A new family of statistical tests for responses in point-event and time-series data for one- and two-sample comparisons

Authors

J. Alexander Heimel[1], Guido T. Meijer[2], Jorrit S. Montijn[1]

Affiliations

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

2 Donders Centre for Neuroscience, Radboud University, Nijmegen, The Netherlands

Contact Information

Correspondence should be addressed to J.S.M. (jsmontijn@gmail.com).

Abstract

Quantifying whether and when signals are modulated by autonomous or external events is ubiquitous in the field of neuroscience. Existing statistical approaches, however, are not ideally suited to do this, especially when the signals under scrutiny show temporal autocorrelations. For example, a standard approach in the analysis of calcium imaging data is to use a t-test on predetermined time-windows to quantify whether neurons respond (differently) to an event of interest. While this is attractive because of its simplicity, only average signal differences can be detected. In practice, neurons often show complex response dynamics which are missed by conventional statistical tests. To solve this issue, we present an improved version of the recently developed ZETA-test which implements support for analysing time-series data. Furthermore, it includes a two-sample test to detect a difference in neural responses between two conditions. We show that our method has a statistical sensitivity superior to t-tests and ANOVAs and works well with temporally correlated data. Open-source code for implementations in MATLAB and Python is available on GitHub.

Keywords: Neural data analysis, Statistics, Calcium imaging, Electrophysiology

Introduction

Neurophysiological studies depend on a reliable quantification of whether and when neural signals are modulated by autonomous or external events, such as the presentation of a sensory stimulus. In the analysis of calcium imaging data, a standard approach is to use a t-test on a predetermined time-window to include only the cells that are significantly modulated. While this approach is attractive because of its simplicity, statistical tests such as the t-test only detect differences in signal averages between the time windows of interest. On the other hand, more advanced model-based methods are rarely used, as they require fitting and manual hyperparameter tuning. These advanced methods can be very computation and/or labour intensive, and may not be feasible for large data sets. A middle ground between these approaches is two perhaps the bin-wise ANOVA, when used in combination with an automatic bin-width determining algorithm (Shimazaki & Shinomoto, 2007). While the ANOVA is an improvement over the t-test in many cases, it is ill suited to analysis of temporally-dependent signals, as we will show later. New recording techniques, such as wide-field and multi-plane imaging, yield increasingly large numbers of cells, and such data-sets may benefit from a test for neuronal responsiveness that requires no arbitrary parameters, binning, or manual curation, and is not negatively affected by temporally correlated signals.

Recent work has proposed such a statistical test for electrophysiological spiking data: the ZETA-test (Montijn et al., 2021). Several studies have already been published that use this approach. It is particularly well suited to detecting whether cells are driven by optogenetic stimulation in opto-tagging experiments (Dudok et al., 2021; Schneider et al., 2023; Spyropoulos et al., 2023; Szadzinska et al., 2021). It has also been used to detect response onset latencies (Oude Lohuis et al., 2022), and quantify somatosensory and visual stimulus responsiveness (Burnett et al., 2023; Montijn et al., 2023; Qin et al., 2023; Ziegler et al., 2023). While the ZETA-test works well for spiking data, an important shortcoming is that it can only be applied to point events, such as spike times. Another limitation of the ZETA-test is that it cannot be used to determine differences between two conditions.

To address these limitations, we developed three new statistical methods inspired by this previous work: 1) a ZETA-test that can be applied to time-series data, like from calcium imaging experiments. 2) A two-sample ZETA-test to determine whether there exists a difference in neuronal spiking activity between two conditions. 3) A two-sample ZETA-test for time-series data. We have tested the performance of these methods on real and synthetic data,

and found that it outperforms common approaches such as t-tests and ANOVAs. We expect that our procedures may be of interest to statisticians and theoreticians, but we have written this manuscript specifically with experimental neuroscientists in mind, who can use our open-source code in MATLAB (https://github.com/JorritMontijn/zetatest) and Python (https://github.com/JorritMontijn/zetapy) to obtain higher yields and more reliable results from their neurophysiological data.

Results

The original ZETA-test and the addition of a data-stitching step

The Zenith of Event-based Time-locked Anomalies test (ZETA-test) has been described in detail previously (Montijn et al., 2021) and in the Methods section, but we will concisely summarize its procedure here. 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. 1B). The deviation of this curve from a linear baseline represents whether the neuron has a higher or lower spiking density than a non-modulated spiking rate (fig. 1C, blue curve). We compare this pattern to the likelihood of observing it by chance by running multiple randomized bootstraps. In each bootstrap iteration, we jitter all stimulus-onset times to generate a single null hypothesis sample (fig. 1C, grey curves). After scaling the experimentally observed curve to the variation in the null hypothesis distribution obtained from all bootstraps, we transform it into a p-value by using the direct quantile position, or by approximation with the Gumbel distribution. 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.

For all statistical benchmarking analyses hereafter, we will use a Receiver Operating Characteristic (ROC) analysis to quantify how well different statistical tests can discriminate whether a neuron is responsive to visual stimulation with drifting gratings. For the spike-based data sets, we recorded 1504 cells in the primary visual cortex (V1) of seven awake mice using Neuropixels (Montijn et al., 2023). To estimate the false positive rate, we randomly jittered the stimulus onset times and reran the procedure. In an ROC analysis, the true-positive rate (significantly responsive cells) is plotted on the y-axis, as a function of the false-positive rate (x-axis), with each point corresponding to one p-value threshold. The curve therefore follows

the ratio of true-positives/false-positives for various values of α, ranging from the very lowest to highest p-value in the combined set of both true and false positives. For a properly calibrated statistical test, the false positive rate is equal to the threshold value α. A perfect discriminator has an area under the curve (AUC) of 1.0, while a random discriminator has an AUC of 0.5.

The first topic we investigated was a shortcoming in the original ZETA-test. The ZETA-test's statistical model assumes that the real and null-hypothesis samples are taken from the same set of point events (spikes). While this holds true when stimulus events occur at fixed intervals, this will be false with heterogeneous inter-event durations, such as for self-generated behavioural events. In this latter case, jittering the onsets may lead to varying inclusion and exclusion of parts of the data (see Methods). We have therefore modified the ZETA-test to include an optional data-stitching step that ensures data conformity between the real and onset-jittered procedures (fig. 1D). While this change may seem significant from a purely theoretical point of view, it had surprisingly little effect on the statistical sensitivity in real-world experimental data with mild heterogeneity of inter-event durations (fig. 1E). We found the following statistical sensitivies: ZETA-test with stitching AUC=0.897, ZETA-test without stitching AUC = 0.899, ANOVA AUC=0.867, T-test AUC=0.785. While the AUC of the ZETA-test was significantly higher than of the ANOVA (Mann-Whitney U-test, $p=2.4 \times 10^{-6}$) and the t-test ($p= 7.4 \times 10^{-54}$), there was no difference in AUC between the stitched and non-stitched ZETA-tests (p=0.80). For the ANOVA, we used the Shimazaki&Shinomoto procedure to calculate the optimal binning size (Shimazaki & Shinomoto, 2007), while for the t-test we tested compared the mean firing rate between the stimulus period (0.0 – 1.0 s) and pre-stimulus inter-trial interval (-0.5 – 0.0 s).

Next, we tested the performance of these four tests under four different synthetic benchmarks using simulated neurons with exponential inter-spike intervals. For each trial, we varied the average inter-spike interval between stimulus and baseline (Biphasic neurons; fig. 1F), or between onset, stimulus and baseline (Triphasic neurons; fig. 1G,H). Finally, we ran a worst-case scenario for the ZETA-test by sampling the number of spikes from a discretized Gaussian distribution and scattering the spike times following uniform random distribution across a 1-second window (fig. 1I). In simulations for fig. 1F,G, we varied the inter-trial intervals from 0.2 – 2.0 s to investigate the effect of the stitching procedure with heterogeneous inter-event durations. Under these conditions we found little difference between the stitching and non-stitching ZETA-tests: Biphasic, 0.645 vs 0.629 ($p=2.7 \times 10^{-5}$); Triphasic, 0.925 vs

0.919 ($p=9.4 \times 10^{-4}$) for with vs without stitching respectively. Next, we simulated a case where the window of interest (τ) was chosen such that it discarded the transitions into and out of baseline-level activity (fig. 1H). In this case, we found a large difference ($p<10^{-100}$) in statistical sensitivity when comparing stitching (AUC=0.882) with no-stitching (AUC=0.694). This result shows that stitching can be a critical step under some specific circumstances, especially for relatively sparse events where the window of interest may be unclear (for example, the duration of the period following licking). Moreover, we found that the ZETA-test outperformed the optimal ANOVA in all cases, except for the Gaussian-noise model. Under these conditions, which we optimized for the ANOVA's performance, and which are highly unlikely to occur in experimental neuroscience data, the difference between the ZETA-test and the optimal ANOVA was small (ZETA-test AUC=0.653, ANOVA AUC=0.674), but significant ($p=1.6 \times 10^{-8}$).

