High frequency spike inference with particle Gibbs sampling

Giovanni Diana[1*], B. Semihcan Sermet[1], David A. DiGregorio[1,2*]

[1]Institut Pasteur, University of Paris, CNRS UMR 3571, Synapse and Circuit Dynamics Laboratory, Paris, France;
[2]Department of Physiology and Biophysics, University of Colorado School of Medicine, Aurora, CO 80045, USA
[*] giovanni.diana@pasteur.fr (GD); david.digregorio@pasteur.fr (DD)

Abstract

Fluorescent calcium indicators are indispensable tools for monitoring the spiking activity of large neuronal populations in animal models. However, despite the plethora of algorithms developed over the last decades, accurate spike time inference methods for spike rates greater than 20 Hz are lacking. More importantly, little attention has been devoted to the quantification of statistical uncertainties in spike time estimation, which is essential for assigning confidence levels to inferred spike patterns. To address these challenges, we introduce (1) a statistical model that accounts for bursting neuronal activity and baseline fluorescence modulation and (2) apply a Monte Carlo strategy (particle Gibbs with ancestor sampling) to estimate the joint posterior distribution of spike times and model parameters. Our method is competitive with state-of-the-art supervised and unsupervised algorithms by analyzing the CASCADE benchmark datasets. The analysis of fluorescence transients recorded using an ultrafast genetically encoded calcium indicator, GCaMP8f, demonstrates the ability of our method to infer spike time intervals as short as five milliseconds. Overall, our study describes a Bayesian inference method to detect neuronal spiking patterns and their uncertainty. The use of particle Gibbs samplers allows for unbiased estimates of spike times and all model parameters, and it provides a flexible statistical framework to test more specific models of calcium indicators.

1 Introduction

Fluorescence indicators of calcium activity allow us to monitor the dynamics of neuronal populations both in vivo and in vitro. In the last decade there has been a proliferation of new methods to identify single spikes from fluorescence time series using template matching (Dyer et al., 2013; Grewe et al., 2010; Kerr et al., 2005; Quan et al., 2010), linear deconvolution (Holekamp et al., 2008; Tubiana et al., 2020; Wei et al., 2020; Yaksi and Friedrich, 2006), finite rate of innovation (Oñativia and Dragotti, 2015; Oñativia et al., 2013), independent component analysis (Mukamel et al., 2009), non model-based signal processing (Sebastian et al., 2019), supervised learning (Hoang et al., 2020; Rupprecht et al., 2021; Sasaki et al., 2008; Sebastian et al., 2021; Theis et al., 2016), constrained non-negative matrix factorization (Pnevmatikakis et al., 2014, 2016; Zhou et al., 2018), active set methods (Friedrich and Paninski, 2016; Friedrich et al., 2017), convex and non-convex optimization methods (Jewell and Witten, 2018; Jewell et al., 2020; Malik et al., 2011; Ranganathan and Koester, 2010; Stern et al., 2020), and interior point method (Vogelstein et al., 2010).

Model-based approaches aim to describe the mechanistic link between spikes and fluorescence. This is typically achieved by introducing a set of time-dependent "state" variables such as calcium level and baseline modulation, and their temporal dynamics. The temporal evolution of the state variables usually depends on additional parameters that remain constant over time and define the dynamics. Depending on the model, state variables might not be fully observable, in which case we need to infer them from the data. In statistics, this class of models is referred to as state-space models, and they are used in time series analysis to describe the probabilistic dependence between data and unobserved variables. Previous works have used tractable state-space models to derive maximum-a-posteriori estimates of spike times (Deneux et al., 2016; Fletcher and Rangan, 2014; Kazemipour et al., 2017; Tsunoda et al., 2010b).

The vast majority of spike inference methods provide single estimates of spike times by minimizing a cost function defined by the underlying model and constraints. This optimization approach does not provide information about the statistical uncertainty associated with single-point estimates. To address this issue, previous studies proposed Bayesian inference methods (Greenberg et al., 2018; Huys and Paninski, 2009; Im et al., 2019; Mishchencko et al., 2011; Mishchenko and Paninski, 2011; Pnevmatikakis et al., 2013; Rahmati et al., 2016; Shibue and Komaki, 2020; Speiser et al., 2017; Theis et al., 2013; Tsunoda et al., 2010a; Vogelstein et al., 2009) which give access to the full probability distribution of the unknowns given the data.

