Supervised learning sets benchmark for robust spike detection from calcium imaging signals

Lucas Theis[1,2]*, Philipp Berens[$*1,2,3,4], Emmanouil Froudarakis[4], Jacob Reimer[4], Miroslav Román Rosón[1,5], Tom Baden[1,3,5], Thomas Euler[1,3,5], Andreas Tolias[3,4,6], Matthias Bethge[1,2,3]

[1] Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany
[2] Institute of Theoretical Physics, University of Tübingen, Tübingen, Germany
[3] Bernstein Center for Computational Neuroscience, University of Tübingen, Tübingen, Germany
[4] Department of Neuroscience, Baylor College of Medicine, Houston, USA
[5] Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany
[6] Department of Computational and Applied Mathematics, Rice University, Houston, USA

* These authors contributed equally to this work.

$ To whom correspondence should be addressed:
Philipp Berens, philipp.berens@uni-tuebingen.de

We present a new data-driven approach to inferring spikes from calcium imaging signals using supervised training of non-linear spiking neuron models. Our technique yields a substantially better performance compared to previous generative modeling approaches, reconstructing spike trains accurately at high temporal resolution even from previously unseen datasets. Future data acquired in new experimental conditions can easily be used to further improve its spike prediction accuracy and generalization performance.

Over the past two decades, two-photon imaging has become one of the most widely used techniques for studying information processing in neural populations in vivo (Kerr and Denk, 2008; Denk et al., 1990). Typically, a calcium indicator such as the synthetic dye Oregon green BAPTA-1 (OGB-1) (Stosiek et al., 2003) or the genetically encoded GCamp6 (Chen et al., 2013) is used to image a large fraction of cells in a neural tissue. Individual action potentials lead to a fast rise in fluorescence, followed by a slow decay with a time constant of several hundred milliseconds (Kerr et al., 2005; Chen et al., 2013). Commonly, neural population activity from dozens or hundreds of cells is imaged using relatively slow scanning speeds (<15 Hz), but novel fast scanning methods (Cotton et al., 2013; Grewe et al., 2010) (up to several 100 Hz) have opened additional opportunities for studying neural population activity at increased temporal resolution.

A fundamental challenge has been to infer the timing of action potentials from the measured noisy calcium fluorescence traces. To solve the problem of spike inference, several methods have been proposed, including template-matching (Greenberg et al., 2008; Grewe et al., 2010; Oñativia et al., 2013) and deconvolution (Yaksi and Friedrich, 2006; Vogelstein et al., 2010, 2009). All these methods have in common that they assume a forward generative model of calcium signal generation which is then inverted to infer spike times (for a notable exception see Sasaki et al., 2008). A crucial shortcoming of this approach is that the forward models rely on a-priori assumptions about the shape of the calcium fluorescence signal induced by a single spike and the statistics of the noise.

In addition, the question how well we can reconstruct the spikes of neurons given calcium measurements has mostly been studied theoretically or using simulated datasets (Wilt et al., 2013; Lütcke et al., 2013). While such studies offer the advantage that many model parameters are under the control of the investigator, they do not answer the question how well we can reconstruct spikes from actual measurements. Unfortunately, most published algorithms have only been evaluated on relatively small experimental datasets. Therefore, a detailed quantitative comparison of different algorithms for reconstructing spikes from calcium traces on large datasets of *in vitro* and *in vivo* population imaging data has been lacking.

Here, we advocate a data-driven approach based on flexible probabilistic models to infer spikes from calcium fluorescence traces. We model the probabilistic relationship between a segment of the fluorescence trace $\boldsymbol{x}_t$ and the number of spikes k_t in a small time bin, assuming they are Poisson distributed with rate $\lambda(\boldsymbol{x}_t)$:

$$p(k_t \mid \boldsymbol{x}_t) = \frac{\lambda(\boldsymbol{x}_t)^k}{k!} e^{-\lambda(\boldsymbol{x}_t)}$$

.

Instead of relying on a specific forward model, we modeled the firing rate $\lambda(\boldsymbol{x}_t)$ using a recently introduced extension of generalized linear models, the factored spike-triggered mixture (STM) model (Theis et al., 2013) (Fig. 1a; see Methods):

$$\lambda_{\mathrm{STM}}(\boldsymbol{x}_t) = \sum_{k=1}^{K} \exp\left(\sum_{m=1}^{M} \beta_{km}(\boldsymbol{u}_m^{\top}\boldsymbol{x}_t)^2 + \boldsymbol{w}_k^{\top}\boldsymbol{x}_t + \boldsymbol{b}_k\right),$$

This model learns a set of K linear and M quadratic features $\boldsymbol{w}_k$ and $\boldsymbol{u}_m$ from the fluorescence trace, which predict the occurrence of spikes. Importantly, it is sufficiently flexible to capture non-linear relationships between fluorescence traces and spikes, but at the same time is sufficiently restricted to avoid overfitting when little data is available. Using a

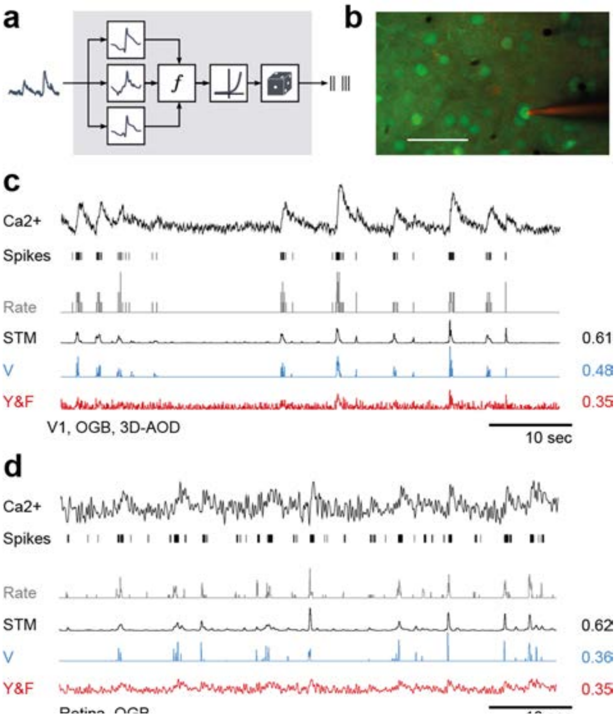


Figure 1: Spike inference from calcium measurements using flexible probabilistic models