Time-series ZETA test

The original formulation of the ZETA-test could be applied to any two sets of point events, such as spike times and stimulus onsets. While its application to electrophysiological data was therefore straightforward, many other methods produce time-series data, such as calcium imaging. We previously applied the ZETA-test to transient-detected calcium imaging data and found that it performed quite well (Montijn et al., 2021). Nevertheless, a better and more generic solution would be a ZETA-test for time-series data, as this could also be applied to data obtained with EEG, patch recordings, fMRI, etc. The Methods section describes an alternative formulation of such a time-series ZETA-test, which we applied to real data and various synthetic benchmarks to test its performance. In a nutshell, the modification is that we replace the cumulative distribution of spikes by a cumulative sum of data values and perform some extra steps, such as data scaling. An extra trick we employ is that if samples are acquired with varying latencies relative to the events of interest, this fact can be used to construct a super-resolution time-series average over trials. This allows the time-series ZETA-test to also effectively deal with data sampled with heterogeneous intervals.

We found the time-series ZETA-test outperformed t-tests and optimally-binned ANOVAs on both GCaMP6f (2430 cells) and OGB 1-AM data (1204 cells) (fig. 2). We ran a similar analysis as described above, using a receiver-operating characteristic (ROC) analysis. However, we found that including all 80 trials saturated the performance of the ZETA-test to

>0.99 for the OGB data. The results we present here are therefore from data sets where we used only the first 8 trials of each data set. We observed that the statistical sensitivity was higher for the time-series ZETA-test than the ANOVA (OGB data ZETA-test AUC=0.944 vs ANOVA AUC=0.924, $p=6.4 \times 10^{-5}$; GCaMP data ZETA-test AUC=0.818 vs ANOVA AUC=0.786, $p=2.4 \times 10^{-7}$). We also noticed a strong discrepancy in the false-alarm rate for ANOVAs that was absent in the ZETA-test. For relatively large values of alpha, the false positive rate was of the shuffle-control was close to the theoretical norm. Below an alpha of approximately 0.05, however, the ANOVA's false positive rate started to diverge from the norm and became excessively liberal (fig. 2I-J). For example, to obtain an empirical false positive rate of 0.001 using the ANOVA, one would require a p-value threshold of around 10^{-8} in OGB data and 10^{-13} in GCaMP data rather than the expected 10^{-3} if p-value scaling had followed the theoretical norm. While some deviation in empirical thresholds from the theoretical norm is expected, a discrepancy of 10 orders of magnitude is excessive.

We suspected this discrepancy is caused by the ANOVA's model assuming statistical independence between adjacent bins of the peri-stimulus time histograms (PSTHs). As calcium indicators are low-pass filters, and even the underlying spiking rate itself is not temporally independent, this assumption is violated. To confirm that the aforementioned effect was not due to other, more complex properties of the experimental data, we ran a simulation with 10000 quadriphasic cells. We generated spiking times as described in the previous section, and applied a Gaussian-filter on spike counts of 40-ms bins (25 Hz) to simulate the effect of the calcium indicator. The results of this simulation were similar to those we found in real data: the AUC of the T-ZETA and ANOVA were similar (fig. 2H), but the false-positive rate of the ANOVA was very high (fig. 2K). We therefore strongly advise against using the ANOVA, or other bin-based analyses, when analyzing temporally-autocorrelated time-series data. We also note that the time-series ZETA-test was somewhat conservative, but it still yielded higher inclusion rates than the t-test owing to its superior sensitivity (GCaMP inclusion rate at $\alpha=0.05$: T-ZETA-test=0.325, t-test=0.251; OGB: T-ZETA-test=0.607, t-test=0.445). The AUC of the T-ZETA test was significantly higher than that of the t-test (GCaMP data T-ZETA-test AUC=0.819 vs t-test AUC=0.788, $p=1.1 \times 10^{-29}$; OGB data T-ZETA-test AUC=0.944 vs t-test AUC=0.842, $p=3.5 \times 10^{-60}$).

Two-sample ZETA test

We developed the ZETA-test to detect whether a neuron's activity was modulated by the occurrence of a set of events. This approach works especially well to, for example, determine the genetic subtype of neurons in an opto-tagging data set (Dudok et al., 2021; Schneider et al., 2023; Szadzinska et al., 2021). Ideally, however, one would also be able to use a binless statistical test to determine whether two neurons behave differently in response to the same stimulus, or whether the same neuron differs in response between two stimuli. We have therefore developed a two-sample zeta-test which can be used to assess the difference in response between two conditions. In short, rather than comparing the response of a neuron to a linear baseline rate, we define a temporal spiking-density deviation vector as the difference in normalized cumulative spike rates between two conditions (fig. 3A-D). We compare the maximum absolute deviation to those we obtain using a trial-swapping procedure where we re-generate a deviation vector in every shuffle-iteration in which we randomly assign trials to one or the other condition. The statistical significance of the test is then the likelihood to find the real data within the distribution of randomized shuffle controls. The one-sample ZETA-test can be seen as a special case of the two-sample ZETA-test where one of the two conditions is taken to be an infinite set of stimulus onsets over the same period as the real data.

We tested the performance of the two-sample ZETA-test (ZETA2-test) using multiple benchmarks. First, we used data from Neuropixels recordings in primary visual cortex (Montijn et al., 2023). We ran two comparisons on these data: first, instead of detecting whether a neuron was visually responsive to gratings drifting, we tested whether the response of two neurons to those stimuli differed (n=1000 randomly selected pairs of neurons) (fig. 3E). A possible use case for this type of comparison would for example be during pre-processing of electrophysiological data, when deciding to merge to two clusters of spikes. We compared the performance of the two-sample zeta-test to that of the two-sample t-test, and a two-way ANOVA. For the ANOVA, we took the p-value to be the lower of the two Bonferroni-corrected p-values for the main effect between conditions (i.e., whether there was a mean-rate difference) and the bin x condition interaction effect. We determined the optimal bin size for the ANOVA using the Shimazaki & Shinomoto procedure (Shimazaki & Shinomoto, 2007). Curiously, the ANOVA displayed the lowest AUC of the three types of tests (optimally-binned two-way ANOVA, AUC=0.926), with the t-test performing intermediately (two-sample t-test, AUC=0.968), and the ZETA2-test performing the best (ZETA2-test, AUC=0.980). All

differences in AUC were significant (Mann-Whitney U test: ZETA2 vs t-test, $p=7.6x10^{-4}$, ZETA2 vs ANOVA, $p=1.6x10^{-30}$; t-test vs ANOVA, $p=3.7x10^{-15}$). Second, we compared how the tests fared when determining whether neurons responded differently to stimulus directions 0 and 90 degrees (fig. 3F). Absolute AUCs were lower than before, but the ranking remained the same: the ZETA2-test performed best (AUC=0.687), followed closely by the t-test (AUC=0.677) and less closely by the ANOVA (AUC=0.622).

Next, we simulated two cases to highlight the strong and weak points of the ZETA2-test. We generated 1000 pairs of cells with a randomly assigned background spiking activity level (mean=1.0 Hz) following an exponential inter-spike interval distribution. Both cells always had the same average background spiking rate, but on top of these spikes we added a single spike in 25% of 240 trials for neuron 1 and 50% of 120 trials for neuron 2 (with a temporal delay of exactly 55 ms and a trial-to-trial jitter of 1 ms). The only way to discriminate these cells is therefore the difference in spike counts during the response peak. The optimal ANOVA could therefore select a bin width that maximized the spike-count differences and yielded an AUC of 0.931 (fig. 3G). The ZETA2-test followed with 0.860 and the t-test with 0.802. Finally, we simulated a similar situation where there was no difference in peak height, but in peak time. For neuron 1 we set the peak time at 53 ms and for neuron 2 at 55 ms, with both cells receiving a single spike in 25% of all trials. Perhaps unsurprisingly, the ZETA2-test was well able to differentiate these cells with an AUC of 0.757 (fig. 3H). The optimal ANOVA was barely above chance with an AUC of 0.516, and the t-test gave an AUC of exactly 0.500.

