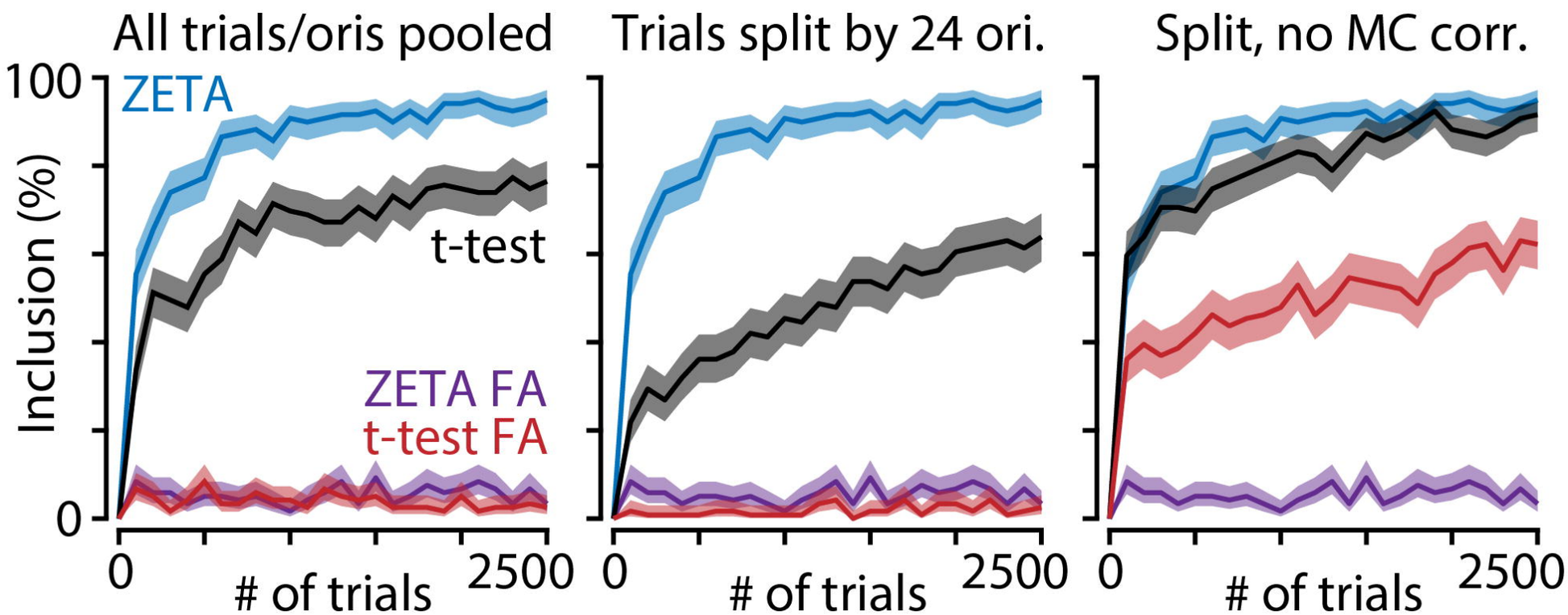

A
SU193, V1, first 480 grating trials
Trial number
Time after stim onset (s)
B
ZETA, p = 1.1 x 10^-14
x 10^-2
Spiking anomaly (s)
Time after stim onset (s)
C
ZETA-IFR
Inst. firing rate (Hz)
Time after stim onset (s)
D
1-ms bin PSTH
Average binned rate (Hz)
Time after stim onset (s)
E
MIMI, p = 2.1 x 10^-23
Fitted spiking rate (Hz)
Time after stim onset (s)
F
MIMI, binwise fit, R^2 = 0.89
Fitted spiking rate (Hz)
Real spiking rate (Hz)
G
ROC V1, 119 neurons, 480 trials
Inclusion fraction
False positive fraction
Test AUC
ZETA 0.996
T-test 0.900
MIMI-1k 0.898
MIMI 0.749
H
Median comput. time
0.084 s
47 s
x 557
ZETA
MIMI
Number of cells (count)
Computation time per cell (s)
I
False positives (jittered onsets)
ZETA
T-test
MIMI-1k