a) Schematic of the probabilistic STM model.
b) Simultaneous recording of spikes and calcium fluorescence traces in primary visual cortex of anesthetized mice. Green: Cells labeled with OGB-1 indicator. Red: Patch pipette filled with Alexa Fluor 594 Scale bar: 50 µm.
c) Example cell recorded from V1. From top to bottom: Calcium fluorescence trace, spikes, spike rate at 25 Hz (grey), inferred spike rate using the STM model (black), the method by Vogelstein et al.(Vogelstein et al., 2010) (V; blue) and Yaksi & Friedrich(Yaksi and Friedrich, 2006) (Y&F; red). All traces were scaled independently for better visibility. On the right, correlation between the inferred and the original spike rate is shown.
d) Example cell recorded from the retina. For legend, see c).

probabilistic model in this way not only provides us with a point estimate of the expected firing rate, $\lambda(\boldsymbol{x}_t)$, but also with easy access to a full distribution over spike counts. In contrast to previous approaches, this allows us to estimate the uncertainty in the predictions and to generate example spikes trains without spending considerable computational resources.

To quantitatively evaluate our framework and compare it to other approaches, we acquired a benchmark dataset with a total of 56 neurons from which we simultaneously recorded calcium traces and spikes (Fig. 1b), recorded with different scanning methods and at different sampling rates (in total ~ 65,000 spikes; see Table 1 and Methods): Dataset 1 consisted of 16 neurons recorded *in-vivo* in primary visual cortex of anesthetized mice using fast 3D AOD-based imaging (Cotton et al., 2013) at ~320 Hz (V1, 3D). Dataset 2 consisted of 31 neurons recorded *in-vivo* in primary visual cortex of anesthetized mice using line scanning at ~12 Hz (V1, 2D). Finally, dataset 3 consisted of 9 retinal ganglion cells recorded *in-vitro* at ~8 Hz (Briggman and Euler, 2011) (Ret, 2D). We resampled calcium traces from all three datasets to a common resolution of 100 Hz. All three datasets were acquired at a zoom factor commonly used in population imaging and thus the signal quality should match well that commonly encountered in these preparations (see Table1).

We trained our algorithm on all but one cell from a dataset and tested it on the one remaining neuron (cross-validation; see Methods). We compared its performance to that of the two most widely used techniques in the literature, deconvolution (Yaksi and Friedrich, 2006) (referred to as *YF*) and ‘fast-oopsi’, an algorithm which performs approximate inference in a biophysical model (Vogelstein et al., 2010) (*V*). To provide a baseline performance, we evaluated how closely the calcium trace followed the spike train without any further processing (*raw*). We computed three different complimentary measures of spike reconstruction performance to provide an extensive quantitative evaluation of the different techniques: (i) accuracy of spike reconstruction as measured by the area under the ROC curve (AUC), (ii) correlation between original and reconstructed spike train and (iii) relative information gain (see Methods).

We found that the spike trains predicted by the probabilistic STM model matched the true spike trains of the neurons very well, both for cortical and retinal neurons (Fig. 1c and d). Indeed, our method reconstructed the true spike trains significantly more accurately than its competitors, yielding a higher correlation and relative information gain for all three datasets as well (Fig. 2a and Supp. Fig. 1a, b; evaluated at 25 Hz; AUC = 0.92, 0.88, 0.81 for the three datasets respectively; for statistics see figures). We next evaluated the performance at a wide range of sampling rates corresponding to time bins between 10 and several hundreds of milliseconds (see Methods). Again, our approach performed better than the other methods, especially at very high sampling rates (Fig. 2b-d; also Suppl. Fig. 2a-f). Interestingly, the AUC of the STM model was highest for neurons with low firing rates (Fig. 2e; Spearman’s rho: -0.67, p<0.001, N=56), while we observed no correlation between the performance of our algorithm and the firing rate of the neurons for the other two performance measures (rho=0.02, p=0.87; rho=-0.01, p=0.92; for correlation and information gain, respectively).

The STM model also allows one to look at the features in the calcium trace that were predictive of the occurrence of a spike. The linear filters resemble of the typical Ca^{2+}-transient, while the quadratic features seem to provide an estimate of the trace variance before and shortly after the spike (Suppl. Fig. 3a). In addition, we estimated a non-linear function to optimally map the rates predicted by the different algorithms to the true rates. We found that this function was the close to the identity for the STM (Suppl. Fig. 3b). The fast-oopsi algorithm consistently underestimated the true rate, while the deconvolution method required an approximately exponential non-linearity.