Two-sample time-series ZETA test

The fourth member of the ZETA-family of tests is the two-sample version of the time-series ZETA-test (T-ZETA2-test). We benchmarked its performance by applying it to the same calcium imaging data sets as for the one-sample time-series ZETA-test, now calculating whether there is a difference in the response of one neuron to its preferred drifting grating direction and the stimulus orthogonal to it. First, we tested whether each neuron was direction- or orientation-tuned by fitting a double von Mises curve and selecting neurons for further analysis only if the R^2 of the fit was significant (von Mises R^2, $p<0.05$). We found 1713/2430 (70%) GCaMP cells and 673/1204 (56%) OGB cells were significantly tuned. To compute the baseline false-positive rate, we split the preferred stimulus trials into two random sets of trials and calculated whether there was a difference in response between them. As for the spike-based

ZETA2-test, we compared the performance between a two-way ANOVA, two-sample t-test and T-ZETA2-test. First, we found that the AUC was highest for the T-ZETA2-test in both data sets (GCaMP, AUC=0.745; OGB, AUC=0.754). The AUC for the ANOVA and t-test were similar (ANOVA: GCaMP, AUC=0.701; OGB, AUC=0.689; t-test: GCaMP, AUC=0.688; OGB, AUC=0.712) (Fig. 6). However, the false positive rate (FPR) for the ANOVA was again exceptionally high: at an α=0.05 the FPR was 0.683 for GCaMP and 0.733 for OGB, whereas the FPRs for the T-ZETA2-test were 0.039 and 0.052 and for the t-test were 0.045 and 0.046 (for GCaMP and OGB respectively). As we mentioned above, the ANOVA is the wrong statistical model for the analysis of time-series data due to the temporal autocorrelation of signals across bins. Overall, the performance of the two-sample time-series ZETA-test was excellent: it showed a false-positive rate close to the theoretical norm and a superior statistical sensitivity to both t-tests and ANOVAs.

Discussion

We created new members of the ZETA family of statistical tests that can be applied to two-sample comparisons and time-series data. The family of ZETA-tests is built upon a statistical method for determining whether neuronal responses are modulated by the occurrence of events. The method is sensitive to arbitrarily complex response patterns and robust in the face of temporally autocorrelated signals. In all experimental data, the ZETA-test showed markedly improved statistical sensitivity compared to established and powerful statistical techniques, such as t-tests and optimally-binned ANOVAs. Moreover, the performance of all types of ZETA-tests was without exception equal or superior to those of t-tests, even in unrealistic synthetic benchmarks where there was only a mean-rate difference (for example fig. 3G). Finally, the family of ZETA-tests is even easier to use than established methods, as it can be applied directly to raw spike times or time-series data and stimulus onsets, and the lack of parameter selection naturally lends itself to the bulk-analysis of large numbers of cells.

We described the procedure of data-stitching to deal with heterogeneous inter-event durations. While this procedure is in most cases as good as, or better than, using the unstitched data, there are exceptions. In the situation where events with heterogeneous inter-event durations lead to short responses in an otherwise stationary background activity, data stitching may reduce the statistical sensitivity of the ZETA-test. Extending the jittered data into periods of stationary inter-event epochs means the variance of the jittered data is reduced, compared to

stitched data where larger jitter magnitudes can include the response of the preceding or following event. However, when a set of control data exists where the only difference from the experimental data is the presence of the events of interest, it might be preferable to instead use the two-sample ZETA-test and directly compare these two conditions with surrogate events for the control condition.

The T-ZETA-test described in this paper may superficially resemble the Kolmogorov-Smirnov (KS) test, so one could wonder how these approaches differ. In our original derivation of ZETA, we perform an in-depth comparison between the KS-test and ZETA-test, so we refer to the reader to this earlier work for more details (Montijn et al., 2021). In short, the main difference between ZETA and KS is that the KS-test is very sensitive to any difference in the cumulative distribution between two conditions, even if that difference is not time-locked to a stimulus but results from intrinsic differences in the distribution of neuronal activity. This makes the KS-test unsuitable for application to neuronal activity, as it generates many false positives. The ZETA method takes only the maximum deviation as a metric for "differentness", and as a result becomes less sensitive to the exact shape of the cumulative distribution.

The ZETA2 test is also much more sensitive than the two-sample Kolmogorov-Smirnov test for data with multiple trials, as the Kolmogorov-Smirnov test does not consider the variability across trials. The ZETA2 test also shows some similarities to a standard permutation test over binned data. Here, the main difference is that each null hypothesis sample in the case of the ZETA2 test is the maximum deviation over all time points rather than a value for each sample time, permuted over trials. The ZETA2 test therefore circumvents the multiple-comparison problem that would arise if one would, for example, perform a permutation test for each sample time. Because of this invariance to the number of sample points, we were also able to use some other tricks, such as creating a super-resolution reference time vector to allow for its application to data with heterogeneous sampling intervals.

In conclusion, the family of ZETA-tests are simpler, more statistically powerful, and less error-prone tools than bin-based PSTHs, t-tests and ANOVAs. ANOVAs can be powerful tools when combined with optimal-binning algorithms, but assume statistical independence between bins. As our results have shown, this causes the ANOVA to output misleading p-values in the case of calcium imaging data. The expanded utility of the ZETA-test to two-sample comparisons and time-series data can therefore provide an attractive, more robust, and more statistically sensitive alternative to other approaches. Implementations of the four ZETA-tests

described in this paper are available in MATLAB (https://github.com/JorritMontijn/zetatest) and Python (https://github.com/JorritMontijn/zetapy).

Methods

A summary of the ZETA-test

The Zenith of Event-based Time-locked Anomalies (ZETA) test has been described in detail previously (Montijn et al., 2021), but for completeness we will describe it below. In the paragraphs thereafter, we will present a modification that addresses a potential shortcoming of the original ZETA-test, explain how we constructed a test for time-series data, and present statistical procedures to perform two-sample comparisons for both spike-based and time-series data.

The metric ζ is computed on a vector of $i = [1 \ldots n]$ spike times $\boldsymbol{x}$, and a vector of $k = [1 \ldots q]$ event times $\boldsymbol{w}$. 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:

$$v_i = x_i - w_k \qquad 1$$

Where k is chosen such that

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

Next, we remove all spike times that are larger than a cut-off value τ, for example the trial-to-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 $\boldsymbol{v}$:

$$g_i = i/n \qquad 3$$

Therefore, $\boldsymbol{g}$ 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:

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

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 shown previously, using δ_i to compute ZETA would make it dependent on the choice of onset times (Montijn et al., 2021). Therefore, we create $\boldsymbol{d}$, a time-invariant 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_r) as the most extreme value, i.e. the maximum of the absolute values:

$$\zeta_r \equiv \max\left(|\boldsymbol{d}|\right) \tag{8}$$

Having calculated ZETA from the temporal deviation vector $\boldsymbol{d}$, we wish to quantify its statistical significance. We therefore construct a null hypothesis distribution by repeating the above procedure M times with shifted event-times $\boldsymbol{w'}$, where for each jitter iteration m and event k, we move each event time by a random amount sampled from the interval [-τ, τ]:

$$w'_{m,k} = w_k + \varepsilon \tag{9}$$

Where each ε is independently drawn from a uniform distribution U:

$$\varepsilon \sim U(-\tau, \tau) \tag{10}$$

We repeat this process M times; where for each jitter iteration m, we calculate $\boldsymbol{\delta}'(m)$:

$$\boldsymbol{\delta}'(m) = \boldsymbol{g}'(m) - \boldsymbol{b}'(m) \tag{11}$$

As before, we mean-normalize $\boldsymbol{\delta}'(m)$ and define a null-hypothesis ZETA sample m as:

$$\zeta'(m) \equiv \max\left(\left|\boldsymbol{\delta}'(m) - \bar{\delta}'(m)\right|\right) \tag{12}$$

Having a way to generate null-hypothesis samples, we can now calculate the significance of ζ_r either directly, using the quantile position of ζ_r in $\boldsymbol{\zeta'}$, or we can approximate it using the Gumbel distribution (Montijn et al., 2021). Either way, we obtain an estimate that asymptotically converges to a deterministic value as the number of jitter iterations M grows. We then use the standard normal's quantile function Φ^{-1} to obtain a corrected ZETA metric ζ that is interpretable as a z-score:

$$\zeta = -\Phi^{-1}\left(\frac{1}{2} - \frac{1}{2}\int_{-\infty}^{\zeta_r} \zeta'(x)\, dx\right) \tag{13}$$

With its corresponding p-value defined by:

$$p = 2 - 2\,\Phi(\zeta\,) \qquad 14$$

Where Φ is the cumulative normal distribution. 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.

