

On the use of calcium deconvolution algorithms in practical contexts

Mathew H. Evans^{1,2}, Rasmus S. Petersen² & Mark D. Humphries^{1,2}

¹School of Psychology, University of Nottingham, UK

²Faculty of Biology, Medicine and Health, University of Manchester, UK

July 25, 2019

Abstract

Calcium imaging is a powerful tool for capturing the simultaneous activity of large populations of neurons. Studies using it to address scientific questions of population dynamics and coding often use the “raw” time-series of changes in calcium fluorescence at the soma. But somatic calcium traces are both contaminated with multiple noise sources and are non-linearly related to spiking. A suite of methods are available to recover spike-evoked events from the raw calcium, from simple deconvolution to inferring the spikes themselves. Here we explore the extent to which our choice of raw or deconvolved calcium time-series affects the scientific inferences we can draw. Our results show the choice qualitatively changes the potential scientific inferences we draw about neural activity, coding, and correlation structure. We show that a substantial fraction of the processing methods fail to recover simple features of population activity in barrel cortex already established by electrophysiological recordings. Raw calcium time-series contain an order of magnitude more cells tuned to task features; yet there is also qualitative disagreement between deconvolution methods on which neurons are tuned. Finally, we show that raw and processed calcium time-series qualitatively disagree on the structure of correlations within the population and the dimensionality of its joint activity. We suggest that quantitative results obtained from population calcium-imaging be verified across multiple forms of the calcium time-series.

1 Introduction

Calcium imaging is a wonderful tool for high yield recordings of large neural populations (Harris et al., 2016; Stringer et al., 2019; Ahrens et al., 2013; Portugues et al., 2014). Many pipelines are available for moving from pixel intensity across frames of video to a time-series of calcium fluorescence in the soma of identified neurons (Mukamel et al., 2009; Vogelstein et al., 2010; Kaifosh et al., 2014; Pachitariu et al., 2016; Deneux et al., 2016; Pnevmatikakis et al., 2016; Friedrich et al., 2017; Keemink et al., 2018; Giovannucci et al., 2019).

But raw calcium fluorescence is nonlinearly related to spiking, and contains noise from a range of sources. These issues have inspired a wide range of deconvolution algorithms (Theis et al., 2016; Berens et al., 2018; Stringer and Pachitariu, 2018), which attempt to turn raw somatic calcium into something more closely approximating spikes. We address here the question facing any systems neuroscientist using calcium imaging: do we use the raw calcium, or attempt to clean it up? Thus our aim is to understand if our choice matters: how do our scientific inferences depend on our choice of raw or deconvolved calcium time-series.

36 Deconvolution algorithms themselves range in complexity from simple deconvolution
37 with a fixed kernel of the calcium response (Yaksi and Friedrich, 2006), through detecting
38 spike-evoked calcium events (Jewell and Witten, 2018; Pachitariu et al., 2016), to directly
39 inferring spike times (Vogelstein et al., 2010; Lütcke et al., 2013; Deneux et al., 2016).
40 This continuum of options raise the further question of the extent to which we should
41 process the raw calcium signals.

42 We proceed in two stages. In order to use deconvolution algorithms, we need to choose
43 their parameters. We'd like to know whether it is worth taking this extra step: how good
44 can these algorithms be in principle, and how sensitive their results are to the choice of
45 parameter values. We thus first evaluate qualitatively different deconvolution algorithms,
46 by optimising their parameters against ground truth data with known spikes. With our
47 understanding of their parameters in hand, we then turn to our main question, by analysing
48 a large-scale population recording from the barrel cortex of a mouse performing a whisker-
49 based decision task. We compare the scientific inferences about population coding and
50 correlations we obtain using either raw calcium signals, or a range of time-series derived
51 from those calcium signals, covering simple deconvolution, event detection, and spikes.

52 We find contrasting answers. A substantial fraction of the methods used here fail
53 to recover basic features of population activity in barrel cortex established from electro-
54 physiology. The inferences we draw about coding qualitatively differ between raw and
55 deconvolved calcium signals. In particular, coding analyses based on raw calcium signals
56 detect an order of magnitude more cells tuned to task features. Yet there is also qualitative
57 disagreement between deconvolution methods on which neurons are tuned. The inferences
58 we draw about correlations between neurons do not distinguish between raw and decon-
59 volved calcium signals, but can qualitatively differ between deconvolution methods. Our
60 results thus suggest care is needed in drawing inferences from population recordings of so-
61 matic calcium, and that one solution is to replicate all results in both raw and deconvolved
62 calcium signals.

63 2 Results

64 2.1 Performance of deconvolution algorithms on ground-truth data-sets

65 We select here three deconvolution algorithms that infer discrete spike-like events, each
66 an example of the state of the art in qualitatively different approaches to the problem:
67 Suite2p (Pachitariu et al., 2016), a peeling algorithm that matches a scalable kernel to the
68 calcium signal to detect spike-triggered calcium events; LZero (Jewell and Witten, 2018), a
69 change-point detection algorithm, which finds as events the step-like changes in the calcium
70 signal that imply spikes; and MLspike (Deneux et al., 2016), a forward model, which fits
71 an explicit model of the spike-to-calcium dynamics in order to find spike-evoked changes
72 in the calcium signal, and returns spike times. We emphasize that these methods were
73 chosen as exemplars of their approaches, and are each innovative takes on the problem;
74 we are not here critiquing individual methods, but using an array of methods to illustrate
75 the problems and decisions facing the experimentalist when using calcium imaging data.

76 We first ask if these deconvolution methods work well in principle. We fit the parame-
77 ters of each method to a data-set of 21 ground-truth recordings (Chen et al., 2013), where
78 the spiking activity of a cell is recorded simultaneously with 60 Hz calcium imaging using
79 high-signal-to-noise juxtacellular recording techniques (Figure 1a). To fit the parameters
80 for each recording, we sweep each method's parameter space to find the parameter value(s)
81 with the best match between the true and inferred spike train.

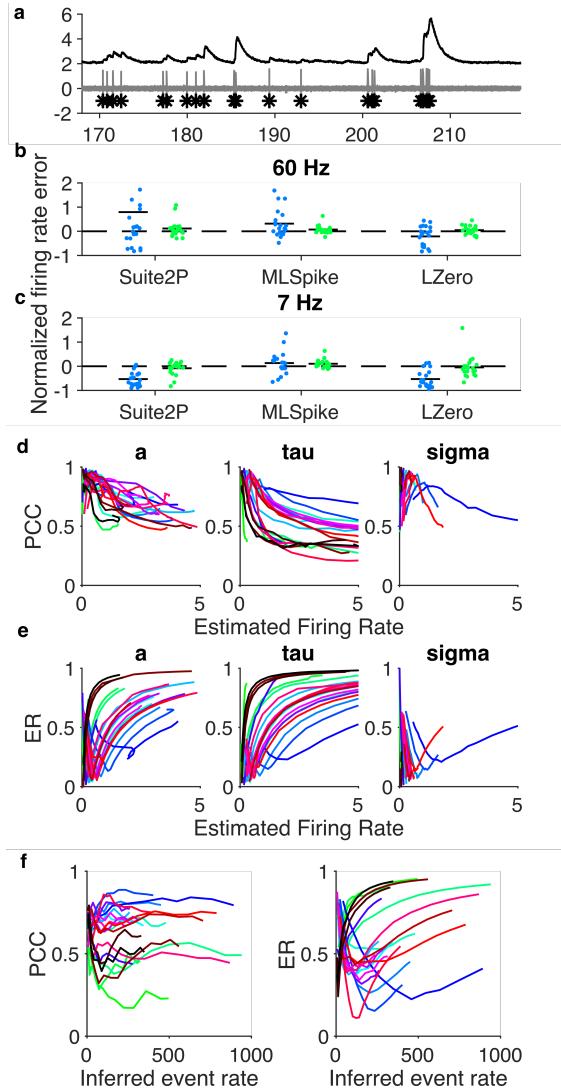


Figure 1: Ground truth data analysis.

- (a) Example simultaneous recording of somatic voltage (grey) and calcium activity (black) imaged at 60Hz. Spikes are marked with asterisks.
- (b) Error in estimating the true firing rate when using optimised parameters, across all three methods. One symbol per recording. We separately plot errors for parameters optimised to maximise the correlation coefficient (PCC), the errors for parameters optimised to minimise the error rate (ER). Horizontal black bars are means. Error is computed relative to the true firing rate: $(Rate_{true} - Rate_{estimated})/Rate_{true}$. For LZero and Suite2p, $Rate_{estimated}$ is computed from event times.
- (c) As for (b), but with the somatic calcium down-sampled to 7Hz before optimising parameters for the deconvolution methods.
- (d) Dependence of MLspike's deconvolution performance on the firing rate of the inferred spike train. For each of ML Spike's free parameters, we plot the correlation coefficient between true and inferred spikes as a function of the firing rate estimated from the inferred spikes. One line per recording. Parameters: A : calcium transient amplitude per spike ($\Delta F/F$); τ calcium decay time constant (s); σ : background (photonic) noise level ($\Delta F/F$)
- (e) as in (d), but using Error Rate between the true and inferred spikes.
- (f) Dependence of Suite2p's deconvolution performance on the firing rate of the inferred event train as a detection threshold parameter is varied. Left: correlation coefficient; right: Error Rate.