Since supervised learning requires training data, we evaluated how the performance of the STM model scaled with the number of available neurons. We found that performance saturates rather quickly between about 7 and 10 neurons, with additional neurons providing only small improvements (Fig. 2f and Suppl. Fig. 4a). We also tested whether a simple linear-nonlinear model would suffice to produce the performance of our algorithm; or whether a

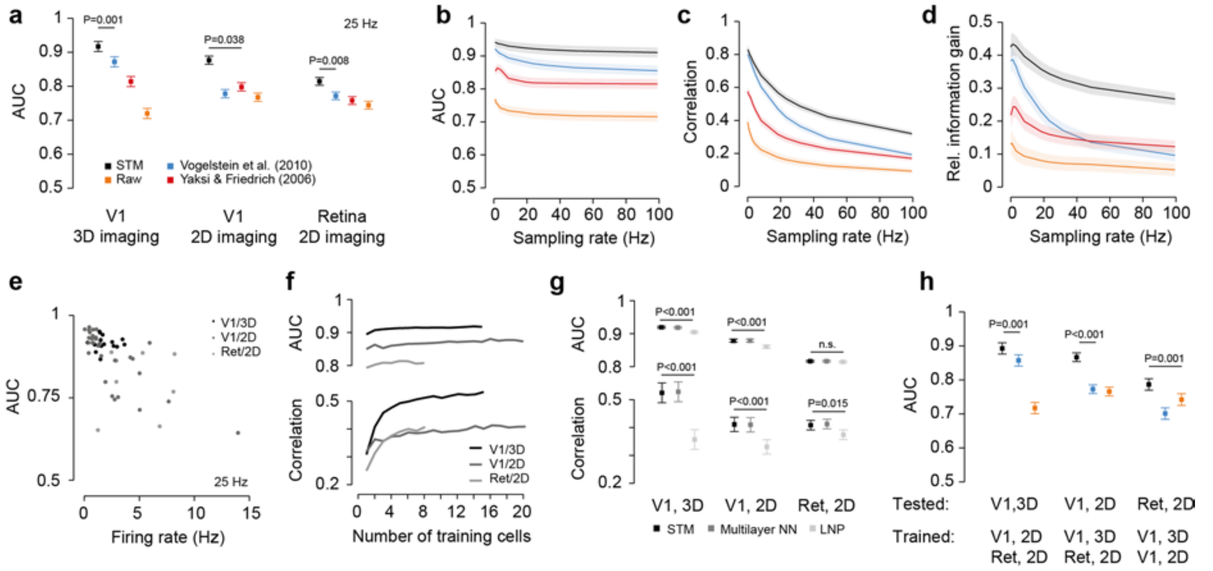


Figure 2: Quantitative evaluation of spike inference performance

a) Accuracy (measured as AUC) for the STM model (black) and the algorithms by Vogelstein et al. (blue) and Yaksi & Friedrich (red) evaluated on the three datasets (see Table 1). Additionally, accuracy obtained by the raw calcium trace (raw, orange) is shown as a baseline. Markers denote mean ± 2 standard error of the mean for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with its best competitor (see Methods).

b) Accuracy (measured as AUC) of the three algorithms and raw calcium as a function of sampling rate evaluated on the V1, 3D imaging dataset. For accuracy on the other datasets, see Suppl. Fig. 2. Colors as in a). Lines denote mean ± 2 SEM for repeated measure designs.

c) Correlation between inferred and true spike rate for the three algorithms and raw calcium as a function sampling rate evaluated on the V1, 3D imaging dataset. For correlation on the other datasets, see Suppl. Fig. 2. Colors as in a). Lines denote mean ± 2 standard error of the mean for repeated measure designs.

d) Relative information gain for the three algorithms and raw calcium as a function sampling rate evaluated on the V1, 3D imaging dataset. For relative information gain on the other datasets, see Suppl. Fig. 2. Colors as in a). Lines denote mean ± 2 SEM for repeated measure designs.

e) Dependence of accuracy (AUC) on firing rate of the neuron across the three datasets.

f) Accuracy (AUC) and correlation of the STM algorithm as a function of the number of training cells. Since the datasets have different size, the lines stop at different values.

g) Accuracy (AUC) and correlation using the STM nonlinearity (black), a more flexible multilayer neural-network (dark grey) or a more restricted linear-nonlinear model (light grey). Markers denote mean ± 2 SEM for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with the LNP model (see Methods). All differences between multilayer NN and STM are not significant (P>0.05 for each comparison).

h) Accuracy (AUC) for STM model, Vogelstein et al.'s algorithm and raw calcium signal trained on two datasets (bottom row) and evaluated on the third dataset (top row), mimicking the situation in which no simultaneous spike and calcium measurements are available. Colors as in a). Markers denote mean ± 2 SEM for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with its best competitor (see Methods).

more flexible non-linearity could improve it by using a multilayer neural network instead of the STM. We found that the STM model provided a good compromise between flexibility and generalization (Fig. 2g and Suppl. Fig. 4b): Its performance was significantly better than that of the simple LNP model (for the retina dataset and accuracy as performance measure the increase was not significant), but could only be marginally improved by using a multi-layer neural network.

Finally, we wanted to know how well our supervised learning-based algorithm can generalize to new types of data that are recorded under different conditions than the data used for training. To do so, we trained the algorithm on two of our datasets and evaluated it on the remaining one. Remarkably, it still was significantly more accurate than its competitors (Fig. 2g and Suppl. Fig. 5a,b; for all but the V1/3D dataset with correlation as performance measure the performance increase was significant), indicating that the algorithm may be directly applied on novel datasets without need for further training. In addition, any labeled data that will be collected in the future can readily be used to further improve spike prediction of the algorithm and its generalization to new experimental conditions (e.g. different neural systems and indicators). In fact, our choice of the spike triggered mixture model for estimating spikes from calcium traces is motivated by its ability to automatically switch between different sub-models whenever the statistics of the data changes (Theis et al., 2013). In this way, our approach constitutes a viable data-driven framework that promises steady improvement of performance in the future as more data becomes available and provides a benchmark dataset on which new algorithms can be readily evaluated..

Acknowledgements

We would like to thank J. Vogelstein and R. Friedrich for providing the code for their algorithms.

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Center for Computational Neuroscience (FKZ 01GQ1002 to T.E. and M.B.); the Deutsche Forschungsgemeinschaft (DFG) through grant BE3848-1 to M.B.; the Werner Reichardt Centre for Integrative Neuroscience Tübingen (EXC307); grants NEI R01-EY018847, NEI P30-EY002520-33, and the NIH-Pioneer award DP1-OD008301 to A.S.T.; the McKnight Scholar Award to A.S.T.