MIMI
Real data (unshuffled)
Number of cells (count)
Binned p-values

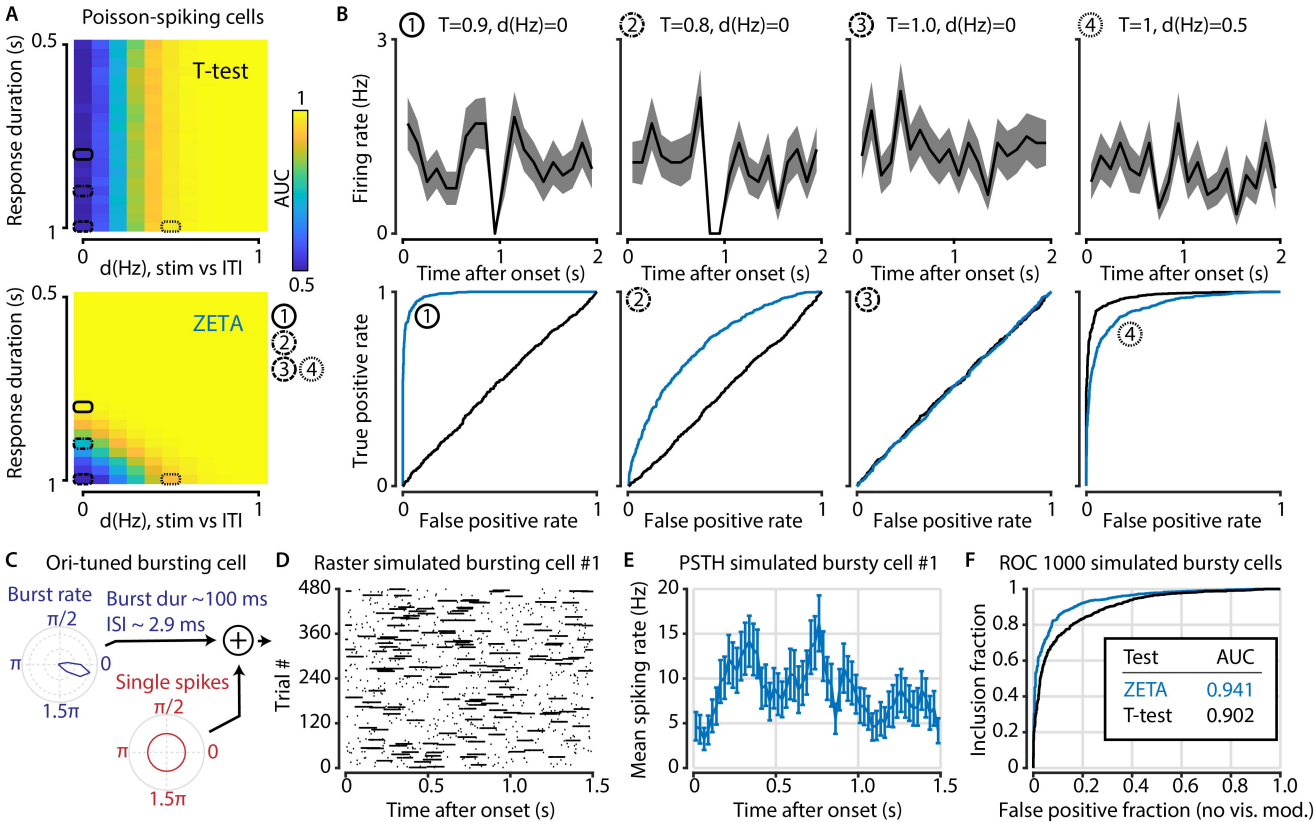

A
ROC V1, control = ISI shuffle
Inclusion fraction
False positive fraction (shuffled ISIs)
Test AUC
ZETA 0.983
ζ-ISI-G 0.986
SISI-∫G 0.968
T-test 0.899
SISI-G 0.844
ζ-ISI-KS 0.985
SISI-KS 0.985
Poiss-KS 0.976
B
ROC V1, control = stim. onset jitter
Inclusion fraction
False positive fraction (jittered onsets)
Test AUC
ZETA 0.984
ζ-ISI-G 0.982
SISI-∫G 0.974
T-test 0.902
SISI-G 0.821
ζ-ISI-KS 0.672
SISI-KS 0.610
Poiss-KS 0.611
C
Comp. time V1, N = 119 neurons
Number of cells (count)
ZETA (binless)
ZETA-ISI (binless)
SISI (binned)
Poisson (binned)
Computation time per cell (s)
D
ROC 1000 simulated bursty cells
Inclusion fraction
False positive fraction (no vis. mod.)