Data-stitching to ensure conformity between the real data and random bootstraps

An important step that was not present in the original description and implementation of the ZETA-test, was to discard any data that was outside all event-window bounds (w_k, $w_k + \tau$) for all events $k=1 \ldots q$. In the original definition, a spike time x_i for which $w_k + \tau < x_i < w_{k+1}$ holds true, was not included in the calculation of the real zeta metric, but could be included in the random resamplings if also $w_k + \tau < x_i < w_k + 2\tau$ or $w_{k+1} - \tau < x_i < w_{k+1}$. Note that this will only occur if τ is chosen to be shorter than the longest inter-event duration. In the original definition of ZETA, we implicitly assumed a fixed inter-event duration equal to τ, but this does not always hold true. For example, when applying the zeta-test to self-generated behavioural events such as licking, inter-event durations will be highly variable.

Following the original ZETA definition, this means that in this case the population of spikes on which the deviation vectors are calculated, differs between the real zeta and random resamplings. Luckily, a simple remedy exists. We can perform "data-stitching" prior to running the random bootstrapping procedure. Specifically, between trial k and $k+1$, we remove all data for which

$$w_k + \tau < x_i \leq w_{k+1} \qquad 15$$

To correct for the removed time, we redefine v as

$$v_i = x_i^* - w_k^* \qquad 16$$

With

$$x_i^* = x_i - \sum_{j=1}^{k-1} e_j \qquad 17$$

And

$$w_k^* = w_k - \sum_{j=1}^{k-1} e_j \qquad 18$$

Where

$$e_j = \begin{cases} w_{j+1} - w_j - \tau & \text{if } w_{j+1} > w_j + \tau \\ 0 & \text{otherwise} \end{cases} \quad 19$$

This way, jittering w can no longer introduce new spikes that were not also used in the calculation of ζ_r, and our assumptions in the derivation of the zeta-test, and specifically its time-invariance, hold (Montijn et al., 2021).

Time-series ZETA to obtain super-sample time resolution

In contrast to the original ZETA-test, where all variables relate to point events v_i, time-series data consist of scalar values y_i sampled at time points t_i relative to some set of events, such as stimulus onsets. Therefore, in order to adapt the ZETA-test for use on time-series data, we first need to construct a reference sample time vector with respect to stimulus onset. In many time-series data sets, the sample times are not strictly synchronized to the event times of interest; for example, one may be recording calcium imaging data at some frequency (e.g., 15.01 Hz) while the screen that presents visual stimuli updates at 59.97 Hz. This means that the delay between the sample acquisitions and the event onsets can be variable across trials. We therefore define the reference time as the set of unique sample times over all q trials, where for each trial k the sample times are relative to the most recent stimulus onset w_k, and we set the tolerance for "uniqueness" to 1/100th of the median inter-sample duration. Note that this "super-resolution" is redundant in the case of slow signals, like calcium imaging, that are recorded at a constant acquisition rate. However, using this approach, the test can also easily be applied to data with a variable acquisition rate. Moreover, it can allow a finer determination of response onsets in cases where the time-limiting factor is the acquisition rate rather than the temporal dynamics of the signal (for example voltage imaging with fast sensors).

We start the time-series ZETA-test by defining a reference time vector $\boldsymbol{T}_k$ as the set of sample times t_i between the k'th event and the k'th event + τ:

$$\boldsymbol{T}_k = \{t_1\ t_2 \dots t_n\} - w_k \quad 20$$

With

$$w_k \leq t_i \leq w_k + \tau \quad 21$$

Therefore, for each event, we obtain a set of sample times, and we define the reference sample-time vector as the union over these sets:

$$\boldsymbol{r} = \{\boldsymbol{T}_1 \cup \boldsymbol{T}_2 \cup \dots \cup \boldsymbol{T}_q\} \quad 22$$

For a single trial with sample times $\boldsymbol{T}$ and values $\boldsymbol{v}$, we linearly interpolate the data to the reference time $\boldsymbol{r}$:

$$y_i = (1 - w)v_{j-1} + wv_j \tag{23}$$

Where

$$w = \frac{r_i - T_{j-1}}{T_j - T_{j-1}} \tag{24}$$

With j chosen such that

$$T_{j-1} \leq r_i \leq T_j \tag{25}$$

We repeat this for each of the q trials, obtaining a [q by n] matrix $\boldsymbol{Y}$ of data values, with n=|$\boldsymbol{r}$| denoting the cardinality of $\boldsymbol{r}$. Taking the mean over trials to obtain $\hat{\boldsymbol{y}}$ and rescaling the range to [0 1], we reduce this to a scaled vector $\boldsymbol{u}$ with n elements:

$$u_i = \frac{\hat{y}_\mathrm{i} - \min(\hat{\boldsymbol{y}})}{\max(\hat{\boldsymbol{y}}) - \min(\hat{\boldsymbol{y}})} \tag{26}$$

Now we replace the fractional spike position in $\boldsymbol{g}$ (Eq. 3) with a cumulative sum $\boldsymbol{s}$ over $\boldsymbol{u}$:

$$s_i = \frac{\sum_{j=1}^{i}(u_j)}{\sum_{h=1}^{n}(u_h)} \tag{27}$$

Furthermore, in the time-series case, if the values would be independent of time, the expected value at any time point is identical and equal to the average value of $\boldsymbol{u}$. Then the linear baseline density vector $\boldsymbol{b}$ (Eq. 4) corresponding to a non-modulated cumulative sum $\boldsymbol{s}$, is simply

$$b_i = \frac{i}{n} \tag{28}$$

Therefore, for the time-series ZETA (T-ZETA), the temporal deviation $\boldsymbol{\delta}$ is given by:

$$\delta_i = s_i - b_i \tag{29}$$

The remaining steps are identical to the original ZETA calculation (Eqs. 6 - 13).

A two-sample ZETA-test

The ZETA-test can be used to detect whether the response of a neuron is modulated by sensory, optogenetic, or behavioural events. But in many cases, researchers might be interested in whether two sets of stimuli lead to a different modulation of the same neuron; or whether two neurons respond differently to the same stimulus. We therefore developed a two-sample version of the ZETA-test. In the following paragraph we will refer to condition α and condition β as two sets of data we are comparing, which can be the response of one neuron to two sets of

events, the response of two neurons to the same set of events, or the response of two neurons during two different sets of events. The only requirement we place on the data is that the temporal window τ is the same for both conditions. We will therefore be working with the event times and event-relative spike times $\boldsymbol{w}^{\alpha}$ and $\boldsymbol{v}^{\alpha}$ of condition α, and $\boldsymbol{w}^{\beta}$ and $\boldsymbol{v}^{\beta}$ of condition β. Note that here, and in the rest of the manuscript, we use the α and β superscripts as condition indicators and not as an exponent.

A first intuitive approach to constructing a two-sample ZETA-test might be to compare the temporal deviations $\boldsymbol{d}$ (Eq. 6) between the two conditions, but this will fail when the two conditions only differ in absolute spiking rates. Since $\boldsymbol{d}$ is based on normalized spiking position $\boldsymbol{g}$, the absolute number of spikes is discarded in Equation 3. However, collapsing all spikes across trials without normalization would lead to differences in total spiking numbers when there is a difference in number of trials, but not spiking rate. We therefore construct a cumulative spiking vector for each condition based on the average number of spikes per event-window:

$$c_i = i/q \tag{30}$$

With q being the number of events and n the total number of spikes, this means that $c_n=n/q$, the average number of spikes per event. Note that dividing c_n by τ would therefore yield the average spiking rate per second. We perform this calculation for both conditions, obtaining $\boldsymbol{c}^{\alpha}$ and $\boldsymbol{c}^{\beta}$. To compare these two vectors, we first create a reference time-vector $\boldsymbol{\rho}$ that contains the spike times of both $\boldsymbol{v}^{\alpha}$ and $\boldsymbol{v}^{\beta}$:

$$\boldsymbol{\rho} = \{\boldsymbol{v}^{\alpha} \cup \boldsymbol{v}^{\beta}\} \tag{31}$$

Then we linearly interpolate $\boldsymbol{c}$ to $\boldsymbol{\rho}$ for condition α:

$$c^{\alpha *} = (1 - w)c_{j-1}^{\alpha} + wc_j^{\alpha} \tag{32}$$

With

$$w = \frac{\rho_i - v_j^{\alpha}}{v_{j-1}^{\alpha} - v_j^{\alpha}} \tag{33}$$

And we repeat the same procedure for $c^{\beta *}$. Now we can compute the difference in cumulative spike counts by simply taking the difference between the two $\boldsymbol{c}^{*}$ vectors:

$$\boldsymbol{\Delta} = \boldsymbol{c}^{\alpha *} - \boldsymbol{c}^{\beta *} \tag{34}$$

We then obtain our raw two-sample ZETA metric Z_r:

$$\mathrm{Z}_r \equiv \max\left(|\boldsymbol{\Delta} - \overline{\boldsymbol{\Delta}}|\right) \tag{35}$$

Similar to the one-sample ZETA-test, we now need to normalize Z_r to the intrinsic variability of $\boldsymbol{v}^{\alpha}$ and $\boldsymbol{v}^{\beta}$. Jittering works well for one-sample comparisons, but this does not satisfy the assumptions of our null hypothesis in the case of two-sample comparisons: we do not wish to know if the difference between condition α and β is larger than if they were both unmodulated. Rather, we wish to know if one modulation pattern is different from the other.