Author contributions

PB, MB and LT designed the project. LT analyzed the data. MF, JR and AST acquired V1 data. MR, TB and TE acquired retinal data. PB wrote the paper with input from all authors. PB and MB supervised the project.

Methods

Datasets

Primary visual cortex (V1)

We recorded calcium traces from neural populations in layer 2/3 of anesthetized wild type mice (male C57CL/6J, age: p40–p60) using a custom-built two-photon microscope using previously described methods(Cotton et al., 2013; Froudarakis et al., 2014). Briefly, the temperature of the mouse was maintained between 36.5 °C and 37.5 °C throughout the experiment using a homeothermic blanket system (Harvard Instruments). While recording we either provided no visual stimulation, moving gratings, or natural and phase scrambled movies as previously described(Froudarakis et al., 2014). A ~1 mm craniotomy was performed over the primary visual cortex of the mouse. The details of surgical techniques and anesthesia protocol have been described elsewhere(Cotton et al., 2013). We then used bolus-loaded Oregon green BAPTA-1 (OGB-1, Invitrogen) as calcium indicator and the injections were performed by using a continuous-pulse low pressure protocol with a glass micropipette to inject ~300 μm below the surface of the cortex. The cortical window was sealed using a glass coverslip. After allowing 1h for the dye uptake we recorded calcium traces using a custom-built two-photon microscope equipped with a Chameleon Ti-sapphire laser (Coherent) tuned at 800 nm and a 20×, 1.0 NA Olympus objective. Scanning was controlled by either a set of galvanometric mirrors (2D imaging) or a custom-built acousto-optic deflector system (AODs; 3D imaging)(Cotton et al., 2013). The average power output of the objective was kept < 50 mW for 2D imaging and 120 mW for 3D imaging. Calcium activity was typically sampled at ~12 Hz with the galvanometric mirrors and at ~320 Hz with the AODs. The field of view was typically 200x200x100μm and 250x250μm for 3D and 2D imaging, respectively, imaging dozens to hundreds of neurons simultaneously(Cotton et al., 2013). To perform simultaneous loose-patch and two-photon calcium imaging recordings, we used glass pipettes with 5–7 MΩ resistance filled with Alexa Fluor 594 (Invitrogen) for targeted two-photon-guided loose cell patching of single cells. Spike times were extracted by thresholding. All procedures performed on mice were conducted in accordance with the ethical guidelines of the National Institutes of Health and were approved by the Baylor College of Medicine IACUC.

Retina

Imaging experiments were performed as described previously(Briggman and Euler, 2011). In short, the retina was enucleated and dissected from dark-adapted wild-type mice (both genders, C57BL/6J, p21-42), flattened, mounted onto an Anodisc (13, 0.1 mm pores, Whatman) with ganglion cells facing up, and electroporated with Oregon green BAPTA-1 (OGB-1, Invitrogen). The tissue was placed under the microscope, where it was constantly perfused with temperated (36°C) carboxygenated (95% O_2, 5% CO_2) artificial cerebral spinal fluid (ACSF). Cells were left to recover for at least 1 hour before recordings were performed. We used a MOM-type two-photon microscope equipped with a mode-locked Ti:sapphire laser (MaiTai-HP DeepSee, Newport Spectra-Physics) tuned to 927 nm(Euler et al., 2009). OGB-1 Fluorescence was detected at 520 BP 30 nm (AHF) under a 20x objective (W Plan-Apochromat, 1.0 NA, Zeiss). Data were acquired with custom software (ScanM by M. Müller and T. Euler running under IgorPro 6.3, Wavemetrics), taking 64 x 64 pixel images at 7.8 Hz. Light stimuli were presented through the objective from a DLP projector (K11, Acer), fitted with band-pass-filtered LEDs (amber, z 578 BP 10; and blue/UV, HC 405 BP 10, AHF/Croma), synchronized with the microscope's scanner. Stimulator intensity (as photoisomerization rate, 10^4 R*/s/cone) was calibrated as described to range from 0.1 (LEDs off) to ~1.3 (ref (Euler et al., 2009)). Mostly due to two-photon excitation of photopigments, an additional, steady illumination component of ~10^4 R*/s/cone was present during the recordings. The field of view was 100x100µm, imaging 50-100 cells in the ganglion cell layer simultaneously(Briggman and Euler, 2011). For juxtacellular spike recordings, OGB-1 labeled somata were targeted with a 5 MΩ glass-pipette under dim IR illumination to establish a loose (<1GΩ) seal. Signals were amplified using an Axopatch 200A amplifier (Molecular Devices) in I=0 mode and digitized at 10 kHz on a Digidata 1440A (Molecular Devices). Imaging and spike data were aligned offline using a trigger signal recorded in both acquisition systems, and spike times were extracted by thresholding. All procedures were performed in accordance with the law on animal protection (Tierschutzgesetz) issued by the German Federal Government and were approved by the institutional animal welfare committee of the University of Tübingen.

Preprocessing

We normalized the sampling rate of all fluorescence traces and spike trains to 100 Hz, resampling to time bins of 10 ms. This allowed us to apply models across datasets independent of which dataset was used for training. We removed linear trends from the fluorescence traces by fitting a robust linear regression with Gaussian scale mixture residuals. That is, for each fluorescence trace F_t, we found parameters a, b, π_k, and σ_k with maximal likelihood under the model

$$F_t = at + b + \varepsilon_t, \qquad \varepsilon_t \sim \sum_{k=1\ldots K} \pi_k \mathcal{N}(\,\cdot\,; 0, \sigma_k^2),$$

and computed $\widetilde{F}_t = F_t - at - b$. We used three different noise components ($K = 3$). Afterwards, we normalized the traces such that the 5th percentile of each trace's fluorescence distribution is at zero, and the 80th percentile is at 1. Normalizing by percentiles instead of the minimum and maximum is more robust to outliers and less dependent on the firing rate of the neuron producing the fluorescence.