82 The best-fit parameters depend strongly on how we evaluate the match between true
83 and inferred spikes. The Pearson correlation coefficient between the true and inferred
84 spike train is a common choice (Brown et al., 2004; Paiva et al., 2010; Theis et al., 2016;
85 Reynolds et al., 2018; Berens et al., 2018), typically with both trains convolved with a
86 Gaussian kernel to allow for timing errors. However, we find that choosing parameters to
87 maximise the correlation coefficient can create notable errors. The inferred spike trains
88 from ML Spike have too many spikes on average (mean error: 31.72%), and the accuracy of
89 recovered firing rates widely varies across recordings (Fig 1b, blue symbols). We attribute
90 these errors to the noisy relationship between the correlation coefficient and the number
91 of inferred spikes (Figure 1c): for many recordings, there is no well-defined maximum
92 coefficient, especially for the amplitude parameter A , so that near-maximum correlation
93 between true and inferred trains is consistent with a wide range of spike counts in the
94 inferred trains. We see the same sensitivity for the event rates from recordings optimised
95 using Suite2p (Figure 1f). If we compare their inferred event rates to true firing rates (Fig
96 1b), we see Suite2p estimates far more events than spikes (mean error 79.47%) and LZero
97 fewer events than spikes (mean error: -21.14%). These further errors are problematic:
98 there cannot be more spike-driven calcium events than spikes, and LZero's underestimate
99 is considerably larger than the fraction of frames with two or more spikes ($<2e^{-4}\%$ frames).

100 To address the weaknesses of the Pearson correlation coefficient, we instead optimise
101 parameters using the Error Rate metric of Deneux et al. (2016). Error Rate returns a
102 normalised score between 0 for a perfect match between two spike trains, and 1 when all
103 the spikes are missed. This comparison between inferred and true spike trains is most
104 straightforward for algorithms like ML Spike that directly return spike times; for the other
105 algorithms, we use here their event times as inferred spikes, a reasonable choice given the
106 low firing rate and well separated spikes in the ground truth data. Choosing parameters
107 to minimise the Error Rate between the true and inferred spike-trains results in excellent
108 recovery of the true number of spikes for all three deconvolution methods (Fig 1b, green
109 symbols), with mean errors of 12% for Suite2P, 7.3% for ML Spike, and 5% for LZero.
110 As we show in Figure 1e for ML Spike and Figure 1f for Suite2p, the Error Rate has a
111 well-defined minima for almost every recording. Consequently, all deconvolution methods
112 can, in principle, accurately recover the true spike-trains given an appropriate choice of
113 parameters.

114 A potential caveat here is that the ground-truth data are single neurons imaged at a
115 frame-rate of 60Hz, an order of magnitude greater than is typically achievable in popula-
116 tion recordings (Peron et al., 2015a). Such a high frame-rate could allow for more accurate
117 recovery of spikes than is possible in population recordings. To test this, we downsample
118 the ground-truth data to a 7Hz frame-rate, and repeat the parameter sweeps for each
119 deconvolution method applied to each recording. As we show in Figure 1c, optimising pa-
120 rameters using the minimum Error Rate still results in excellent recovery of the true spike
121 rate (and interestingly for some recordings reduces the error when using the correlation
122 coefficient). Lower frame-rates need not then be an impediment to using deconvolution
123 methods.

124 **2.2 Parameters optimised on ground-truth are widely distributed and 125 sensitive**

126 What might be an impediment to using deconvolution methods on population recordings
127 is that the best parameter values vary widely between cells. Figure 2a-b plots the best-fit
128 parameter values for each recording across deconvolution methods and sampling rates.
129 Each method has at least one parameter with substantial variability across recordings,

130 varying by an order of magnitude or more. This suggests that the best parameters for one
 131 cell may perform poorly for another cell.

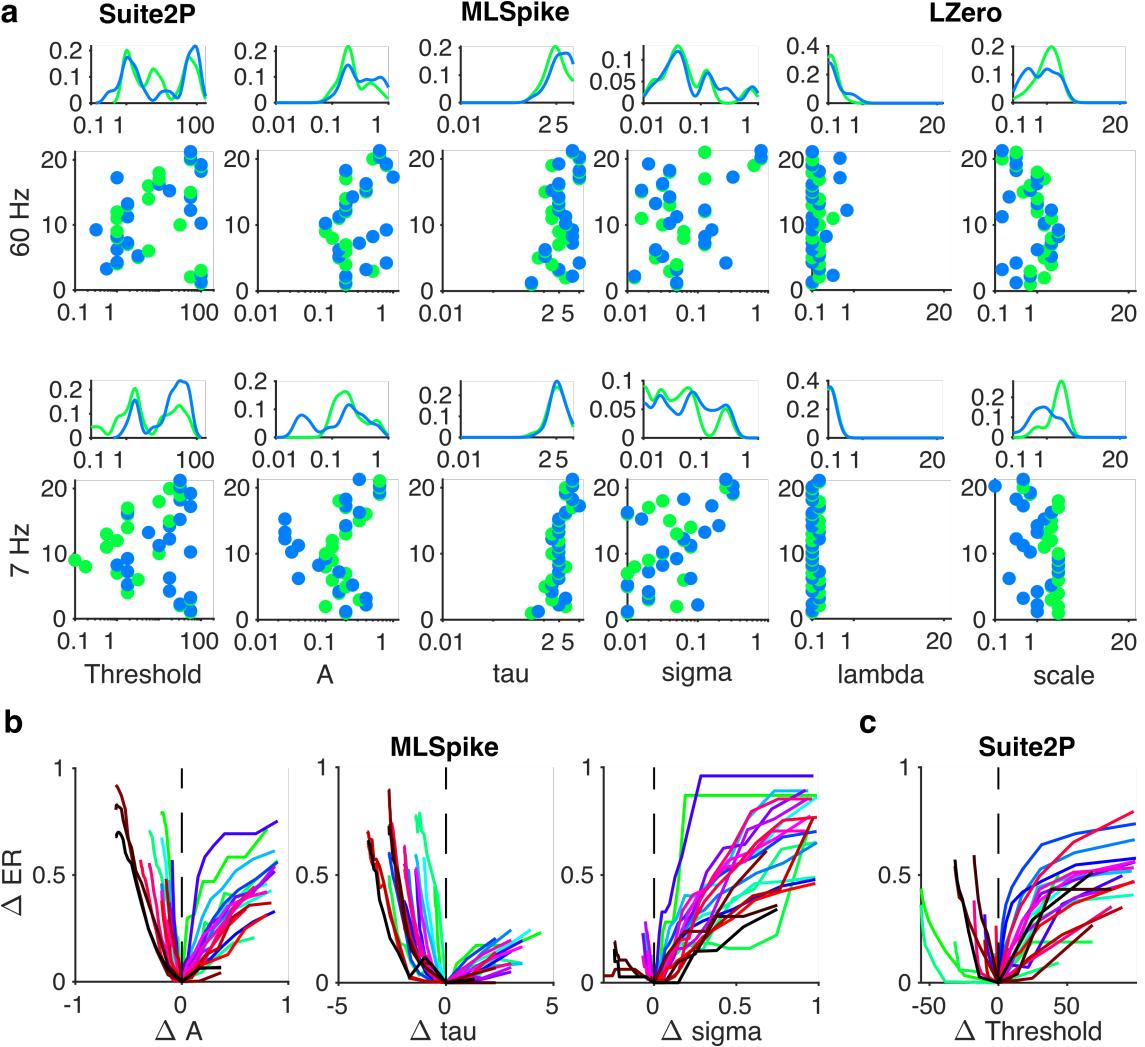


Figure 2: Variation in best-fit spike deconvolution parameters across ground-truth recordings.

(a) Distributions of optimised parameter values across recordings. In each panel, we plot parameter values on the x-axis against the recording ID on the y-axis (in an arbitrary but consistent order). Parameter values are plotted for those optimised using the Error Rate (green). Top row: fits to the original 60 Hz frame-rate data; bottom row: fits to data down-sampled to 7 Hz. [1: link recordings of the same cell what do you mean? Each cell is in its own y-axis position already, so horizontal lines between blue and green dots?]

(b) Change in error rate as a function of the change away from a parameter's optimum value, for each of ML Spike's free parameters. One line per recording.

(c) Change in the error rate with change in Suite2p's threshold value away from its optimum for each recording. One line per recording.

132 The problem of between-cell variation in parameter values would be compensated
 133 somewhat if the quality of the inferred spike or event trains is robust to changes in those
 134 values. However, we find performance is highly sensitive to changes in some parameters.
 135 Figure 2b-c shows that for most recordings the quality of the inferred spike train abruptly
 136 worsens with small increases or decreases in the best parameter. Thus using deconvolution
 137 algorithms on population recordings comes with the potential issues that parameters can

138 be both sensitive and vary considerably across cells.

139 2.3 Deconvolution of population imaging in barrel cortex during a de- 140 cision task

141 We turn now to seeing if and how these issues play out when analysing a large-scale
142 population recording with no ground-truth. The data we use are two-photon calcium
143 imaging time-series from a head-fixed mouse performing a whisker-based two-alternative
144 decision task (Fig. 3a-b), from the study of Peron et al. (2015b). We analyse here a single
145 session with 1552 simultaneously recorded pyramidal neurons in L2/3 of a single barrel
146 in somatosensory cortex, imaged at 7 Hz for just over 56 minutes, giving 23559 frames in
147 total across 335 trials of the task.

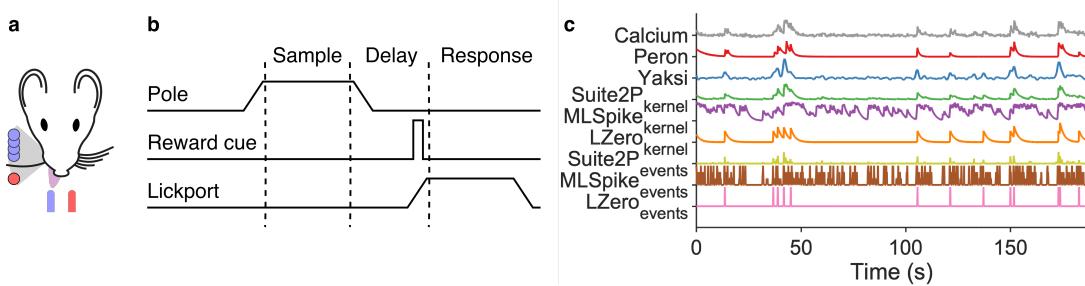


Figure 3: Experimental data from Peron et al. (2015b).