Therefore, to construct a null-hypothesis distribution of ZETA values in the two-sample test, we take random trials from the unified set of α and β. First we separate all spikes into distinct sets for each event k, such that

$$\boldsymbol{a}_k = \boldsymbol{b}_k - w_k \qquad 36$$

Where

$$\boldsymbol{b}_k = \{x_i | w_k < x_i \leq w_{k+1}\} \qquad 37$$

Therefore, each of the q^{α} and q^{β} events correspond to a set of event times $\boldsymbol{a}$, constituting a unified set $\boldsymbol{A}$ from which we will randomly select trials to generate our null-hypothesis samples:

$$\boldsymbol{A} = \{\boldsymbol{a}^{\alpha} \cup \boldsymbol{a}^{\beta}\} \qquad 38$$

To create a random null-hypothesis sample m for condition α, we will take q^{α} random sets from $\boldsymbol{A}$ with replacement and recombine them to create a null-hypothesis version of $\boldsymbol{v}$:

$$\boldsymbol{v}'^{\alpha,m} = \{\boldsymbol{A}_{\sigma_m(1)} \cup \boldsymbol{A}_{\sigma_m(2)} \cup \ldots \cup \boldsymbol{A}_{\sigma_m(q^{\alpha})}\} \qquad 39$$

Here, $\boldsymbol{\sigma}_m$ is a vector containing the q^{α} random integers in the range $[1, q^{\alpha} + q^{\beta}]$. We repeat the procedure for condition β, and for each random sample m, we obtain two shuffle-randomized spike vectors $\boldsymbol{v}'^{\alpha}$ and $\boldsymbol{v}'^{\beta}$. We then plug these into Equations 30 - 35 to generate a null-hypothesis ZETA sample Z', and use Equations 13 - 14 to transform these values into a statistical metric and corresponding p-value.

The one-sample ZETA-test is a special case of the two-sample ZETA-test

One can find an equivalence between the one- and two-sample tests in the case where condition α is a set of events, and condition β represents the linear baseline. This linear baseline can be approximated by an infinite set of events occurring over the same time window as the events in condition α. This means that all temporal modulation is averaged out, and the resulting cumulative spike count of condition β becomes identical to linear baseline density vector $\boldsymbol{b}$ from Equation 4 if sampled at $\boldsymbol{v}^{\alpha}$, and divided by τ. Randomly selecting q^{α} sets from $\boldsymbol{A}$ (Eq. 38), when $q^{\beta} = \infty$, therefore means selecting q^{α} random event-times, which is identical to jittering event onsets with an infinite jittering window. For condition β we take infinite samples from infinite

events, so $\boldsymbol{v}^{\beta}$ will remain identical to linear baseline density vector $\boldsymbol{b}$. Subtracting these vectors yields $\boldsymbol{\delta}'$ (Eq. 11). The steps thereafter (Eq. 12-14) are the same for the one-sample and two-sample tests, so they are therefore equivalent under these conditions.

Note that sampling at $\boldsymbol{v}^{\alpha}$ is only strictly necessary to make the one-sample test computationally tractable, but this does not influence its mathematical behaviour. Moreover, a division by τ can be added to the two-sample test without affecting its statistical properties – it merely scales the values in $\boldsymbol{A}$. The only real difference between the one-sample-equivalent version of the two-sample test, and the actual one-sample test is therefore the width of the jittering window. In data sets that are non-stationary across event repetitions, the local jittering of the one-sample test ensures uniform sampling over the signal's non-stationarity, but this is absent from the two-sample zeta-test. On the other hand, when signals are stationary across events, the one-sample-equivalent two-sample test and the one-sample test are truly mathematically equivalent.

Two-sample time-series ZETA-test

Having described the time-series ZETA-test and the two-sample ZETA-test, the two-sample time-series ZETA-test is a straightforward combination of these procedures. Rather than calculating the deviation vector $\boldsymbol{\delta}$ between the cumulative sum $\boldsymbol{s}$ and linear baseline vector $\boldsymbol{b}$, we calculate a cumulative sum for both condition α and condition β. We replace Eqs. 24-26 with

$$s_i^{\alpha} = \frac{\sum_{j=1}^{i}\left(y_j^{\alpha}\right)}{\sum_{h=1}^{n}\left(y_h^{\alpha}\right)} \qquad 40$$

$$s_i^{\beta} = \frac{\sum_{j=1}^{i}\left(y_j^{\beta}\right)}{\sum_{h=1}^{n}\left(y_h^{\beta}\right)} \qquad 41$$

Then the two-sample deviation vector $\boldsymbol{\Delta}$ becomes:

$$\Delta_i = s_i^{\alpha} - s_i^{\beta} \qquad 42$$

Now our raw two-sample time-series ZETA metric Z_r can be calculated as before by taking the maximum of the absolute of the mean-normalized deviations. To generate null-hypothesis samples, we repeat the above steps, but construct a [q by n] matrix $\boldsymbol{Y}$ for both conditions by randomly selecting q^{α} or q^{β} trials from the unified set of $q^{\alpha} + q^{\beta}$ possible trials for each random sample. Finally, we obtain the statistical metric using Eq. 13 - 14.

Experimental data

The calcium imaging data analysed in this paper are the same as previously described for both the GCaMP6f (Montijn, Meijer, et al., 2016) and OGB-1 AM recordings (Montijn, Goltstein, et al., 2016). We used GCaMP data from 15 recordings in four C57BL/6 mice, and OGB data from 8 recordings in 8 C57BL/6 mice. Cell bodies were detected semi-automatically using an open-source toolbox (https://github.com/JorritMontijn/Preprocessing_Toolbox) and cells were only included for further analysis if their somata were clearly distinguishable from the background neuropil. The electrophysiological data used here are also previously described elsewhere (Montijn et al., 2023). In short, we performed 21 repeated-insertion recordings in seven C57BL/6 mice with Neuropixels. All 19 mice were housed in a 12 h/12 h dark/light cycle with ad libitum access to food and water and were awake during recording. The recording setup presented drifting gratings of 24 directions (spaced in 15-degree steps) and was controlled using Acquipix (Montijn, 2022). Spikes were sorted post-hoc using Kilosort 2.5 (Pachitariu et al., 2023), and electrode location was determined by aligning histological slices and neurophysiological landmarks to the AllenCCF mouse brain atlas (https://github.com/cortex-lab/allenCCF) using the UniversalProbeFinder (Montijn & Heimel, 2022). All code used in the Neuropixels data acquisition and pre-processing is available online in the Acquipix and UniversalProbeFinder repositories on https://github.com/JorritMontijn. We included only clusters of sufficient quality, as quantified by their spike contamination ratio and non-stationarity, for further analysis. For more detailed information, see (Montijn et al., 2023). 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.

Experimental benchmarks one-sample tests

We verified the statistical performance of the one-sample ZETA test and one-sample time-series ZETA test tests by calculating the p-value for stimulus responsiveness for all neurons. The false positive rate was computed by repeating the procedure after jittering the stimulus onset times.

Experimental benchmarks two-sample tests

For the two-sample ZETA test we calculated whether random pairs of neurons differed in their response (n=1000 pairs) to the presentation of a drifting grating (figure 3E) or whether the same

neuron differed in response between two random directions of drifting gratings (figure 3F). In both cases, we used a randomly selected 50% of spikes from either neuron, so we could calculate the false positive rate by instead comparing the same 50% of spikes from the first neuron to the remaining 50% of spikes from the same neuron.