Algorithm

We predict the number of spikes k_t falling in the t-th time bin of a neuron's spike train based on 1000 ms windows of the fluorescence trace centered around t (preprocessed fluorescence snippets $\boldsymbol{x}_t$). To reduce the risk of overfitting and to speed up the training phase of the algorithm, we reduced the dimensionality of the fluorescence windows via PCA, keeping enough principal components to explain at least 95% of the variance (which resulted in 8 to 20 dimensions, depending on the dataset). Keeping 99% of the variance and slightly regularizing the model's parameters gave similar results but was slower.

We assume that the spike counts k_t given the preprocessed fluorescence snippets $\boldsymbol{x}_t$ can be modeled using a Poisson distribution,

$$p(k_t \mid \boldsymbol{x}_t) = \frac{\lambda(\boldsymbol{x}_t)^k}{k!} e^{-\lambda(\boldsymbol{x}_t)}$$

.

We tested three models for the firing rate $\lambda(\boldsymbol{x}_t)$ function:

(1) A simple linear-nonlinear-Poisson (LNP) neuron with exponential nonlinearity,

$$\lambda_{\mathrm{LNP}}(\boldsymbol{x}_t) = \exp(\boldsymbol{w}^\top \boldsymbol{x}_t),$$

where $\boldsymbol{w}$ is a linear filter.

(2) A spike-triggered mixture (STM) model(Theis et al., 2013) with exponential nonlinearity,

$$\lambda_{\mathrm{STM}}(\boldsymbol{x}_t) = \sum_{k=1}^{K} \exp\left(\sum_{m=1}^{M} \beta_{km}(\boldsymbol{u}_m^\top \boldsymbol{x}_t)^2 + \boldsymbol{w}_k^\top \boldsymbol{x}_t + b_k\right),$$

where $\boldsymbol{w}_k$ are linear filters, $\boldsymbol{u}_m$ are quadratic filters weighted by β_{km} for each of K components, and b_k is a offset for each component. We used three components and two quadratic features ($K = 3,\ M = 2$). The performance of the algorithm was not particularly sensitive to the choice of these parameters (we evaluated $K = 1, \dots 4$ and $M = 1, \dots, 4$ in a grid search).

(3) And a multi-layer neural network (ML-NN) with two hidden layers,

$$\lambda_{\mathrm{ML-NN}}(\boldsymbol{x_t}) = \exp(\boldsymbol{w}_3^\top g(\boldsymbol{W}_2 g(\boldsymbol{W}_1 \boldsymbol{x}_t + \boldsymbol{b}_1) + \boldsymbol{b}_2) + b_3)$$

,

where $g(\boldsymbol{y}) = \max(0, \boldsymbol{y})$ is a point-wise rectifying nonlinearity and $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are matrices. We tested MLPs with 10 and 5 hidden units, and 5 and 3 hidden units for the first and second hidden layer, respectively. Again, the performance of the algorithm was not particularly sensitive to those parameters.

Parameters of all models were optimized by maximizing the average log-likelihood for a given training set,

$$\frac{1}{N}\sum_{n=1}^{N} \log p(k_t \mid \boldsymbol{x}_t),$$

using limited-memory BFGS(Byrd et al., 1995), a standard quasi-Newton method. To increase robustness against potential local optima in the likelihood of the STM and the ML-NN, we trained four models with randomly initialized parameters and geometrically averaged their predictions. The geometric average of several Poisson distributions again yields a Poisson distribution whose rate parameter is the geometric average of the rate parameters of the individual Poisson distributions.

Performance evaluation

Typically, we evaluated the performance of the algorithms on spike trains binned at 40 ms resolution, i.e. a sampling rate of 25 Hz. For Fig. 2b-d and Suppl. Fig. 2, we changed the bin width between 10 ms (i.e. 100 Hz) and 500 ms (i.e. 2 Hz). We used cross-validation to evaluate the performance of our framework, i.e. we estimated the parameters of our model on a training set, typically consisting of all but one cell for each dataset, and evaluated its performance on the remaining cell. Results obtained using different splits into training and test sets were subsequently averaged.

Accuracy

By thresholding predictions of an algorithm we obtain a classification of time bins into bins with spikes and bins with no spikes. The ROC curve is obtained by varying the threshold and plotting the false positive rate (bins falsely classified as containing spikes) against the true positive rate (correctly predicted spikes). Finally, the AUC score is given by the area under the ROC curve. Equivalently, the AUC score can be computed as the probability that a randomly picked prediction for a bin containing a spike is larger than a randomly picked prediction for a bin containing no spike(Fawcett, 2006). This formulation generalizes more easily to the case with multiple spikes per bin (bins are picked with probability proportional to the number of spikes) and is the one we used to compute the score. While this is a commonly used score for evaluating spike inference procedures(Vogelstein et al., 2010), it is invariant under arbitrary strictly monotonically increasing transformations and does not take the uncertainty of the prediction into account.

Correlation

We computed the linear correlation coefficient between the true binned spike train and the inferred one. Similar to the AUC score, this a widely used measure with a simple and intuitive interpretation, but is invariant under affine transformations of the predictions and does not take the uncertainty of the predictions into account.

Relative information gain

The information gain quantifies the amount of information about the spike train extracted from the calcium trace. Unlike AUC and correlation, it takes into account the uncertainty of the prediction.