Test AUC
ZETA 0.941
T-test 0.902
ζ-ISI-G 0.845

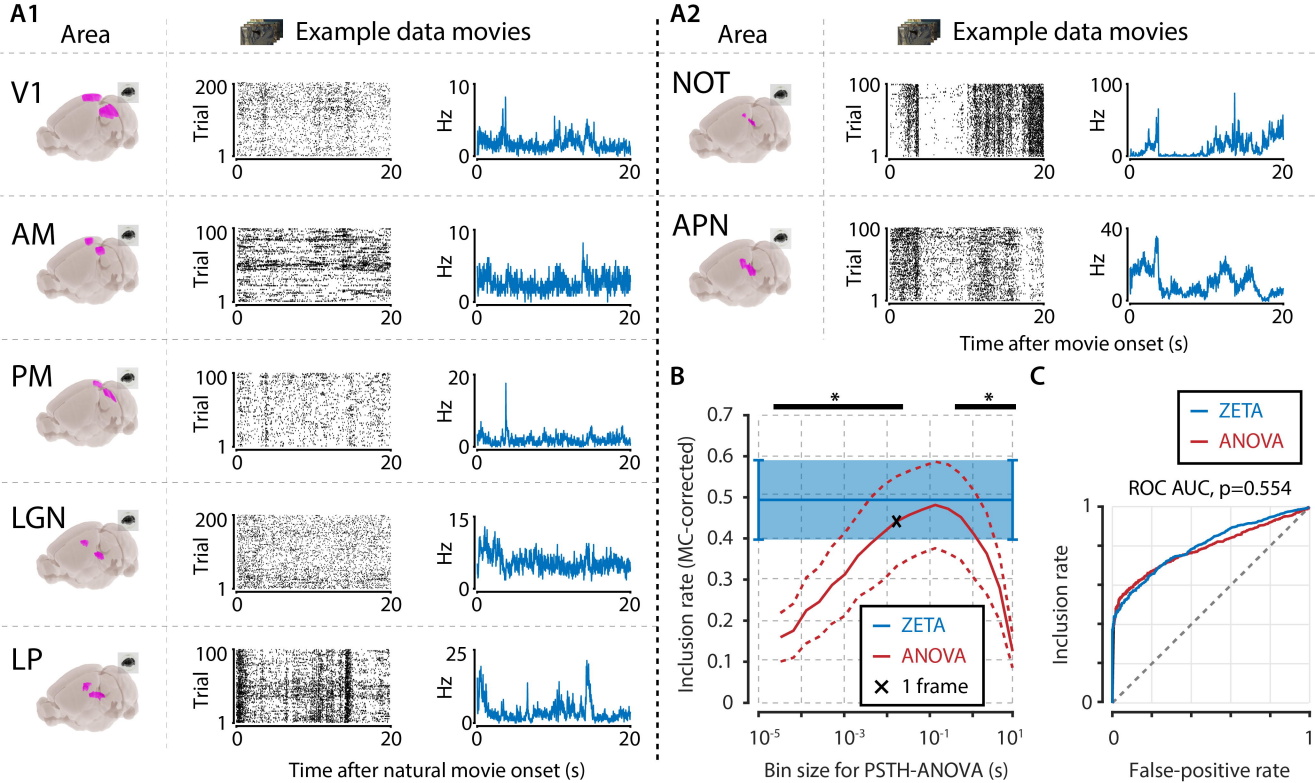
A Trial number and pre-processing dependence

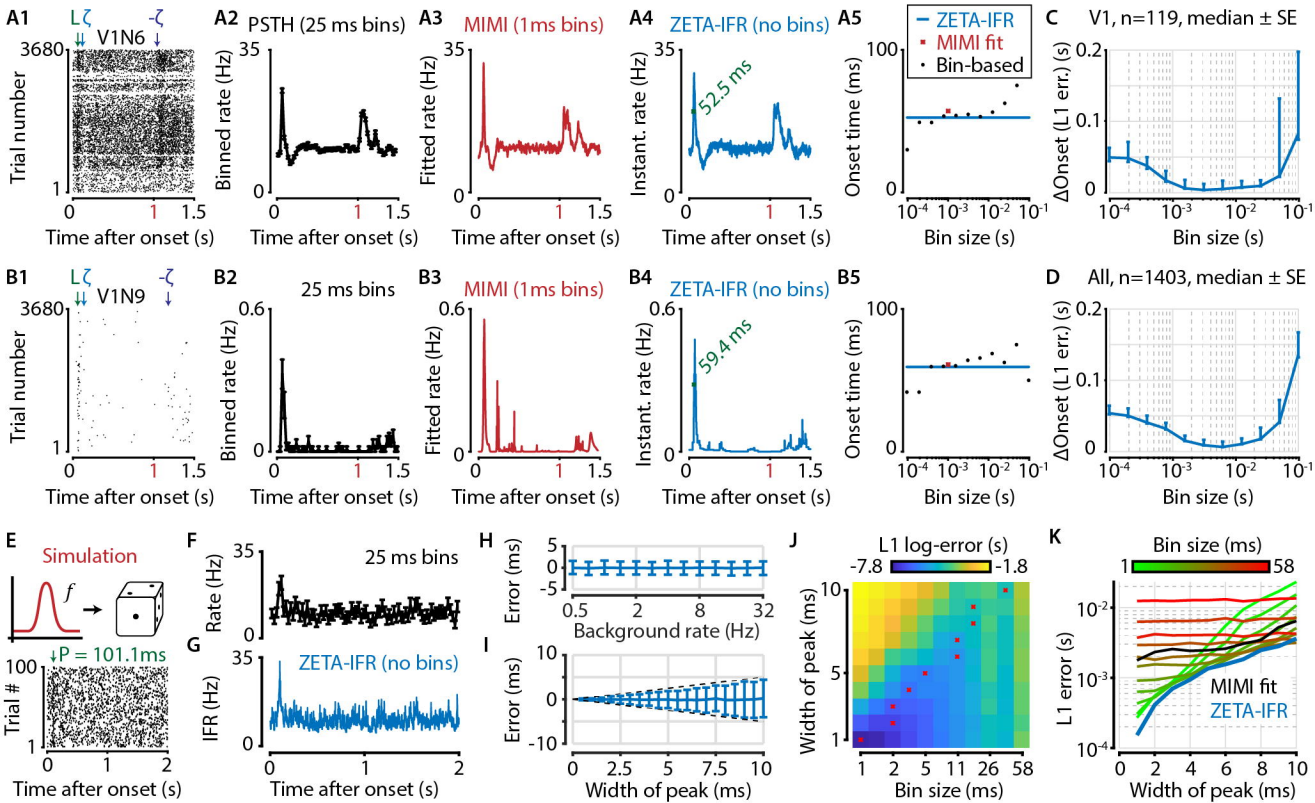


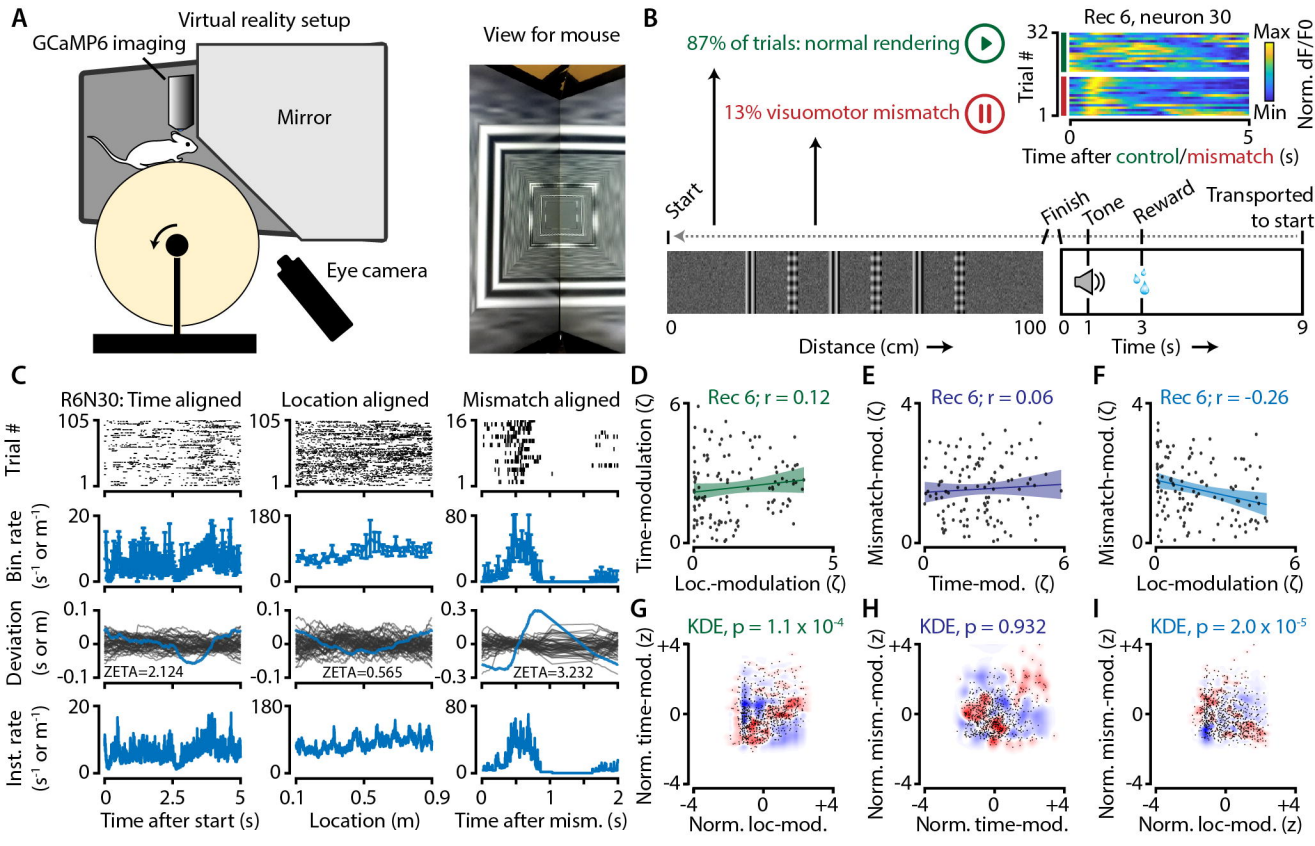