For the time-series-based two-sample tests, we used a similar comparison, except we took 50% of trials rather than spikes (as selecting 50% of samples would run into various difficulties). The false positive rate was then computed by repeating the procedure but comparing half of the trial responses of the first neuron to the remaining half.

Synthetic benchmarks one-sample ZETA-test

We quantified the effect of the data-stitching procedure by generating spiking responses of four types of synthetic neurons. All types of synthetic neurons used exponential inter-spike intervals as firing distributions. For the first type of neuron, its baseline firing rate was determined randomly by sampling from an exponential distribution with a mean of 5.0 Hz. The neuron was assigned a random preferred orientation and preferred stimulus spiking rate ($\lambda pref = \lambda base + Exp(\lambda=5)$). At its non-preferred stimuli, the neuron's firing rate was equal to its baseline rate, and it was equal to $\lambda pref$ for a stimulus of its preferred orientation, with intermediate values following a von Mises curve with random tuning bandwidth (κ=5-10). We generated 20 repetitions of 8 stimulus directions and created heterogeneous inter-stimulus durations by varying these durations from 200ms to 2 seconds. The stimulus-on period was fixed to 1 second, however, leading to both overlapping responses and varying inter-trial intervals.

The second type of simulation was a tri-phasic neuron, for which we generated 160 trials of varying and partially overlapping duration, ranging from 0.5 to 10 seconds. Each trial consisted of an onset period of 100 ms, followed by a 900 ms sustained period, after which the neuron returned to its baseline rate. For each neuron the baseline rate was chosen to be $\lambda base=Exp(0.1)$, the onset rate $\lambda onset= \lambda base+Exp(1)+0.2$, and the sustained rate $\lambda sust= \lambda base+Exp(0.1)$.

The third type was a tri-phasic neuron we constructed to illustrate the effect that stitching versus no-stitching can have. Again, we used 160 trials, all of 1 second in duration, with a 3 second inter-trial period. The base rate of this neuron during inter-trial intervals was Exp(0.1)+0.1, the onset rate (100 ms) was Exp(4)+4, and the sustained rate (900 ms) was Exp(2)+2. Note that by setting the trial duration to 1 second, ZETA's real deviation does not

cover the transition into and out of the baseline firing rate, while the jittered deviations will include either of these transitions.

To illustrate the worst-case scenario for the ZETA-test, we simulated a (rather unrealistic) neuron that had a two-phase firing distribution, where the spike counts were chosen in blocks of 1 second from an i.i.d. Gaussian distribution with a mean of 3 (for 3 out of 4 seconds), or an i.i.d. Gaussian distribution with a mean of 3.1+ Exp(1) (for 1 out of 4 seconds). We dispersed the resulting number of spikes randomly thoroughly the 1-second block following a uniform distribution. This procedure generates statistics that satisfy the assumptions of the ANOVA's model.

Synthetic benchmarks two-sample tests

To further investigate the statistical properties of our methods, we also used two types of generated synthetic data sets. In both cases, we generated neurons with a baseline rate as described above; an exponential inter-spike interval model with a rate randomly chosen from $\lambda \sim$ Exp(1). Furthermore, one additional spike was added to half of all trials on top of these baseline spikes. The timing of these spikes followed a Gaussian distribution with a mean of 55 ms and a standard deviation of 1 ms. This therefore simulates a cell with a sharp onset peak.

For the first benchmark, we then generated a neuron with the same baseline firing rate, but which had an onset spike in all trials – it therefore differed in both peak height and total number of spikes (owing to the higher peak). For the second benchmark, we instead generated a neuron with the same baseline firing rate and same amount of onset spikes, but whose onset peak was instead shifted by 2 ms – this means the neurons differed in onset peak latency but not in total number of spikes. In both cases, the two-sample tests (ZETA, ANOVA, t-test) were then applied to obtain a p-value. False positive rates were calculated by generating the second neuron with the exact same parameters as the first neuron: the same number of onset spikes occurring at the same peak time.

Acknowledgements

We thank Valeria Gazolla and Christian Keysers for suggesting to pursue a ZETA-test for time-series data. This work was funded by a Royal Netherlands Academy of Arts and Sciences (KNAW) Fonds KNAW-Instituten grant.

References

Burnett, L. E., Koppensteiner, P., Symonova, O., Masson, T., Vega-Zuniga, T., Contreras, X., Rülicke, T., Shigemoto, R., Novarino, G., & Jösch, M. (2023). *Subcortical circuit dysfunctions delay perceptual decision-making in autism models* (p. 2022.10.11.511691). bioRxiv. https://doi.org/10.1101/2022.10.11.511691

Dudok, B., Szoboszlay, M., Paul, A., Klein, P. M., Liao, Z., Hwaun, E., Szabo, G. G., Geiller, T., Vancura, B., Wang, B.-S., McKenzie, S., Homidan, J., Klaver, L. M. F., English, D. F., Huang, Z. J., Buzsáki, G., Losonczy, A., & Soltesz, I. (2021). Recruitment and inhibitory action of hippocampal axo-axonic cells during behavior. *Neuron*, *109*(23), 3838-3850.e8. https://doi.org/10.1016/j.neuron.2021.09.033

Montijn, J. S. (2022). *JorritMontijn/Acquipix: Acquipix v0.9.0* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.7104646

Montijn, J. S., Goltstein, P. M., & Pennartz, C. M. (2016). Mouse V1 population correlates of visual detection rely on heterogeneity within neuronal response patterns. *eLife*, *4*, e10163. https://doi.org/10.7554/eLife.10163

Montijn, J. S., & Heimel, J. A. (2022). *A universal pipeline for the alignment of electrode tracks with slice histology and electrophysiological data* (p. 2022.06.20.496782). bioRxiv. https://doi.org/10.1101/2022.06.20.496782

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., Riguccini, V., Levelt, C. N., & Heimel, J. A. (2023). Impaired Direction Selectivity in the Nucleus of the Optic Tract of Albino Mice. *Investigative Ophthalmology & Visual Science*, *64*(11), 9. https://doi.org/10.1167/iovs.64.11.9

Montijn, J. S., Seignette, K., Howlett, M. H., Cazemier, J. L., Kamermans, M., Levelt, C. N., & Heimel, J. A. (2021). A parameter-free statistical test for neuronal responsiveness. *eLife*, *10*, e71969. https://doi.org/10.7554/eLife.71969

Oude Lohuis, M. N., Marchesi, P., Olcese, U., & Pennartz, C. (2022). *Triple dissociation of visual, auditory and motor processing in primary visual cortex* (p. 2022.06.29.498156). bioRxiv. https://doi.org/10.1101/2022.06.29.498156

Pachitariu, M., Sridhar, S., & Stringer, C. (2023). *Solving the spike sorting problem with Kilosort* (p. 2023.01.07.523036). bioRxiv. https://doi.org/10.1101/2023.01.07.523036

Qin, Y., Ahmadlou, M., Suhai, S., Neering, P., de Kraker, L., Heimel, J. A., & Levelt, C. N. (2023). Thalamic regulation of ocular dominance plasticity in adult visual cortex. *eLife*, *12*, RP88124. https://doi.org/10.7554/eLife.88124

Schneider, M., Tzanou, A., Uran, C., & Vinck, M. (2023). Cell-type-specific propagation of visual flicker. *Cell Reports*, *42*(5), 112492. https://doi.org/10.1016/j.celrep.2023.112492

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

Spyropoulos, G., Schneider, M., Kempen, J. van, Gieselmann, M. A., Thiele, A., & Vinck, M. (2023). *Distinct feedforward and feedback pathways for cell-type specific attention effects* (p. 2022.11.04.515185). bioRxiv. https://doi.org/10.1101/2022.11.04.515185

Szadzinska, W., Danielewski, K., Kondrakiewicz, K., Andraka, K., Nikolaev, E., Mikosz, M., & Knapska, E. (2021). Hippocampal Inputs in the Prelimbic Cortex Curb Fear after Extinction. *Journal of Neuroscience*, *41*(44), 9129–9140. https://doi.org/10.1523/JNEUROSCI.0764-20.2021

Ziegler, K., Folkard, R., Gonzalez, A. J., Burghardt, J., Antharvedi-Goda, S., Martin-Cortecero, J., Isaías-Camacho, E., Kaushalya, S., Tan, L. L., Kuner, T., Acuna, C., Kuner, R., Mease, R. A., & Groh, A. (2023). Primary somatosensory cortex bidirectionally modulates sensory gain and nociceptive behavior in a layer-specific manner. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-38798-7