Assuming an average firing rate of λ and a predicted firing rate of λ_t at time t, the expected information gain (in bits per bin) can be estimated as

$$I_g = \frac{1}{T}\sum_t k_t \log_2 \frac{\lambda_t}{\lambda} + \lambda - \frac{1}{T}\sum_t \lambda_t$$

,

assuming Poisson statistics and independence of spike counts in different bins. The estimated information gain is bounded from above by the (unknown) amount of information about the spike train in the calcium trace, and the marginal entropy of the spike train, which can be estimated using

$$H_m = \frac{1}{T}\sum_t \log(k_t\,!) - \lambda \log \lambda + \lambda.$$

Dividing the information gain by the marginal entropy yields the relative information gain, a number between 0 and 1. This can be interpreted as the amount of entropy explained by the model. Since only our method was optimized to yield Poisson firing rates, we allowed each method a single monotonically increasing nonlinear function, which we optimized to maximize the average information gain over all cells. That is, we evaluated

$$\frac{1}{T}\sum_t k_t \log_2 \frac{f(\lambda_t)}{\lambda} + \lambda - \frac{1}{T}\sum_t f(\lambda_t),$$

where f is a piecewise linear monotonically increasing function optimized to maximize the information gain averaged over all cells (using an SLSQP implementation in SciPy). For visualization purposes, we slightly regularized the functions to be smooth.

Other algorithms

Yaksi & Friedrich (2006): This deconvolution algorithm(Yaksi and Friedrich, 2006) assumes that the fluorescence trace has been created by linearly convolving the spike train with an exponential calcium transient. To invert this process, it removes noise by local smoothing and then applies the inverse filter resulting from the calcium transient. Using the cross-validation procedure outlined above, we automatically tuned the algorithm by testing 66 different parameter configurations and choosing the one producing the highest correlation with the measured spike trains of all but one cell and evaluating performance on the remaining cell. The parameters controlled the cutoff frequency of a low-pass filter, a time constant of the filter used for deconvolution, and whether or not an iterative smoothing procedure was applied to the fluorescence traces.

Vogelstein et al. (2010): The fast-oopsi technique or non-negative deconvolution technique assumes a more complex forward model of the fluorescence trace and performs approximate inference in this model(Vogelstein et al., 2009). In contrast to Yaksi & Friedrichs algorithm, it restricts the inferred spike rates to be positive. Using the cross-validation approach outlined above, we adjusted the hyperparameters on the training set by performing a search over a grid of 54 parameter sets controlling the degree of assumed observation noise and the expected number of spikes.

Statistical analysis

We used generalized Loftus & Masson standard errors of the means for repeated measure designs(Franz and Loftus, 2012) and report the mean ± 2 SEM. To assess statistical significance, we compare the performance of the STM model to the performance of its next best competitor, performing a one-sided Wilcoxon signed rank test and report the respective p-value above a line spanning the respective columns.

Code and Data sharing

All analysis was done in Python. We provide a Python implementation of our algorithm as well as the datasets used for evaluating the algorithms online (www.bethgelab.org/code/spikeinference)[1].

[1] Please note that we are also preparing a Matlab implementation which will be released at a later point in time.

Table 1: Datasets

Dataset	Area	n	Indicator	Scan frequency (fps)	Scanning method	#spikes	# of spikes/ cell	Field of view
1	V1	16	OGB-1	322.5 ± 53.2	3D AOD	19,876	1242	200x200 x100 µm
2	V1	31	OGB-1	11.8 ± 0.9	2D galvo scan	32,385	1045	250x250 µm
3	Retina	9	OGB-1	7.8	2D galvo scan	12,488	1387	100x100 µm

References

Briggman, K. L., and Euler, T. (2011). Bulk electroporation and population calcium imaging in the adult mammalian retina. *J. Neurophysiol.* 105, 2601–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21346205 [Accessed July 30, 2013].

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* 16, 1190–1208.

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. a, Orger, M. B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499, 295–300. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3777791&tool=pmcentrez&rendertype=abstract [Accessed March 19, 2014].

Cotton, R. J., Froudarakis, E., Storer, P., Saggau, P., and Tolias, A. S. (2013). Three-dimensional mapping of microcircuit correlation structure. *Front. Neural Circuits* 7, 151. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3794294&tool=pmcentrez&rendertype=abstract [Accessed March 30, 2014].

Denk, W., Strickler, J., and Webb, W. (1990). Two-photon laser scanning fluorescence microscopy. *Science (80-.).* 248, 73–76. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3481855&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Euler, T., Hausselt, S. E., Margolis, D. J., Breuninger, T., Castell, X., Detwiler, P. B., and Denk, W. (2009). Eyecup scope--optical recordings of light stimulus-evoked fluorescence signals in the retina. *Pflugers Arch.* 457, 1393–414. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037819&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014].

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. Available at: http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X [Accessed July 9, 2014].

Franz, V. H., and Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychon. Bull. Rev.* 19, 395–404. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3348489&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2014].

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., Saggau, P., Bethge, M., and Tolias, A. S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat. Neurosci.* 17, 851–857. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24747577 [Accessed May 26, 2014].

Greenberg, D. S., Houweling, A. R., and Kerr, J. N. D. (2008). Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat. Neurosci.* 11, 749–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18552841 [Accessed March 20, 2014].

Grewe, B. F., Langer, D., Kasper, H., Kampa, B. M., and Helmchen, F. (2010). High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision.

Nat. Methods 7, 399–405. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20400966 [Accessed March 19, 2014].

Kerr, J. N. D., and Denk, W. (2008). Imaging in vivo: watching the brain in action. *Nat. Rev. Neurosci.* 9, 195–205. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18270513 [Accessed December 11, 2013].

Kerr, J. N. D., Greenberg, D., and Helmchen, F. (2005). Imaging input and output of neocortical networks in vivo. *Proc. Natl. Acad. Sci. U. S. A.* 102, 14063–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1201343&tool=pmcentrez&rendertype=abstract.

Lütcke, H., Gerhard, F., Zenke, F., Gerstner, W., and Helmchen, F. (2013). Inference of neuronal network spike dynamics and topology from calcium imaging data. *Front. Neural Circuits* 7, 201. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871709&tool=pmcentrez&rendertype=abstract [Accessed September 2, 2014].