B As A, but with 300 ms baseline period

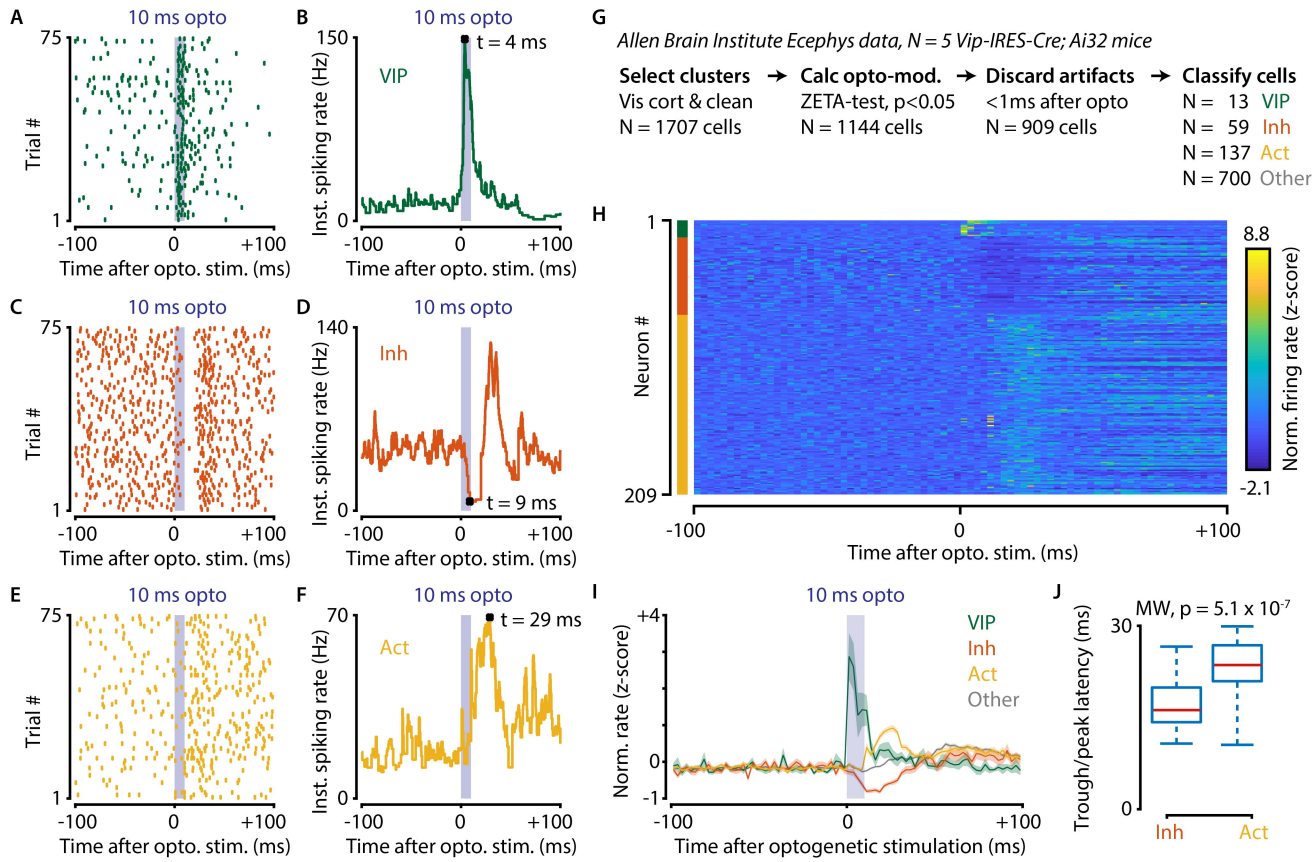



A
Poisson-spiking cells
T-test
ZETA
Response duration (s)
d(Hz), stim vs ITI
AUC
B
T=0.9, d(Hz)=0
T=0.8, d(Hz)=0
T=1.0, d(Hz)=0
T=1, d(Hz)=0.5
Firing rate (Hz)
Time after onset (s)
True positive rate
False positive rate
C
Ori-tuned bursting cell
Burst rate
Burst dur ~100 ms
ISI ~ 2.9 ms
Single spikes
D
Raster simulated bursting cell #1
Trial #
Time after onset (s)
E
PSTH simulated bursty cell #1
Mean spiking rate (Hz)
Time after onset (s)
F
ROC 1000 simulated bursty cells
Inclusion fraction
False positive fraction (no vis. mod.)