- (a) Schematic of task set-up. A pole was raised within range of the single left-hand whisker; its position, forward (red) or backward (blue) indicated whether reward would be available from the left or right lick-port.
- (b) Schematic of trial events. The pole was raised and lowered during the sample period; a auditory cue indicated the start of the response period.
- (c) All deconvolution methods applied to one raw calcium signal from the same neuron.

148 Our primary goal is to understand how the choices of deconvolving these calcium-
149 imaging data alter the scientific inferences we can draw. As our baseline, we use the
150 “raw” $\Delta F/F$ time-series of changes in calcium indicator fluorescence. We use the above
151 three discrete deconvolution methods to extract spike counts (MLSpike), event occurrence
152 (LZero), or event magnitude (Suite2p) per frame. For comparison, we use Peron et al.
153 (2015b)’s own version of denoised calcium time-series, created using a custom version of the
154 peeling algorithm (Lütcke et al., 2013), a greedy template-fitting algorithm with variable
155 decay time constants across events and cells, with parameters chosen to result in the same
156 proportion of silent cells as has been shown previously with unbiased electrophysiology. As
157 an example of simpler methods, we use Yaksi and Friedrich (2006)’s simple deconvolution
158 of the raw calcium with a fixed kernel of the calcium response to a single spike. And finally
159 we create smoothed versions of the discrete-deconvolution methods, by convolving their
160 recovered spikes/events with a fixed spike-response kernel. Figure 3c show an example raw
161 calcium time-series for one neuron, and the result of applying each of these 8 processing
162 methods. We thus repeat all analyses on 9 different sets of time-series extracted from the
163 same population recording.

164 We choose the algorithm parameters as follows. Simple deconvolution (Yaksi and
165 Friedrich, 2006) involves taking a parameterised kernel of the GCaMP6s response to a
166 single spike [Same as Peron’s kernel? no, Peron used a bank of kernels. The kernel matches
167 Suite2P’s default GCaMP6s kernel, and the one used in LZero’s internal cost function].
168 For the three discrete deconvolution methods, we choose the modal values of the best-fit

parameters that optimised the Error Rate over the ground-truth recordings. This seems a reasonably consistent choice, of using the most consistently performing values obtained from comparable data: neurons in the same layer (L2/3) in the same species (mouse), in another primary sensory area (V1). Most importantly for our purposes, choosing the modal values means we avoid pathological regions of the parameter space.

2.4 Deconvolution methods disagree on estimates of simple neural statistics

We first check how well each approach recovers the basic statistics of neural activity event rates in L2/3 of barrel cortex. Electrophysiology has shown that the distribution of firing rates across neurons in a population is consistently long-tailed, and often log-normal, all across rodent cortex (Wohrer et al., 2013); and L2/3 neurons in barrel cortex are no different (O'Connor et al., 2010), with median firing rates less than 1 Hz, and a long right-hand tail of rarer high-firing neurons. We thus expect the calcium event rates or spike rates from our time-series would follow such a distribution. (Event rates for raw calcium, Peron, Yaksi and the continuous (kernel) versions of the data was obtained by thresholding the calcium time-series)

Figure 4a shows that the raw calcium and two of the discrete deconvolution methods (Suite2p, LZero) have qualitatively correct distributions of event rates (median near zero, long right-hand tails). The Peron time-series also have the correct distribution of event rates, which is unsurprising as it was tuned to do so. All other methods give qualitatively wrong distributions of spike rates (MLSpike) or event rates (all other methods). There is also little overlap in the distributions of spike rates between the three discrete deconvolution methods. Applying a kernel to their inferred spikes/events shifts rather than smooths the firing rate distributions ($\text{Suite2P}_{\text{kernel}}$, $\text{MLSpike}_{\text{kernel}}$, $\text{LZero}_{\text{kernel}}$), suggesting noise in the deconvolution process is amplified through the additional steps of convolving with a kernel and thresholding.

Cell-attached recordings in barrel cortex have shown that ~26% of L2/3 pyramidal cells are silent during a similar pole localisation task, with silence defined as emitting fewer than one spike every two minutes (O'Connor et al., 2010). For the nine approaches we test here, six estimated the proportion of silent cells to be less than 1%, including two of the discrete deconvolution methods (Figure 4c). For raw calcium and methods returning continuous time-series, raising the threshold for defining events will lead to more silent cells, but at the cost of further shifting the event rate distributions towards zero. Even for simple firing statistics of neural activity, the choice of time-series gives widely differing, and sometimes wrong, results.

2.5 Inferences of single cell tuning differ widely between raw calcium and deconvolved methods

We turn now to what we can infer about simple properties of neural coding, and how our choice of deconvolution method can alter those inferences. The decision task facing the mouse (Fig. 3a) requires that it moves its whisker back-and-forth to detect the position of the pole, delay for a second after the pole is withdrawn, and then make a choice of the left or right lick-port based on the pole's position (Fig. 3b). As the imaged barrel corresponds to the single spared whisker (on the contralateral side of the face), so the captured population activity during each trial likely contains neurons tuned to different aspects of the task. We show here that the number and identity of such task-tuned neurons in the population differ widely between deconvolution methods.

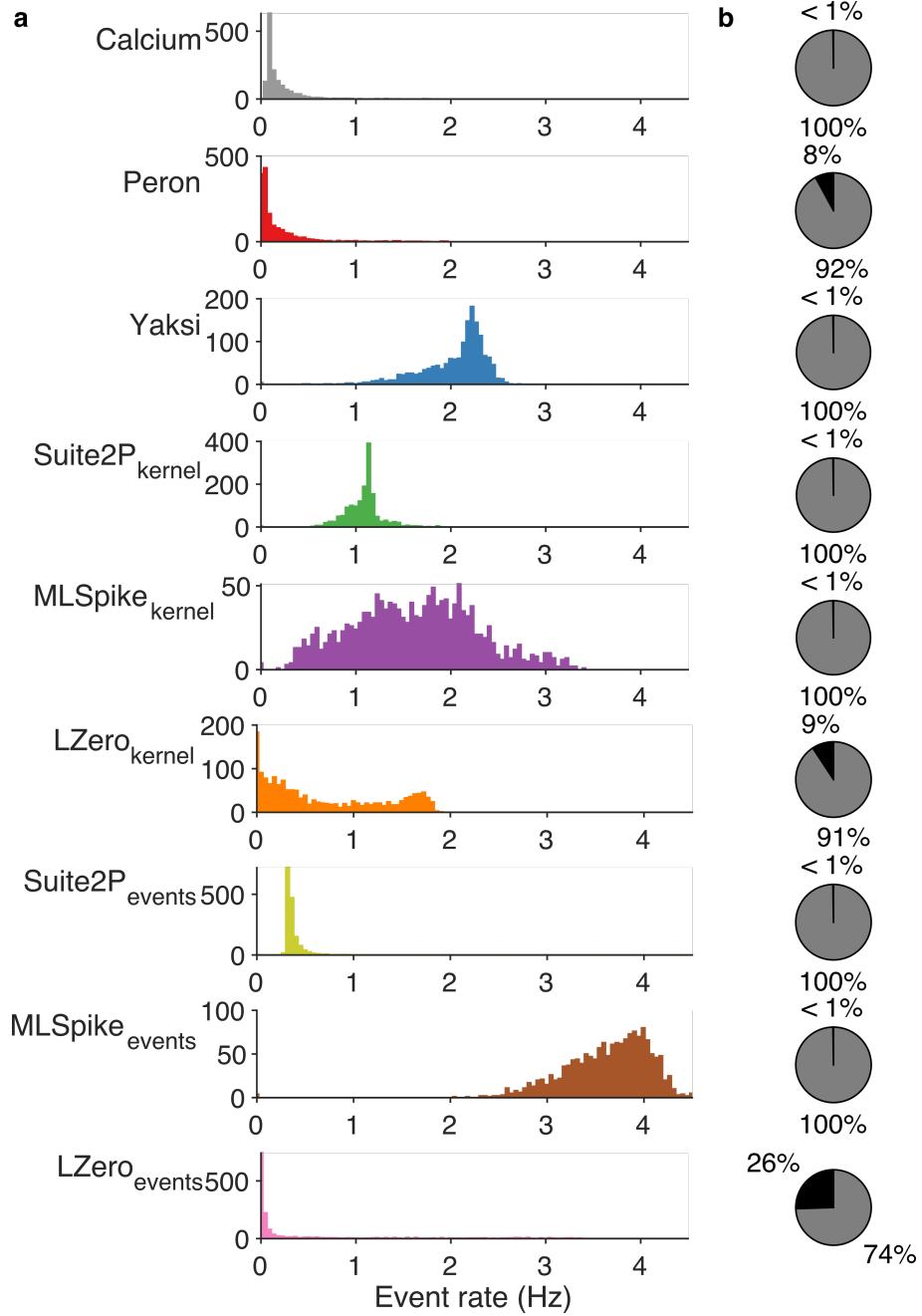


Figure 4: Estimates of population-wide event rates vary qualitatively across deconvolution methods.

(a) The distribution of event rate per neuron across the recorded population, according to each deconvolution method. For raw calcium and the five denoising methods (upper 6 panels), events are detected as fluorescence transients greater in magnitude than three standard deviations of background noise. The discrete deconvolution methods (lower 3 panels) return per frame: a spike count (MLSpike), a binary event detection (LZero), or an event magnitude (Suite2p); these time-series were thus sparse, with most frames empty.