Oñativia, J., Schultz, S. R., and Dragotti, P. L. (2013). A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. *J. Neural Eng.* 10, 046017. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4038919&tool=pmcentrez&rendertype=abstract [Accessed June 2, 2014].

Sasaki, T., Takahashi, N., Matsuki, N., and Ikegaya, Y. (2008). Fast and accurate detection of action potentials from somatic calcium fluctuations. *J. Neurophysiol.* 100, 1668–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18596182 [Accessed March 21, 2014].

Stosiek, C., Garaschuk, O., Holthoff, K., and Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7319–24. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165873&tool=pmcentrez&rendertype=abstract.

Theis, L., Chagas, A. M., Arnstein, D., Schwarz, C., and Bethge, M. (2013). Beyond GLMs: a generative mixture modeling approach to neural system identification. *PLoS Comput. Biol.* 9, e1003356. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3836720&tool=pmcentrez&rendertype=abstract [Accessed March 29, 2014].

Vogelstein, J. T., Packer, A. M., Machado, T. a, Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* 104, 3691–704. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3007657&tool=pmcentrez&rendertype=abstract [Accessed March 27, 2014].

Vogelstein, J. T., Watson, B. O., Packer, A. M., Yuste, R., Jedynak, B., and Paninski, L. (2009). Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophys. J.* 97, 636–55. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2711341&tool=pmcentrez&rendertype=abstract [Accessed March 21, 2014].

Wilt, B. a, Fitzgerald, J. E., and Schnitzer, M. J. (2013). Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing. *Biophys. J.* 104, 51–62.

Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540268&tool=pmcentrez&rendertype=abstract [Accessed August 2, 2014].

Yaksi, E., and Friedrich, R. W. (2006). Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. *Nat. Methods* 3, 377–83. Available at: http://www.nature.com/nmeth/journal/v3/n5/abs/nmeth874.html [Accessed March 21, 2014].

Supplementary Material

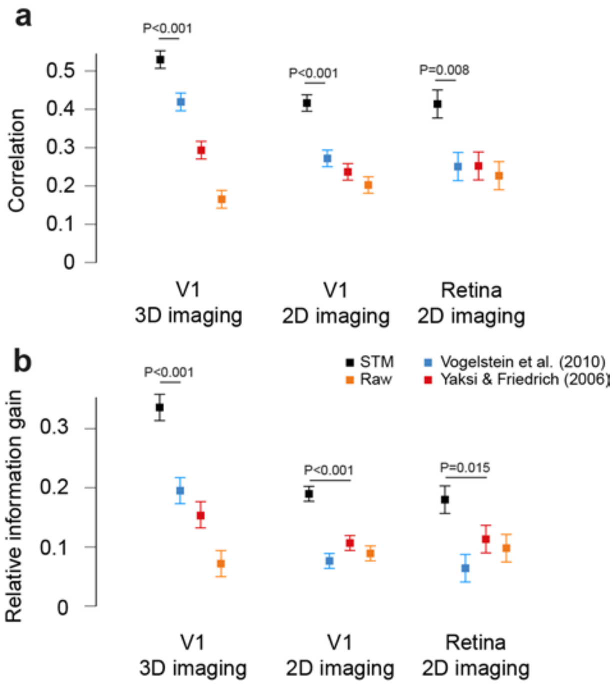
Supervised learning sets benchmark for robust spike detection from calcium imaging signals

Lucas Theis[1,2]*, Philipp Berens[$*1,2,3,4], Emmanouil Froudarakis[4], Jacob Reimer[4], Miroslav Román Rosón[1,5], Tom Baden[1,3,5], Thomas Euler[1,3,5], Andreas Tolias[3,4,6], Matthias Bethge[1,2,3]

[1] Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany
[2] Institute of Theoretical Physics, University of Tübingen, Tübingen, Germany
[3] Bernstein Center for Computational Neuroscience, University of Tübingen, Tübingen, Germany
[4] Department of Neuroscience, Baylor College of Medicine, Houston, USA
[5] Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany
[6] Department of Computational and Applied Mathematics, Rice University, Houston, USA

* These authors contributed equally to this work.

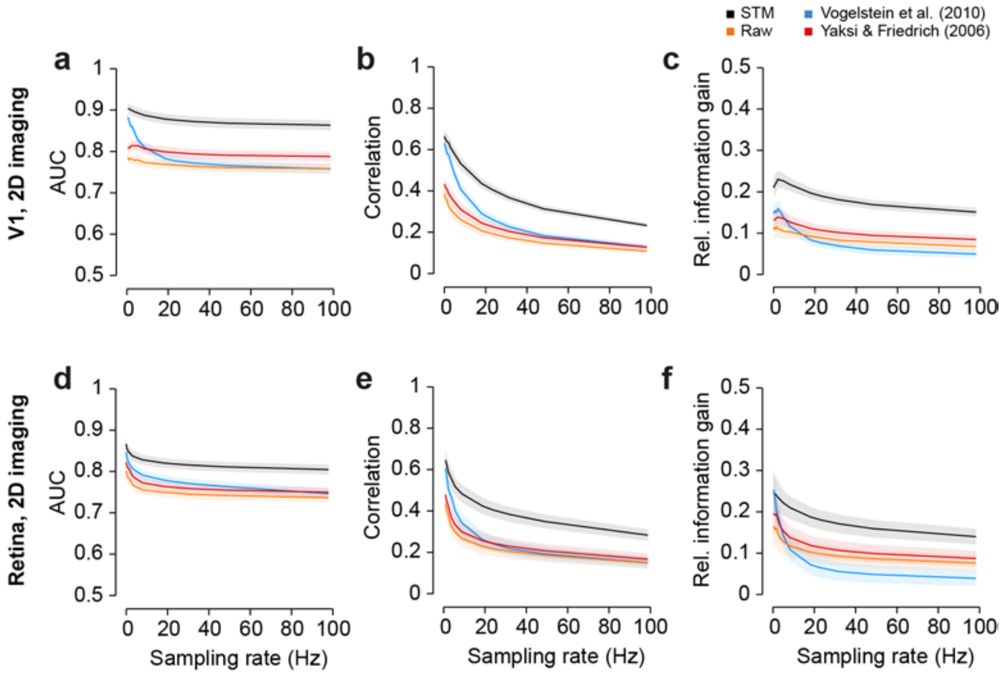
$ To whom correspondence should be addressed:
Philipp Berens, philipp.berens@uni-tuebingen.de