Test AUC
ZETA 0.941
T-test 0.902


A1
Area
Example data movies
V1
AM
PM
LGN
LP
Trial
Hz
Time after natural movie onset (s)
A2
Area
Example data movies
NOT
APN
Time after movie onset (s)
B
Inclusion rate (MC-corrected)
ZETA
ANOVA
1 frame
Bin size for PSTH-ANOVA (s)
C
ZETA
ANOVA
ROC AUC, p=0.554
Inclusion rate
False-positive rate


A1
L ζ
-ζ
V1N6
3680
Trial number
1
0
1
1.5
Time after onset (s)
A2
PSTH (25 ms bins)
35
Binned rate (Hz)
0
Time after onset (s)
A3
MIMI (1ms bins)
Fitted rate (Hz)
Time after onset (s)
A4
ZETA-IFR (no bins)
52.5 ms
Instant. rate (Hz)
Time after onset (s)
A5
ZETA-IFR
MIMI fit
Bin-based
100
Onset time (ms)
Bin size (s)
C
V1, n=119, median ± SE
0.2
0.1
ΔOnset (L1 err.) (s)
Bin size (s)
B1
L ζ
-ζ
V1N9
3680
Trial number
1
Time after onset (s)
B2
25 ms bins
0.6
Binned rate (Hz)
Time after onset (s)
B3
MIMI (1ms bins)
Fitted rate (Hz)
Time after onset (s)
B4
ZETA-IFR (no bins)
59.4 ms
Instant. rate (Hz)
Time after onset (s)
B5
Onset time (ms)
Bin size (s)
D
All, n=1403, median ± SE
ΔOnset (L1 err.) (s)
Bin size (s)
E
Simulation
f
↓P = 101.1ms
100
Trial #
1
Time after onset (s)
F
25 ms bins
Rate (Hz)
G
ZETA-IFR (no bins)
IFR (Hz)
Time after onset (s)
H
Error (ms)
5
0
-5
0.5
2
8
32
Background rate (Hz)
I
Error (ms)
10
0
-10
0
2.5
5
7.5
10
Width of peak (ms)
J
L1 log-error (s)
-7.8
-1.8
Width of peak (ms)
Bin size (ms)
K
Bin size (ms)
1
58
L1 error (s)
MIMI fit
ZETA-IFR
Width of peak (ms)


A
Virtual reality setup
GCaMP6 imaging
Mirror
Eye camera
View for mouse
B
87% of trials: normal rendering
13% visuomotor mismatch
Rec 6, neuron 30
Trial #
Max
Min
Norm. dF/F0
Time after control/mismatch (s)
Start
Finish
Tone
Reward
Transported to start
Distance (cm)
Time (s)
C
R6N30: Time aligned
Location aligned
Mismatch aligned
Trial #
Bin. rate (s⁻¹ or m⁻¹)
Deviation (s or m)
ZETA=2.124
ZETA=0.565
ZETA=3.232
Inst. rate (s⁻¹ or m⁻¹)
Time after start (s)
Location (m)
Time after mism. (s)
D
Rec 6; r = 0.12
Time-modulation (ζ)
Loc.-modulation (ζ)
E
Rec 6; r = 0.06
Mismatch-mod. (ζ)
Time-mod. (ζ)
F
Rec 6; r = -0.26
Mismatch-mod. (ζ)
Loc-modulation (ζ)
G
KDE, p = 1.1 x 10⁻⁴
Norm. time-mod. (z)
Norm. loc-mod.
H
KDE, p = 0.932
Norm. mism.-mod. (z)
Norm. time-mod.
I
KDE, p = 2.0 x 10⁻⁵
Norm. mism.-mod. (z)
Norm. loc-mod. (z)


A
10 ms opto
75
Trial #
1
-100
0
+100
Time after opto. stim. (ms)
B
10 ms opto
150
VIP
t = 4 ms
Inst. spiking rate (Hz)
0
-100
0
+100
Time after opto. stim. (ms)
C
10 ms opto
75
Trial #
1
-100
0
+100
Time after opto. stim. (ms)
D
10 ms opto
140
Inh
t = 9 ms
Inst. spiking rate (Hz)
0
-100
0
+100
Time after opto. stim. (ms)
E
10 ms opto
75
Trial #
1
-100
0
+100
Time after opto. stim. (ms)
F
10 ms opto
70
Act
t = 29 ms
Inst. spiking rate (Hz)
0
-100
0
+100
Time after opto. stim. (ms)
G
Allen Brain Institute Ecephys data, N = 5 Vip-IRES-Cre; Ai32 mice
Select clusters
Vis cort & clean
N = 1707 cells
Calc opto-mod.
ZETA-test, p<0.05
N = 1144 cells
Discard artifacts
<1ms after opto
N = 909 cells
Classify cells
N = 13 VIP
N = 59 Inh
N = 137 Act
N = 700 Other
H
1
Neuron #
209
-100
0
+100
Time after opto. stim. (ms)
8.8
-2.1
Norm. firing rate (z-score)
I
10 ms opto
+4
0
-1
Norm. rate (z-score)
VIP
Inh
Act
Other
-100
0
+100
Time after optogenetic stimulation (ms)
J
MW, p = 5.1 x 10-7
30
0
Trough/peak latency (ms)
Inh
Act