(b) Proportion of active (gray) and silent (black) cells for each method. Silent cells are defined following (Peron et al., 2015b) as those with an event rate less than 0.0083Hz.

Following Peron et al. 2015a, we define a task-tuned cell as one for which the peak in its trial-averaged histogram of activity exceeds the predicted upper limit from shuffled data (Fig.5a; see Methods). When applied to the raw calcium time-series, close to half the neurons are tuned (Fig.5a). This is more than double the proportion found for the next nearest method (Yaksi's simple deconvolution), and at least a factor of 5 greater than the proportion of tuned neurons resulting from any discrete deconvolution method, which each report less than 10% of the neurons are tuned.

Worse, few neurons are detected as tuned in time-series resulting from multiple methods (Fig.5b). Only 104 neurons (6.7%) are labelled as tuned in at least two sets of time-series, and just 21 (1.35%) are labelled as tuned in all nine. Even separately considering the continuous and discrete time-series, we find only 38 cells are tuned across all six continuous methods, and 25 neurons for all three discrete deconvolution methods (Fig.5c). Figure 5d illustrates the diversity of detected tuning even amongst the neurons with the greatest agreement between methods.

These results suggest that raw calcium alone over-estimates tuning in the population, but also that there can be substantial disagreement between deconvolution methods. One solution for robust detection of tuned neurons is to find those agreed between the raw calcium time-series and more than one deconvolution method. In Figure 5e-h, we show how increasing the number of methods required to agree on a neuron's tuned status creates clear agreement between time-series processed with all methods, even if a particular method did not reach significance for that cell. Even requiring agreement between the raw calcium and just two other methods is enough to see tuning of many cells. The identification of unambiguously task-tuned cells could thus be achieved by triangulating the raw calcium with the output of multiple deconvolution methods.

In the pole detection task considered here, neurons tuned to pole contact are potentially crucial to understanding the sensory information used to make a decision. Touch onset is known to drive a subset of neurons to spike with short latency and low jitter (O'Connor et al., 2010; Hires et al., 2015). Detecting such rapid, precise responses in the slow kinetics of calcium imaging is challenging, suggesting discrete-deconvolution methods might be necessary to detect touch-tuned neurons. To test this, in each of the 9 sets of time-series we identify touch-tuned neurons by a significant peak in their touch-triggered activity (Fig 6a). Figure 6b shows that, while all data-sets have touch-tuned neurons, the number of such neurons differs substantially between them. And rather than being essential, discrete deconvolution methods disagree strongly on touch-tuning, with MLSpike (events) finding 45 touch-tuned neurons and LZero (events) finding one. Thus our inferences of the coding of task-wide or specific sensory events crucially depends on our choice of calcium imaging time-series.

2.6 Inconsistent recovery of population correlation structure across deconvolution approaches

The high yield of neurons from calcium imaging is ideal for studying the dynamics and coding of neural populations (Harvey et al., 2012; Huber et al., 2012; Kato et al., 2015). Many analyses of populations start from pairwise correlations between cells, whether as measures of a population's synchrony or joint activity, or as a basis for further analyses like clustering and dimension reduction (Cunningham and Yu, 2014). We now show how our inferences of population correlation structure also depend strongly on the choice of deconvolution method.

Figure 7a shows that the distributions of pairwise correlations qualitatively differ between the sets of time-series we derived from the same calcium imaging data. The con-

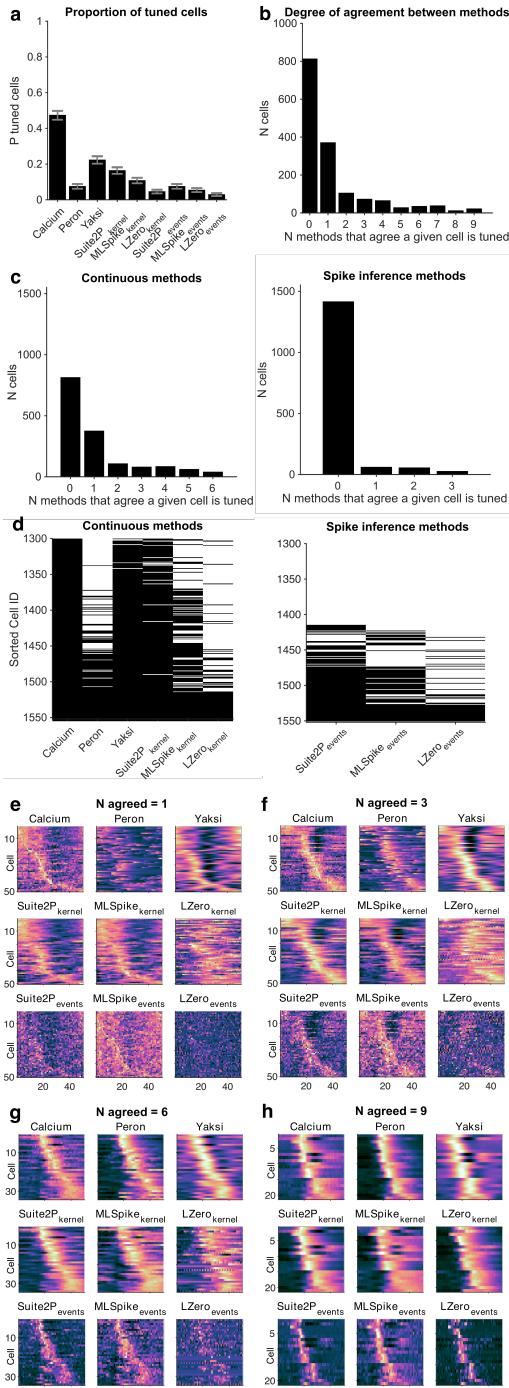


Figure 5: Inferences of single cell tuning show poor agreement between raw calcium and deconvolution methods, and between methods.

- (a0) Examples of a tuned (left) and non-tuned (right) cell from the raw calcium time-series. X: Data; Y: upper 95% interval from shuffled data.
- (a) Number of tuned cells per deconvolution method. Error bars are 95% binomial confidence intervals.
- (b) Agreement between methods. For each neuron, we count the number of methods (including raw calcium) for which it is labelled as tuned. Bars show the number of cells classified as tuned by exactly N methods.
- (c) Similar to (b), but breaking down the cells into: agreement between methods (raw or denoising) resulting in continuous signals (left panel); and agreement between discrete deconvolution methods (right panel).
- (d) Comparison of cell tuning across methods. Each row shows whether that cell is tuned (black) or not (white) under that deconvolution method. Cells are ordered from bottom to top by the number of methods that classify that cell as tuned.
- (e-h) Identifying robust cell tuning. Panel groups (e) to (h) show cells classed as tuned by increasing numbers of deconvolution methods. Each panel within a group plots one cell's normalised (z-scored) trial-average histogram per row, ordered by the time of peak activity. The first panel in a group of 9 shows histograms from raw calcium signals; each of the 8 subsequent panel shows trial-average

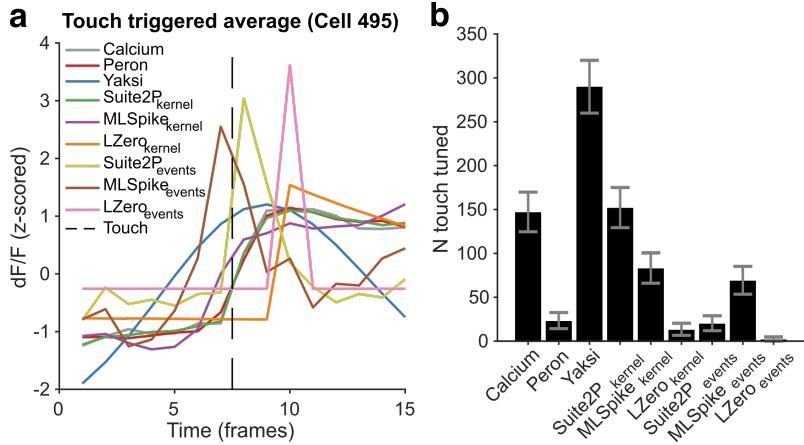


Figure 6: Touch-triggered neuron responses.

- (a) Touch-triggered average activity from one neuron, across all deconvolution methods. The dotted line is the imaging frame in which the whisker touched the pole.
- (b) Number of touch-tuned cells across deconvolution methods. A cell is classed as touch-tuned if its peak touch-triggered activity is significantly greater than shuffled data. Error bars are Jeffreys intervals

siderably narrower distributions from the discrete deconvolution time-series compared to the others is expected, as these time-series are sparse. Nonetheless, there are qualitative differences within the sets of discrete and continuous time-series. Some distributions are approximately symmetric, with broad tails; some asymmetric with narrow tails; the correlation distribution from the Peron method time-series is the only one with a median below zero. These qualitative differences are not due to noisy estimates of the pairwise correlations: for all our sets of time-series the correlations computed on a sub-set of time-points in the session agree well with the correlations computed on the whole session (Figure 7b). Thus pairwise correlation estimates for each method are stable, but their distributions differ between methods.

Looking in detail at the full correlation matrix shows that even for methods with similar distributions, their agreement on correlation structure is poor. Some neuron pairs that appear correlated from time-series processed by one deconvolution method are uncorrelated when processed with another method (Figure 7c). Over the whole population, the correlation structure obtained from the raw calcium, Yaksi and Suite2p (kernel) time-series all closely agree, but nothing else does (Figure 7d): the correlation structure obtained from LZero agrees with nothing else; and the discrete deconvolution methods all generate dissimilar correlation structures (Figure 7e). Our inferences about the extent and identity of correlations within the population will differ qualitatively depending on our choice of imaging time-series.