Supplementary Figure 1: Additional measures of performance

a) Correlation for the STM model (black) and the algorithms by Vogelstein et al. (blue) and Yaksi & Friedrich (red) evaluated on the three datasets (see Table 1). Additionally, correlation obtained by the raw calcium trace (raw, orange) is shown as a baseline. Markers denote mean ± 2 standard error of the mean for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with its best competitor (see Methods).

b) As in a) but for Relative Information Gain.



Supplementary Figure 2: Performance as a function of sampling rate for the two additional datasets

a) Accuracy (measured as AUC) of the three algorithms and raw calcium as a function of sampling rate evaluated on the V1, 2D imaging dataset. Lines denote mean ± 2 SEM for repeated measure designs.

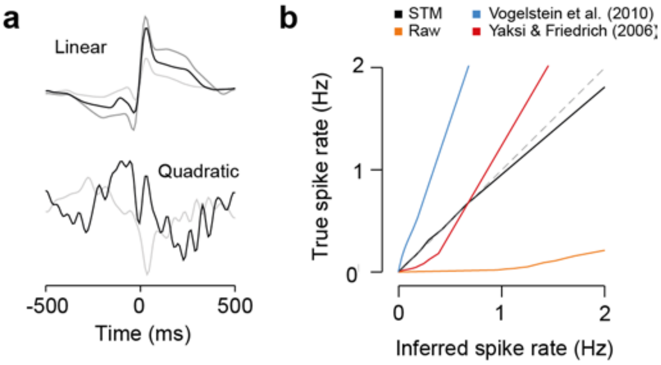
b) As in a, but for correlation as performance measure.

c) As in a, but for relative information gain as performance measure.

d) Accuracy (measured as AUC) of the three algorithms and raw calcium as a function of sampling rate evaluated on the retina, 2D imaging dataset. Lines denote mean ± 2 SEM for repeated measure designs.

e) As in a, but for correlation as performance measure.

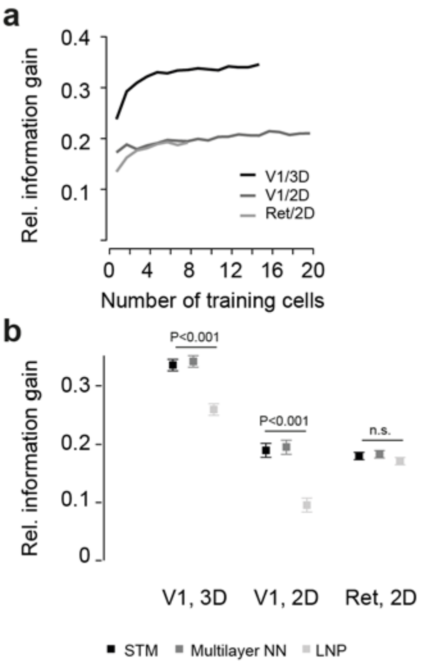
f) As in a, but for relative information gain as performance measure.



Supplementary Figure 3: Features learned by STM model and non-linearities

a) The three linear and two quadratic features learned by the STM model from the data, corresponding to the $\boldsymbol{w}_k$ and $\boldsymbol{u}_m$ in the definition of the STM.

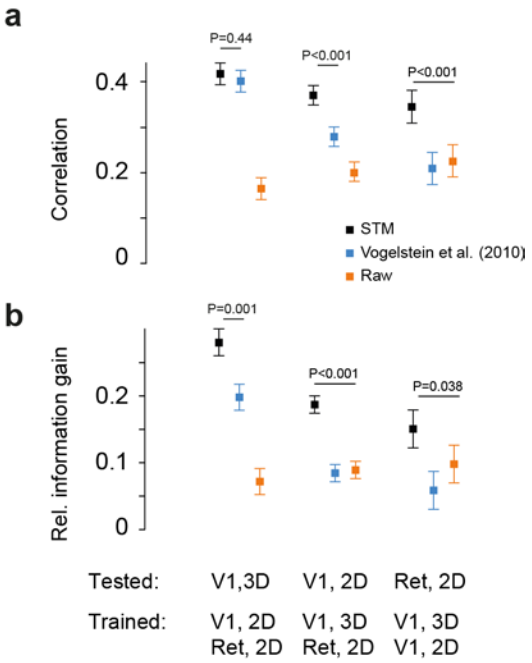
b) Piecewise linear functions inferred to model the non-linearity between the inferred and the true spike rate for the different compared algorithms.



Supplementary Figure 4: Training set size and model flexibility

a) Relative information gain of the STM algorithm as a function of the number of training cells. Since the datasets have different size, the lines stop at different values.

b) Relative information gain using the STM model (black), a more flexible multilayer neural-network (dark grey) or a more restricted linear-nonlinear model (light grey). Markers denote mean ± 2 SEM for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with the LNP model.



Supplementary Figure 5: Generalization performance

a) Correlation for STM model, Vogelstein et al.’s algorithm and raw calcium signal trained on two datasets (bottom row) and evaluated on the third dataset (top row), mimicking the situation in which no simultaneous spike and calcium measurements are available. Markers denote mean ± 2 SEM for repeated measure designs (see Methods). P-value results from a comparison of the performance of the STM with its best competitor (see Methods).

b) As in a. for relative information gain.