Figures

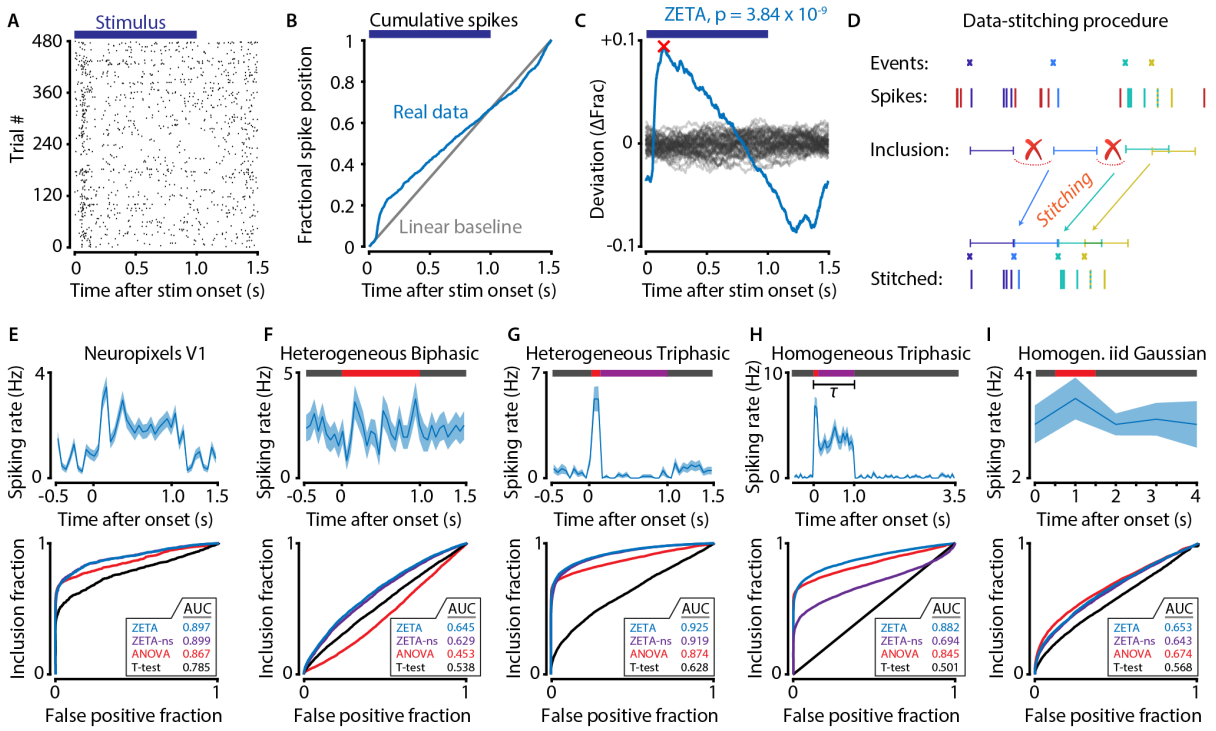

Figure 1. Improvements to the ZETA-test. A) Raster plot of an example neuron. B) Calculating the statistical metric for the ZETA-test depends on the difference between the real fractional spike positions (blue) and the null-hypothesis expectation from a constant, stimulus-unmodulated rate (grey). C) The Zenith of Event-based Time-locked Anomalies (ZETA, red cross) defines the significance after normalizing for the neuron's intrinsic variability (grey curves). D) Overview of the new data-stitching procedure, where time periods are removed if they were not used for the calculation of the real deviation curve in panel B. E-I) Under specific conditions data stitching can dramatically improve ZETA's performance. E) Experimental Neuropixels data shows the ZETA-test is superior to an optimal ANOVA and t-test, but stitching has little effect. ZETA indicates test with stitching, ZETA-ns without stitching. F,G) Two neuron models responding to stimuli with heterogeneous inter-trial intervals all show the ZETA-test performs excellently, and the ZETA-test with data-stitching shows a small, but significant improvement. H) If the window of interest (τ) is chosen such that it discards transitions into and out of baseline activity, the sensitivity of the ZETA-test with no stitching is mediocre. Performing Data-stitching improves the sensitivity of the ZETA-test to be superior to the ANOVA. I) Even under statistical conditions optimized for the ANOVA, the ZETA-test's sensitivity remains high.

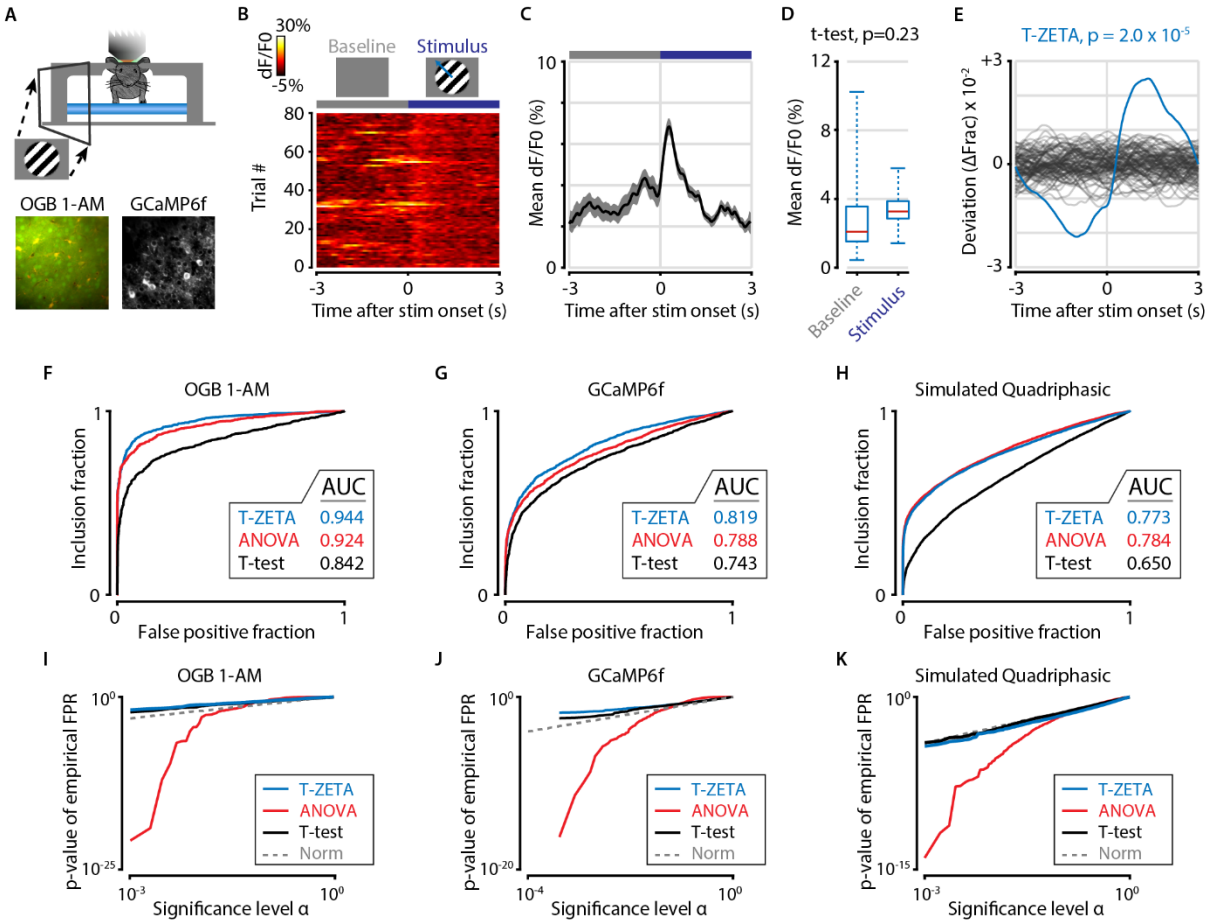

Figure 2. The time-series ZETA-test outperforms the t-test and ANOVA on statistical sensitivity, while the ANOVA applied to calcium data is excessively liberal. A) The time-series ZETA-test can be applied to data recorded with calcium imaging. B) An example cell's calcium activity (dF/F0) recorded with OGB in response to drifting gratings. C) This cell shows a clear onset response to the stimulus (period indicated by the blue bar), but also displays variable spontaneous activity and a lack of a sustained response. D) These factors reduce the difference in mean dF/F0 between the 3 s pre-stimulus baseline and 3 s stimulus periods, leading a t-test to erroneously classify this cell as non-responsive (paired t-test, n = 80 trials, p=0.23). E) Our alternative method (T-ZETA) does not use window-averages and that detects any time-locked deviations in neural activity. The blue curve shows the true deviation from a static level, and the grey curves show 100 bootstraps of deviations obtained by randomly jittering the stimulus onsets. F-K) ROC analyses to benchmark the statistical sensitivity under different conditions. F) Performance benchmark on OGB data showing the statistical sensitivity of the T-ZETA test (blue), an optimally binned ANOVA (red), and a t-test of 3s pre- vs 3s post- stimulus onset activity (black). G) Same as F, but for GCaMP6f data. H) Same as F, but for simulated quadri-