2.7 Deconvolution methods show the same population activity is both low and high dimensional

Dimensionality reduction techniques, like principal components analysis (PCA), allow researchers to make sense of large scale neuroscience data (Chapin and Nicolelis, 1999; Briggman et al., 2005; Churchland et al., 2012; Harvey et al., 2012; Cunningham and Yu, 2014; Kobak et al., 2016), by reducing the data from N neurons to $d < N$ dimensions. Key to such analyses is the choice of d , a choice guided by how much of the original data we can capture. To assess such inferences of population dimensionality, we apply PCA to

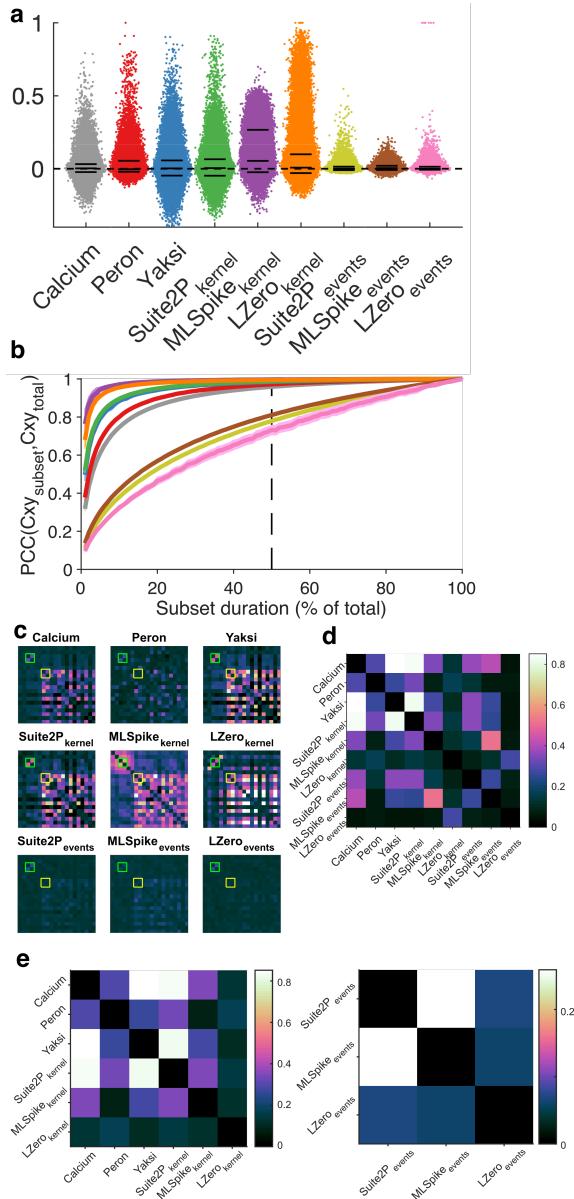


Figure 7: Effects of deconvolution on pairwise correlations between neurons.

- (a) Distributions of pairwise correlations between all cells, for each deconvolution method (one dot per cell pair, x-axis jitter added for clarity). Solid black lines are 5th, 50th and 95th percentiles.
- (b) Stability of correlation structure in the population. We quantify here the stability of the pairwise correlation estimates, by comparing the correlation matrix constructed on the full data ($C_{xy\text{total}}$) to the same matrix constructed on a subset of the data ($C_{xy\text{subset}}$). Each data-point is the mean correlation between $C_{xy\text{total}}$ and $C_{xy\text{subset}}$; one line per deconvolution method. Shaded error bars are one standard deviation of the mean across 100 random subsets.
- (c) Examples of qualitatively differing correlation structure across methods. Each panel plots the pairwise correlations for the same 50 neurons on the same colour scale. As examples, we highlight two pairs of cells: one consistently correlated across different methods (green arrow and boxes); the other not (yellow arrow and boxes).
- (d) Comparison of pairwise correlation matrices between deconvolution methods. Each square is the Spearman's rank correlation between the full-data correlation matrix for that pair of methods.
- (e) as in (d), but split to show continuous methods (left) or discrete deconvolution methods (right).

291 our 9 sets of imaging time-series to estimate the dimensionality of the imaging data (which
 292 for PCA is the variance explained by each eigenvector of the data's covariance matrix).

293 Figure 8a plots for each deconvolution method the cumulative variance explained when
 294 increasing the number of retained dimensions. Most deconvolution methods qualitatively
 295 disagree with the raw calcium data-set on the relationship between dimensions and vari-
 296 ance. This relationship is also inconsistent across deconvolution methods; indeed the
 297 discrete deconvolution methods result in the shallowest ($\text{MLSpike}_{\text{events}}$) and amongst the
 298 steepest ($\text{LZero}_{\text{events}}$) relationships between increasing dimensions and variance explained.
 299 The number of dimensions required to explain 80% of the variance in the data ranges
 300 from $d = 125$ (Peron) to $d = 1081$ ($\text{MLSpike}_{\text{events}}$), a jump from 8% to 70% of all pos-
 301 sible dimensions (Fig 8b). Thus we could equally infer that the same L2/3 population
 302 activity is low dimensional (<10% dimensions required to explain 80% of the variance)
 303 or high-dimensional (>50% of dimensions required) depending on our choice of imaging
 304 time-series.

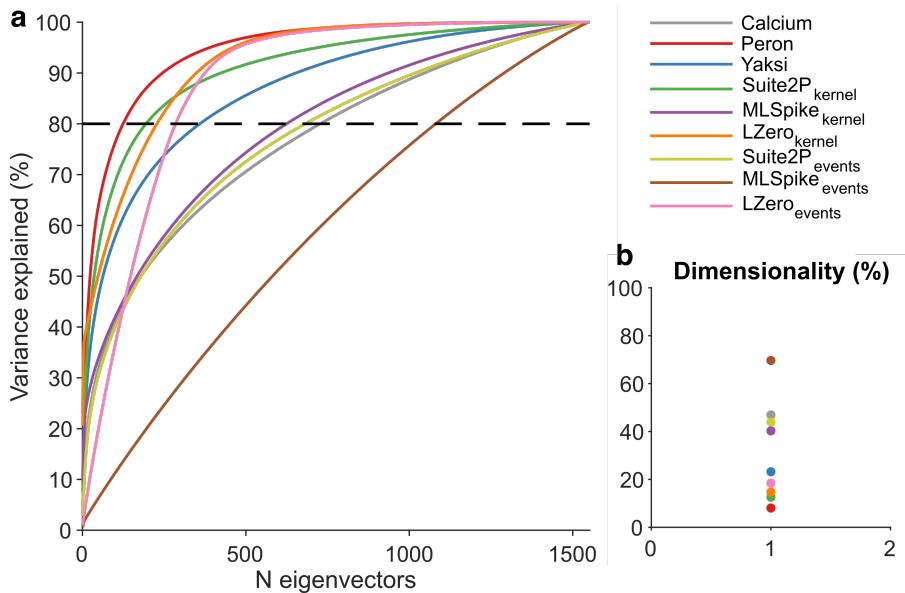


Figure 8: Dimensionality of population activity.

(a) Cumulative variance explained by each dimension of the data's covariance matrix, one line per deconvolution method. Dimensions are obtained from principal components analysis, and are ordered by decreasing contribution to the total variance explained. Dashed line is the 80% threshold used in panel (b).

(b) Proportion of dimensions required to explain 80% of the variance in the data.

305 **3 Discussion**

306 Imaging of somatic calcium is a remarkable tool for capturing the simultaneous activity of
307 hundreds to thousands of neurons. But the time-series of each neuron’s calcium fluores-
308 cence is inherently noisy and non-linearly related to its spiking. We sought here to address
309 how our choice of corrections to these time-series – to use them raw, deconvolve them into
310 continuous time-series, or deconvolve them into discrete events – affect the quality and
311 reliability of the scientific inferences drawn.

312 Our results show the choice qualitatively changes the potential scientific inferences we
313 draw about neural activity, coding, and correlation structure. We consistently observe
314 that the analysis results differ sharply between the raw calcium and most, if not all, of the
315 processed time-series. However, the deconvolved time-series also consistently disagreed
316 with each other, even between methods of the same broad class (continuous or discrete
317 time-series).

318 **3.1 Accurate discrete deconvolution is possible, but sensitive**

319 We find much that is encouraging. In fitting discrete deconvolution methods to ground-
320 truth data, we found they can in principle accurately recover neural activity. A caveat
321 here is that the choice of metric for evaluation and fitting of parameters is of critical
322 importance. The widely-used Pearson correlation coefficient is a poor choice of metric
323 as it returns inconsistent results with small changes in algorithm parameters, and leads
324 to poor estimates of simple measures such as firing rate when used across methods and
325 sampling rates. By contrast, the Error Rate metric (Deneux et al., 2016; Victor and
326 Purpura, 1996) resulted in excellent recovery of ground-truth spike trains. Other recently
327 developed methods for comparing spike-trains based on information theory (Theis et al.,
328 2016) or fuzzy set theory (Reynolds et al., 2018), may also be appropriate.

329 However, while good estimates of neural activity can be achieved with modern discrete
330 deconvolution methods (Berens et al., 2018; Pachitariu et al., 2018), the best parameters
331 vary substantially between cells, and small changes in analysis parameters result in poor
332 performance. This variation and sensitivity of parameters played out as widely-differing
333 results between the three discrete deconvolution methods in analyses of neural activity,
334 coding, and correlation structure.

335 **3.2 Choosing parameters for deconvolution methods**

336 A potential limitation of our study is that we use a single set of parameter values for
337 each discrete deconvolution method applied to the population imaging data from barrel
338 cortex. But then our situation is the same as that facing any experimentalist: in the
339 absence of ground-truth, how do we set the parameters? Our solution here was to use
340 the modal parameter values from ground-truth fitting, as these values are candidates for
341 the most general solutions. We also felt these were a reasonable choice for the population
342 imaging data from barrel cortex, given that the ground-truth recordings came from the
343 same species (mouse) in the same layer (2/3) of a different bit of primary sensory cortex
344 (V1).

345 Rather than use the most general parameters values, another solution would be to
346 tune the parameters to obtain known gross statistics of the neural activity. This was
347 the approach used by Peron and colleagues (Peron et al., 2015b) to obtain the denoised
348 “Peron” time-series we included here. But as we’ve seen, this approach can lead to its own
349 problems: for example, in the Peron time-series, it created a distributions of correlations

350 that differed from any other set of time-series. Indeed, finding good parameter values may
351 be an intractable problem, as it is possible each neuron requires individual fitting, to reflect
352 the combination of its expression of fluorescent protein, and its particular non-linearity
353 between voltage and calcium.

354 **3.3 Ways forward**

355 The simplest solution to the inconsistencies between different forms of time-series is to
356 triangulate them, and take the consensus across their results. For example, our finding of
357 a set of tuned neurons across multiple methods is strong evidence that neurons in L2/3
358 of barrel cortex are responsive across the stages of the decision task. Further examples
359 of such triangulation in the literature are rare; Klaus and colleagues (Klaus et al., 2017)
360 used two different pipelines to derive raw $\Delta F/F$ of individual neurons from one-photon
361 fibre-optic recordings in the striatum, and replicated all analyses using the output of both
362 pipelines. Our results encourage the further use of triangulation to create robust inference:
363 obtaining the same result in the face of wide variation increases our belief in its reliability
364 (Munafò and Davey Smith, 2018).

365 There are caveats to using triangulation. For single neuron analyses, triangulation
366 inevitably comes at the price of reducing the yield of neurons to which we can confidently
367 assign roles. A further problem for triangulation is how to combine more complex analyses,
368 such as pairwise correlations; the alternative is to rely on qualitative comparisons.

369 Many studies use the raw calcium signal as the basis for all their analyses (Harvey et al.,
370 2012; Huber et al., 2012) [cite others], perhaps assuming this is the least biased approach.
371 Our result show this is not so: the discrepancy between raw and deconvolved calcium on
372 single neuron coding suggests an extraordinary range of possible results, from about half
373 of all neurons tuned to the task down to less 5 percent. The qualitative conclusion – there
374 is coding – is not satisfactory. Thus our results should not be interpreted as a call to
375 abandon deconvolution methods; rather they serve to delimit how we can interpret their
376 outputs.

377 Instead, we need deconvolution solved: as sensors with faster kinetics (though funda-
378 mentally limited by kinetics of calcium release itself) and higher signal-to-noise ratios are
379 developed (Badura et al., 2014; Dana et al., 2016, 2019), so the accuracy and robustness of
380 de-noising and deconvolution should improve; and as the neuron yield continues to increase
381 (Stringer et al., 2019; Ahrens et al., 2013), so the potential for insights from inferred spikes
382 or spike-driven events grows. Developing further advanced deconvolution algorithms will
383 harness these advances, but are potentially always limited by the lack of ground-truth to
384 fit their parameters. Our results may provide impetus for a different direction of research,
385 focussing on how we can get consensus among the output of different algorithms, and thus
386 provide robust scientific inferences about neural populations.

387 **4 Methods**

388 **Ground truth data**

389 Ground truth data was accessed from crcns.org (Svoboda, 2015), and the experiments have
390 been described previously (Chen et al., 2013). Briefly, mouse visual cortical neurons ex-
391 pressing the fluorescent calcium reporter protein GCaMP6s were imaged with two-photon
392 microscopy at 60Hz. Loose-seal cell-attached recordings were performed simultaneously
393 at 10kHz. The data-set contains twenty one recordings from nine cells.

394 **Population imaging data description**

395 Population imaging data was accessed from crcns.org and have been described previ-
396 ously (Peron et al., 2015b). Briefly, volumetric two photon calcium imaging of primary
397 somatosensory cortex (S1) was performed in awake head-fixed mice performing a whisker-
398 based object localisation task. In the task a metal pole was presented on one of two loca-
399 tions and mice were motivated with fluid reward to lick at one of two lick ports depending
400 on the location of the pole following a brief delay. Two photon imaging of GCaMP6s
401 expressing neurons in superficial S1 was performed at 7Hz. Images were motion corrected
402 and aligned, before regions of interest were manually set and neuropil-subtracted. A single
403 recording from this dataset was used for population analysis. The example session had
404 1552 neurons recorded for a total of 23559 frames (56 minutes).

405 **List of deconvolution methods**

406 **MLSpike**

407 MLSpike (Deneux et al., 2016) was accessed from <https://github.com/mlspike>. MLSpike
408 uses a model-based probabilistic approach to recover spike trains in calcium imaging data
409 by taking baseline fluctuations and cellular properties into account. Briefly, MLSpike
410 implements a model of measured calcium fluorescence as a combination of spike-induced
411 transients, background (photonic) noise and drifting baseline fluctuations. A maximum
412 likelihood approach determines the probability of the observed calcium at each time step
413 given an inferred spike train generated through a particular set of model parameters.
414 MLSpike returns a maximum a posteriori spike train (as used here), or a spike probability
415 per time step. MLSpike also returns an estimate of the drifting background fluorescence
416 which is ignored in this work.

417 MLSpike has a number of free parameters, of which we optimise three: A , the mag-
418 nitude of fluorescence transients caused by a single spike; τ , calcium fluorescence decay
419 time; σ , background (photonic) noise level. MLSpike also has parameters for different
420 calcium sensor kinetics (for OBG, GCaMP3, GCaMP6 and so on) which we fix to default
421 values for GCaMP6.

422 For our analysis of event rate MLSpike's spike train was counted (mean event count
423 per second), and for subsequent analyses was converted to a dense array of spike counts
424 per imaging frame.

425 **Suite2P**

426 Suite2P (Pachitariu et al., 2016, 2018) was accessed from [https://github.com/cortex-lab/Suite2P](https://github.com/cortex-
427 lab/Suite2P). Suite2P was developed as a complete end-to-end processing pipeline for
428 large scale 2-photon imaging analysis - from image registration to spike extraction and

429 visualization - of which we only use the spike extraction step. The spike deconvolution
430 of Suite2P uses a sparse non-negative deconvolution algorithm, greedily identifying and
431 removing calcium transients to minimise the cost function

$$C = \|F - s * k\|^2,$$

432 where the cost C is the squared norm of fluorescence F minus a reconstruction of that
433 signal comprising a sparse array of spiking events s multiplied by a parameterised calcium
434 kernel k . The kernel was parameterised following defaults for GCaMP6s (exponential
435 decay of 2 seconds, though it has been shown the precise value of this parameter does not
436 affect performance for this method (Pachitariu et al., 2018)).

437 Suite2P has a further free parameter which sets the minimum spike event size, the
438 *Threshold*, which determines the stopping criteria for the algorithm.

439 Elements of s are of varying amplitude corresponding to the amplitude of the calcium
440 transients at that time. For ground truth firing rate analysis we are interested in each
441 algorithm's ability to recover spike trains, therefore we treat each event as a 'spike' and
442 optimise the algorithm appropriately. For our analysis of event rate Suite2P's event train
443 was counted (mean event count per second), and for subsequent analyses was converted
444 to a dense array of varying amplitude events (i.e. s) per imaging frame.

445 **LZero**

446 The method we refer to as LZero was written in Matlab based on an implementation
447 in *R* accessed at <https://github.com/jewellsean/LZeroSpikeInference>. A full description is
448 available in the paper of Jewell and Witten (2018). Briefly, in LZero spike detection is cast
449 as a change-point detection problem, which could be solved with an l_0 optimization algo-
450 rithm. Working backwards from the last time point the algorithm finds time points where
451 the calcium dynamics abruptly change from a smooth exponential rise. These change
452 points correspond to spike event times. Spike inference accuracy is assessed similarly to
453 Suite2P by measuring the fit between observed fluorescence and a reconstruction based
454 on inferred spike times and a fixed calcium kernel.

455 LZero has two free parameters - *lambda*, a tuning parameter that controls the trade-off
456 between the sparsity of the estimated spike event train and the fit of the estimated calcium
457 to the observed fluorescence; and *scale*, the magnitude of a single spike induced change in
458 fluorescence.

459 For our analysis of event rate LZero's spike train was counted (mean event count per
460 second), and for subsequent analyses was converted to a dense array of spikes per imaging
461 frame (maximum one spike per imaging frame due to limitations of the algorithm).

462 **Yaksi**

463 Yaksi is an implementation of the deconvolution approach of Yaksi and Friedrich (2006).
464 The fluorescence time series is low-pass filtered (4th order butterworth filter, 0.7Hz cutoff)
465 to remove noise before having a calcium kernel (exponential decay of 2 seconds, as used
466 in Suite2P and LZero above) linearly deconvolved out of the signal using Matlab's `deconv`
467 function. The output of Yaksi is a continuous signal approximating spike density per unit
468 time.

469 **Peron events**

470 Peron events refer to the de-noised calcium event traces detailed in the original Peron
471 et al. (2015b) paper. Here a version of the ‘peeling’ algorithm (Lütcke et al., 2013) was
472 developed, a template-fitting algorithm with variable decay time constants across events
473 and cells. This algorithm was tuned by the authors to generate a low number of false
474 positive detections (a rate of 0.01Hz) on ground truth data, matching firing rate statistics
475 from cell-attached electrophysiology and leading to a hit rate of 54%. The output for
476 analysis is a continuous signal approximating de-noised calcium concentration per unit
477 time.