phasic neurons with distinct baseline, onset, sustained, and offset responses where we filtered generated spike times with an exponential filter to simulate the effect of the slow dynamics of a calcium indicator. I-K) Control analyses show the false positive rate (FPR) as a function of the threshold value alpha. I,J) In real data, the T-ZETA test is somewhat conservative, as lies above the theoretical norm (dotted line), but owing to its higher statistical sensitivity, it still shows higher inclusion rates than a t-test for all alphas. On the other hand, the ANOVA is a poor statistical model when data points are not statistically independent: it is too liberal by many orders of magnitude, as the filtering properties of the calcium indicators induce temporal dependencies in the signal. K) Simulations confirm the cause of the ANOVA's poor performance is unrelated to particulars of real data. Here, the T-ZETA and t-test are close to the theoretical norm, but the ANOVA is still overly liberal by a factor of 10^{11} in the case of an alpha of 0.001.

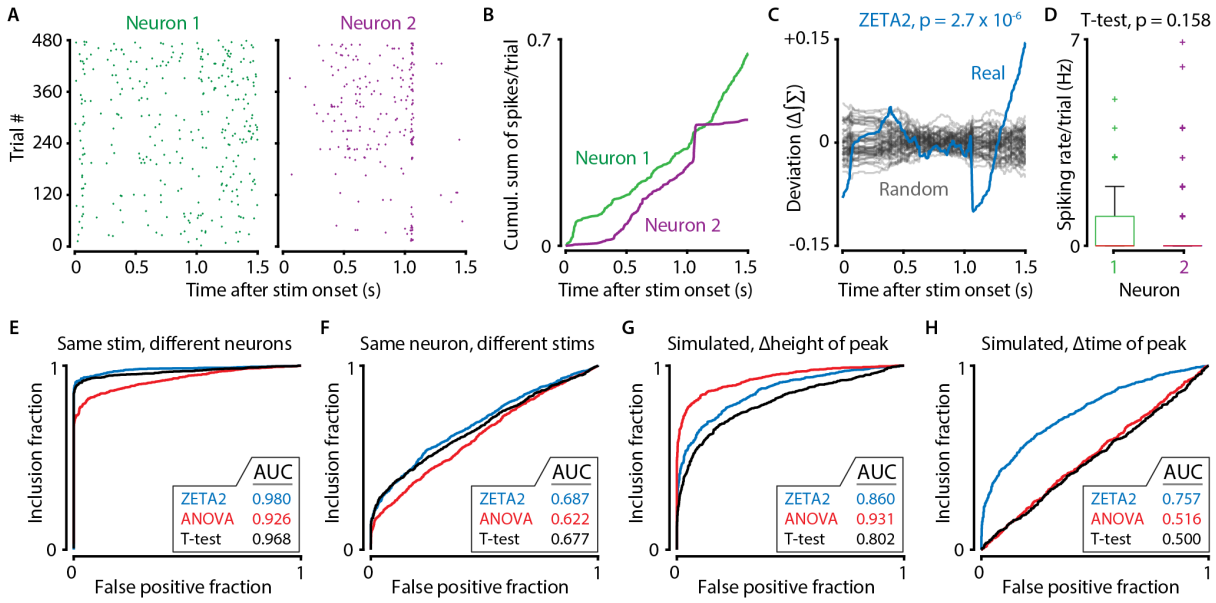

Figure 3. The two-sample ZETA-test quantifies response differences between two conditions. A) Example spike raster plots of two neurons responding drifting grating stimuli. B) The statistical metric for the two-sample ZETA-test depends on the difference in the cumulative sum of the spikes per trial for the two conditions. C) The real difference (blue) is compared to the difference obtained using resamplings where trials are randomly assigned to condition 1 (here neuron 1) or 2. D) The two-sample ZETA-test (ZETA2) detects a difference between these two example neurons ($p=2.7 \times 10^{-6}$), but a two-sample t-test does not (p=0.158). E-H) Various benchmarks to compare the performance of the ZETA2-test to an optimally-binned two-way ANOVA and t-test. E) Discrimination of 1000 pairs of V1 neurons recorded with Neuropixels in response to the same stimuli; the ZETA2-test performs best, followed by the t-test and ANOVA. F) Discrimination between the responses to drifting gratings in the 0 and 90 degree directions for all 1504 experimentally recorded cells; the ZETA2-test performs best, followed by the t-test and ANOVA. G) Simulation of the best-case scenario for the ANOVA, where the difference between two conditions is defined only by the number of spikes in a short response peak. H) Worst-case scenario for the ANOVA, where the difference between two conditions is defined only by a 2 ms difference in peak time.

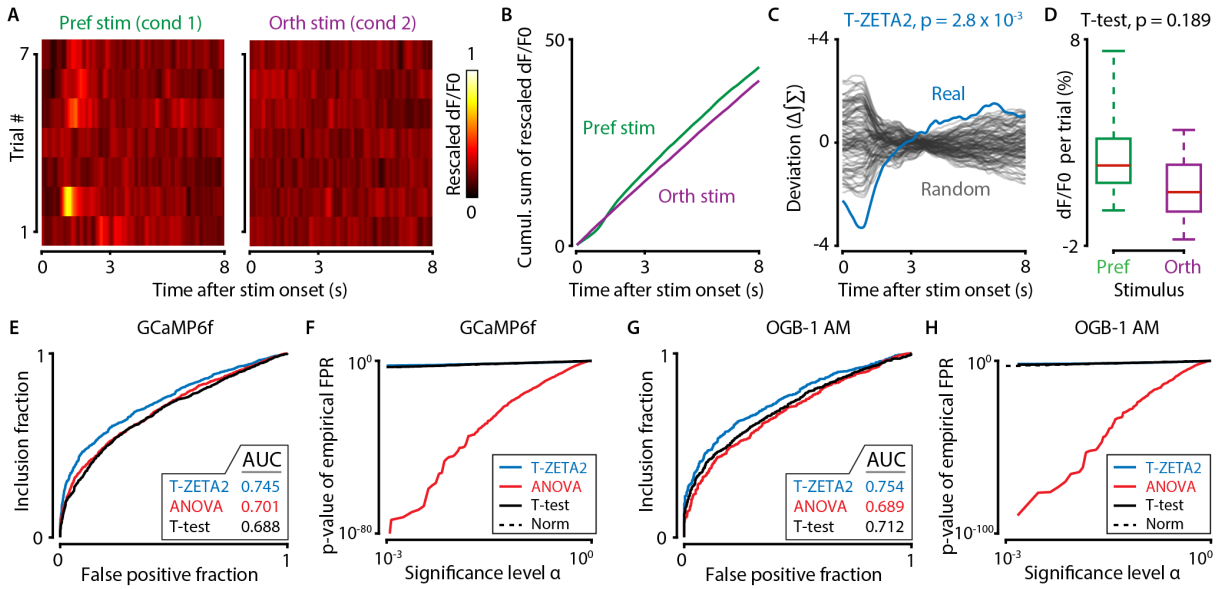

Figure 4. The two-sample time-series ZETA-test (T-ZETA2) can discriminate neural activity in calcium imaging data. A) Heat maps of one example V1 neuron's response to drifting gratings moving in its preferred and orthogonal direction. B) The T-ZETA2-test uses the cumulative sum of rescaled neural activity (here dF/F0). C) The deviation in cumulative sum defines the ZETA-metric and is compared to deviation curves obtained after randomly combining trials from either condition, similar to the spike-based ZETA2-test. D) In this example, the T-ZETA2-test detected a significant difference in responses between the two conditions ($p=2.8 \times 10^{-3}$) but a t-test did not (p=0.189). E) Benchmark discriminating between the preferred and orthogonal stimulus responses of all GCAMP neurons. The T-ZETA2 performed best, followed by the ANOVA and t-test. F) The false-positive rate of the T-ZETA2 and t-test were close to the theoretical norm, but the ANOVA was excessively liberal, showing the ANOVA is unsuitable for time-series data with temporal correlations. G,H) As E,F, but for OGB data, showing almost identical results.