478 *Events and kernel versions of spike inference methods*

479 Where a spike inference method returns spike counts per time point, these are plotted
480 as Method_{events}. To compare to other methods that return a de-noised dF/F or firing
481 rate estimates, these event traces are convolved with a calcium kernel and plotted as
482 Method_{kernel}. The kernel used is consistent with that used as a default for GCaMP6s
483 in ML Spike, Suite2P and LZero, namely an exponential decay of two seconds duration
484 normalised to have an integral of 1.

485 **Ground truth spike train metrics**

486 Pearson correlation coefficient was computed between the ground truth and inferred spikes
487 (ML Spike) or events (Suite2P, LZero) following convolution of both with a gaussian kernel
488 (61 samples wide, 1.02 seconds).

489 Error Rate was computed between the ground truth and inferred spikes/events using
490 the Deneux et al. (2016) implementation of normalised error rate, derived from Victor
491 and Purpura (1996) Error Rate (code available <https://github.com/MLspike>). Briefly,
492 the error rate is 1 - F1-score, where the F1-score is the harmonic mean of sensitivity and
493 precision (Davis and Goadrich, 2006),

$$\begin{aligned} \text{sensitivity} &= 1 - \frac{\text{misses}}{\text{total spikes}}, \\ \text{precision} &= 1 - \frac{\text{false detections}}{\text{total detections}}, \\ \text{ErrorRate} &= 1 - 2 \frac{\text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}}. \end{aligned}$$

494 Hits, misses and false detections were counted with a temporal precision of 0.5 seconds.
495 For normalised estimation of errors in firing/event rate we compute,

$$\frac{\text{estimated rate} - \text{true rate}}{\text{true rate}},$$

496 where spike/event rates are measured in Hz.

497 **Parameter fitting**

498 For each method the best parameters for each cell were determined by brute force search
499 over an appropriate range (i.e. at least two orders of magnitude encompassing full param-
500 eter ranges used in the original publications for each method). The parameter ranges were
501 explored on a log scale as follows: ML Spike A (0.01:1, 21 values), tau (0.01:5, 21 values),

502 sigma (0.01:1, 21 values); Suite2P Threshold (0.1:100, 13 values); LZero lambda (0.1:20,
503 23 values), scale (0.1:20, 23 values).

504 The modal best parameters, as determined using Error Rate on downsampled data,
505 were then fixed for the population imaging data analysis. These were: MLSpike A: 0.1995,
506 tau: 1.9686, sigma: 0.0398; Suite2P Threshold: 1.7783; LZero sigma: 0.1; lambda: 3.1623.

507 **Downsampling**

508 Ground truth calcium data was downsampled from 60Hz to 7Hz in Matlab by up-sampling
509 by `7 - interp(ca,7)` and downsampling the resultant 420Hz signal by 60 as Matlab's
510 downsampling must be done in integer steps. [??? how was the downsampling done then?
511 By taking every Nth frame from the interpolated data? Or averaging over every approx
512 8.5 frames of the 60Hz signal?]. Yes by downsampling the interpolated data. Is this not
513 clear from what I've re-written?. The 60Hz signal is upsampled by 7 to 420 Hz, which is
514 then down sampled by a factor of 60 to 7Hz.

515 **4.1 Event rate estimation**

516 Spike inference methods (`Suite2Pevents`, `MLSpikeevents`, `LZeroevents`) return estimated spike
517 times (`MLSpike`), or event times (`Suite2P/LZero`) which were converted into mean event
518 rates (Hz) per cell.

519 The event rate for continuous methods (Calcium, Peron, Yaksi, `Suite2Pkernel`, `MLSpikekernel`,
520 `LZerokernel`) for each cell was determined by counting activity/fluorescence transients
521 greater than three standard deviations of the background noise. Background noise was
522 calculated by subtracting a smoothed four-point moving average of the fluorescence from
523 the raw data to result in a 'noise only' trace. This operation was done separately for each
524 cell and each method. Event rate was then computed in Hz.

525 [what does this all mean? That each data-point for the background noise was an
526 average over 4 adjacent frames, and the standard deviation of the noise was computed
527 from those data-points? Why smooth the noise? And what segment of the data was
528 treated as "background"? (i.e. how many data-points)? Or does this mean that the whole
529 Ca2+ trace for each cell was smoothed using a 4-frame average (shifting 1 frame?), and
530 the SD of that smoothed signal was used as an estimate of background noise?] No. I
531 have re-written for clarity. I'm trying to isolate the high-frequency noise by subtracting
532 the low-frequency component first. So the df/f is smoothed, and this smoothed signal
533 is subtracted from the original data to leave a noise-only trace. This noise-only trace is
534 what the detection threshold is computed from. Does this logic make sense, and does my
535 explanation now explain what I did?

536 Silent cells were defined as cells with event rates below 0.0083Hz (or fewer than one
537 spike per two minutes of recording) as in O'Connor et al. (2010).

538 **4.2 Task-tuned cells**

539 Task-tuning was determined for each neuron using the model-free approach of Peron et al.
540 (2015b). Neurons were classed as task-tuned if their peak trial-average activity exceeded
541 the 95th percentile of a distribution of trial-average peaks from shuffled data (10000 shuf-
542 fles). The shuffle test was done separately for correct lick-left and lick-right trials and cells
543 satisfying the tuning criteria in either case were counted as task-tuned.

544 Tuned cell agreement was calculated as the number of methods that agreed to the
545 tuning status of a given cell, for all methods and separately for continuous and spike

546 inference methods.

547 4.3 Touch-related responses

548 Touch-tuned cells were determined by computing touch-triggered average activity for each
549 cell, before calculating whether the data distribution of peak touch-induced activity ex-
550 ceeds the expected activity of shuffled data. In more detail, the time of first touch -
551 between the mouse's whisker and the metal pole - on each trial was recorded. For each
552 touch time, one second of activity (seven data samples) was extracted before and after
553 the frame closest to touch (15 samples total); taking the mean of these gave the average
554 touch response for the cell. To determine whether a cell was touch tuned or not the time
555 of peak touch-triggered average activity was calculated, and a Wilcoxon rank sum test
556 (Benjamini Hochberg corrected, alpha 0.05) between the true data distribution at peak
557 time and a matched random sample of data from the same cell was performed.

558 [unclear what the pairwise tests were between; the mean activity at the peak time, and
559 a N-length vector of mean activity as the same time obtained from N shuffled datasets?]

560 The test was between the distribution around the peak (e.g. 100 values from 100
561 touches) and a matched random sample of the data (e.g. 100 other points taken at random
562 from that cell's time series).

563 [Shuffled how many times? Shuffled how?].

564 Not really a shuffle. it's a matched (N data points for N touches, 133 in this case)
565 random sample. I think this is equivalent to a shuffle, no?

566 4.4 Pairwise correlations

567 Pairwise correlations (Pearson correlation coefficients, Fig. 7a) were calculated for all pairs
568 of neurons in Matlab (corrcoef) at the data sampling rate (7Hz).

569 [State what forms of time-series were correlated - from debugging doc I've put the
570 format of each method in the method description section, and put this section before the
571 analysis sections. Does this address this point appropriately?]

572 Stability of correlation estimates (Fig. 7b) at the recording durations used was assessed
573 by computed the similarity between correlation distributions for the intact dataset to
574 those from subsets of the dataset. For each deconvolution method, we computed the
575 pairwise correlation matrix using the entire session's data, as above. We also sampled a
576 subset of time-points (1%-100%) of the full dataset at random without replacement and
577 computed a matrix of pairwise correlations for this subset. We then compute the similarity
578 between the total and subset matrices using Pearson's correlation coefficient. This process
579 was repeated 100 times and the mean (line) and standard deviation (shading) of the 100
580 repeats were plotted.

581 4.5 Correlations between correlation matrices

582 Correlations between correlation matrices (Fig. 7c-e) were computed using Spearman's
583 rank correlation between the unique pairwise correlations from each method (i.e. the
584 upper triangular entries of the correlation matrix).

585 4.6 Dimensionality

586 To determine the dimensionality of each dataset we performed eigendecomposition of the
587 covariance matrix of each dataset. [Computed as per the pairwise correlations above?
588 not sure what you mean? In terms of data format yes...] The resultant eigenvalues

589 were sorted into descending order, and the cumulative variance explained plotted, and the
590 number of eigenvectors required to explain 80% of the variance recorded.

591 References

- 592 Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller.
593 Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat.*
594 *Methods*, 10(5):413–420, May 2013.
- 595 Aleksandra Badura, Xiaonan Richard Sun, Andrea Giovannucci, Laura A Lynch, and
596 Samuel S-H Wang. Fast calcium sensor proteins for monitoring neural activity. *Neuro-*
597 *photonics*, 1(2):025008, October 2014.
- 598 Philipp Berens, Jeremy Freeman, Thomas Deneux, Nicolay Chenkov, Thomas McColgan,
599 Artur Speiser, Jakob H Macke, Srinivas C Turaga, Patrick Mineault, Peter Rupprecht,
600 Stephan Gerhard, Rainer W Friedrich, Johannes Friedrich, Liam Paninski, Marius Pa-
601 chitariu, Kenneth D Harris, Ben Bolte, Timothy A Machado, Dario Ringach, Jasmine
602 Stone, Luke E Rogerson, Nicolas J Sofroniew, Jacob Reimer, Emmanouil Froudarakis,
603 Thomas Euler, Miroslav Román Rosón, Lucas Theis, Andreas S Tolias, and Matthias
604 Bethge. Community-based benchmarking improves spike rate inference from two-photon
605 calcium imaging data. *PLoS Comput. Biol.*, 14(5):e1006157, May 2018.
- 606 K L Briggman, H D I Abarbanel, and W B Kristan, Jr. Optical imaging of neuronal
607 populations during decision-making. *Science*, 307(5711):896–901, February 2005.
- 608 Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data
609 analysis: state-of-the-art and future challenges. *Nat. Neurosci.*, 7(5):456–461, May 2004.
- 610 J K Chapin and M A Nicolelis. Principal component analysis of neuronal ensemble activity
611 reveals multidimensional somatosensory representations. *J. Neurosci. Methods*, 94(1):
612 121–140, December 1999.
- 613 Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy
614 Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, Loren L
615 Looger, Karel Svoboda, and Douglas S Kim. Ultrasensitive fluorescent proteins for
616 imaging neuronal activity. *Nature*, 499(7458):295–300, July 2013.
- 617 Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul
618 Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during
619 reaching. *Nature*, 487(7405):51–56, July 2012.
- 620 John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural
621 recordings. *Nat. Neurosci.*, 17(11):1500–1509, November 2014.
- 622 Hod Dana, Boaz Mohar, Yi Sun, Sujatha Narayan, Andrew Gordus, Jeremy P Hasseman,
623 Getahun Tsegaye, Graham T Holt, Amy Hu, Deepika Walpita, Ronak Patel, John J
624 Macklin, Cornelia I Bargmann, Misha B Ahrens, Eric R Schreiter, Vivek Jayaraman,
625 Loren L Looger, Karel Svoboda, and Douglas S Kim. Sensitive red protein calcium
626 indicators for imaging neural activity. *Elife*, 5, March 2016.
- 627 Hod Dana, Yi Sun, Boaz Mohar, Brad K Hulse, Aaron M Kerlin, Jeremy P Hasseman,
628 Getahun Tsegaye, Arthur Tsang, Allan Wong, Ronak Patel, John J Macklin, Yang
629 Chen, Arthur Konnerth, Vivek Jayaraman, Loren L Looger, Eric R Schreiter, Karel

- 630 Svoboda, and Douglas S Kim. High-performance calcium sensors for imaging activity
631 in neuronal populations and microcompartments. *Nat. Methods*, 16(7):649–657, July
632 2019.
- 633 Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC
634 curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML
635 '06, pages 233–240, New York, NY, USA, 2006. ACM.
- 636 Thomas Deneux, Attila Kaszas, Gergely Szalay, Gergely Katona, Tamás Lakner, Amiram
637 Grinvald, Balázs Rózsa, and Ivo Vanzetta. Accurate spike estimation from noisy calcium
638 signals for ultrafast three-dimensional imaging of large neuronal populations *in vivo*.
639 *Nat. Commun.*, 7:12190, July 2016.
- 640 Johannes Friedrich, Pengcheng Zhou, and Liam Paninski. Fast online deconvolution of
641 calcium imaging data. *PLoS Comput. Biol.*, 13(3):e1005423, March 2017.
- 642 Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown,
643 Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou,
644 Baljit S Khakh, David W Tank, Dmitri B Chklovskii, and Eftychios A Pnevmatikakis.
645 CaImAn an open source tool for scalable calcium imaging data analysis. *Elife*, 8, January
646 2019.
- 647 Kenneth D Harris, Rodrigo Quian Quiroga, Jeremy Freeman, and Spencer L Smith. Im-
648 proving data quality in neuronal population recordings. *Nat. Neurosci.*, 19(9):1165–
649 1174, August 2016.
- 650 Christopher D Harvey, Philip Coen, and David W Tank. Choice-specific sequences in
651 parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–68, April
652 2012.
- 653 Samuel Andrew Hires, Diego A Gutnisky, Jianing Yu, Daniel H O'Connor, and Karel
654 Svoboda. Low-noise encoding of active touch by layer 4 in the somatosensory cortex.
655 *Elife*, 4, August 2015.
- 656 Daniel Huber, D A Gutnisky, S Peron, D H O'Connor, J S Wiegert, L Tian, T G Oertner,
657 L L Looger, and K Svoboda. Multiple dynamic representations in the motor cortex
658 during sensorimotor learning. *Nature*, 484(7395):473–478, April 2012.
- 659 Sean Jewell and Daniela Witten. Exact spike train inference via ℓ_0 optimization. *Ann.*
660 *Appl. Stat.*, 12(4):2457–2482, December 2018.
- 661 Patrick Kaifosh, Jeffrey D Zaremba, Nathan B Danielson, and Attila Losonczy. SIMA:
662 Python software for analysis of dynamic fluorescence imaging data. *Front. Neuroinform.*,
663 8:80, September 2014.
- 664 Saul Kato, Harris S Kaplan, Tina Schrödel, Susanne Skora, Theodore H Lindsay, Eviatar
665 Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dynamics embed the motor
666 command sequence of *caenorhabditis elegans*. *Cell*, 163(3):656–669, October 2015.
- 667 Sander W Keemink, Scott C Lowe, Janelle M P Pakan, Evelyn Dylda, Mark C W van
668 Rossum, and Nathalie L Rochefort. FISSA: A neuropil decontamination toolbox for
669 calcium imaging signals. *Sci. Rep.*, 8(1):3493, February 2018.

- 670 Andreas Klaus, Gabriela J Martins, Vitor B Paixao, Pengcheng Zhou, Liam Paninski, and
671 Rui M Costa. The spatiotemporal organization of the striatum encodes action space.
672 *Neuron*, 95(5):1171–1180.e7, August 2017.
- 673 Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam
674 Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Chris-
675 tian K Machens. Demixed principal component analysis of neural population data. *Elife*,
676 5, April 2016.
- 677 Henry Lütcke, Felipe Gerhard, Friedemann Zenke, Wulfram Gerstner, and Fritjof Helm-
678 chen. Inference of neuronal network spike dynamics and topology from calcium imaging
679 data. *Front. Neural Circuits*, 7:201, December 2013.
- 680 Eran A Mukamel, Axel Nimmerjahn, and Mark J Schnitzer. Automated analysis of cellular
681 signals from large-scale calcium imaging data. *Neuron*, 63(6):747–760, September 2009.
- 682 Marcus R Munafò and George Davey Smith. Robust research needs many lines of evidence.
683 *Nature*, 553(7689):399–401, January 2018.
- 684 Daniel H O'Connor, Simon P Peron, Daniel Huber, and Karel Svoboda. Neural activity
685 in barrel cortex underlying vibrissa-based object localization in mice. *Neuron*, 67(6):
686 1048–1061, September 2010.
- 687 Marius Pachitariu, Carsen Stringer, Sylvia Schröder, Mario Dipoppa, L Federico Rossi,
688 Matteo Carandini, and Kenneth D Harris. Suite2p: beyond 10, 000 neurons with stan-
689 dard two-photon microscopy. *BioRxiv*, Preprint at <http://dx.doi.org/10.1101/061507>,
690 2016.
- 691 Marius Pachitariu, Carsen Stringer, and Kenneth D Harris. Robustness of spike deconvolu-
692 tion for neuronal calcium imaging. *J. Neurosci.*, August 2018.
- 693 António R C Paiva, Il Park, and José C Príncipe. A comparison of binless spike train
694 measures. *Neural Comput. Appl.*, 19(3):405–419, April 2010.
- 695 Simon Peron, Tsai-Wen Chen, and Karel Svoboda. Comprehensive imaging of cortical
696 networks. *Curr. Opin. Neurobiol.*, 32:115–123, June 2015a.
- 697 Simon P Peron, Jeremy Freeman, Vijay Iyer, Caiying Guo, and Karel Svoboda. A cellular
698 resolution map of barrel cortex activity during tactile behavior. *Neuron*, 86(3):783–799,
699 May 2015b.
- 700 Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh
701 Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha
702 Ahrens, Randy Bruno, Thomas M Jessell, Darcy S Peterka, Rafael Yuste, and Liam
703 Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging
704 data. *Neuron*, January 2016.
- 705 Ruben Portugues, Claudia E Feierstein, Florian Engert, and Michael B Orger. Whole-
706 brain activity maps reveal stereotyped, distributed networks for visuomotor behavior.
707 *Neuron*, 81(6):1328–1343, March 2014.
- 708 Stephanie Reynolds, Therese Abrahamsson, Per Jesper Sjöström, Simon R Schultz, and
709 Pier Luigi Dragotti. CosMIC: A consistent metric for spike inference from calcium
710 imaging. *Neural Comput.*, 30(10):2726–2756, October 2018.

- 711 Carsen Stringer and Marius Pachitariu. Computational processing of neural recordings
712 from calcium imaging data. *Curr. Opin. Neurobiol.*, 55:22–31, December 2018.
- 713 Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Ken-
714 neth D Harris. High-dimensional geometry of population responses in visual cortex.
715 *Nature*, June 2019.
- 716 K Svoboda. Simultaneous imaging and loose-seal cell-attached electrical recordings from
717 neurons expressing a variety of genetically encoded calcium indicators. *GENIE project,*
718 *Janelia Farm Campus, HHMI; CRCNS.org*, 2015.
- 719 Lucas Theis, Philipp Berens, Emmanouil Froudarakis, Jacob Reimer, Miroslav
720 Román Rosón, Tom Baden, Thomas Euler, Andreas S Tolias, and Matthias Bethge.
721 Benchmarking spike rate inference in population calcium imaging. *Neuron*, 90(3):471–
722 482, May 2016.
- 723 J D Victor and K P Purpura. Nature and precision of temporal coding in visual cortex:
724 a metric-space analysis. *J. Neurophysiol.*, 76(2):1310–1326, August 1996.
- 725 Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi,
726 Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train infer-
727 ence from population calcium imaging. *J. Neurophysiol.*, 104(6):3691–3704, December
728 2010.
- 729 Adrien Wohrer, Mark D Humphries, and Christian K Machens. Population-wide distri-
730 butions of neural activity during perceptual decision-making. *Prog. Neurobiol.*, 103:
731 156–193, April 2013.
- 732 Emre Yaksi and Rainer W Friedrich. Reconstruction of firing rate changes across neuronal
733 populations by temporally deconvolved ca2+ imaging. *Nat. Methods*, 3(5):377–383, May
734 2006.