However, the state-space models used in these approaches do not take into account the possibility of burst firing and slow changes in the fluorescence baseline, which is known to be an important issue for the analysis of in-vivo

recordings. Moreover, current Bayesian methods do not treat time-independent model parameters (e.g. rate constants) and dynamic variables equally. Instead, they require additional optimization procedures to calibrate model parameters, thus neglecting how the uncertainty about model parameters propagates into spike times.

Inference on non-linear and non-Gaussian state-space models is analytically intractable, requiring the application of Monte Carlo methods to obtain unbiased approximations of the posterior distributions. Because the number of unknowns in these models is of the order of the number of time steps in the fluorescence time-series, the analysis of long time series requires efficient strategies to sample from high-dimensional spaces. A major breakthrough in analyzing state-space models has been the introduction of sequential Monte Carlo methods (Chopin and Papaspiliopoulos, 2020). These algorithms can sample efficiently from the latent space by approximating sequentially the target posterior distribution by combining importance sampling and resampling techniques. In particular, the particle Gibbs algorithm can be used to obtain unbiased estimates of the joint distribution of time-independent model parameters and dynamical variables, but it has never been applied in the context of spike inference.

In this work we employ the particle Gibbs (PG) sampler on a bursting autoregressive (BAR) model of time series calcium-dependent fluorescence. Our generative model accounts for periods of high firing rates between periods of baseline (lower) firing rates. By quantifying the performance of our method (PGBAR) on the CASCADE benchmark dataset (Rupprecht et al., 2021) we showed that our approach is competitive with existing techniques. Finally, we tested PGBAR on in-vitro recordings of cerebellar granule cells using the ultrafast GCaMP8f calcium indicator, showing that our method permits the detection of spikes reliably even for high firing rates ($\sim 200Hz$).

2 Results

2.1 The model

We used a generative model of time-dependent fluorescence changes that will be used for spike inference. We model the normalized fluorescence $F_{1:T}$ as the sum of a calcium-dependent fluorescence level c_t (hereinafter referred to as calcium level for brevity), a time-varying baseline b_t and Gaussian noise η_t

$$F_t = c_t + b_t + \eta_t, \qquad t = 1, \cdots, T \tag{1}$$

where the fluorescence noise η_t is normally distributed with zero mean and variance σ^2. To describe bursts of calcium transients, we introduce firing states $q_t = 0, 1$, associated respectively with low and high firing rates, r_0 and r_1. We allow for stochastic transitions between these two states with rates $w_{0\to1}$ and $w_{1\to0}$. The probability of switching from q to q' within a sampling period Δ is given by the transition matrix

$$W = \begin{bmatrix} 1 - w_{0\to1}\Delta & w_{0\to1}\Delta \\ w_{1\to0}\Delta & 1 - w_{1\to0}\Delta \end{bmatrix} \tag{2}$$

The number of spikes at time t, s_t, is modeled by a Poisson distribution with rate r_1 when $q_t = 1$ otherwise with baseline firing rate r_0 when $q_t = 0$. The dynamics of the calcium level in response to a spike train is modeled as a second order autoregressive process

$$c_t = \begin{cases} c_0 + As_1 & t = 1 \\ \gamma_1 c_1 + As_2 & t = 2 \\ \gamma_2 c_{t-2} + \gamma_1 c_{t-1} + As_t & t > 2 \end{cases} \tag{3}$$

where c_0 is the initial calcium level and A controls the calcium increase upon single action potential. Note that in Eq. (3) the calcium level at time t depends on the previous calcium levels up to c_{t-2}. The dynamics of c_t in response to a single spike (kernel response) is characterized by a finite rise time (time to peak response) and exponential decay (see Section Response kernel for a derivation).

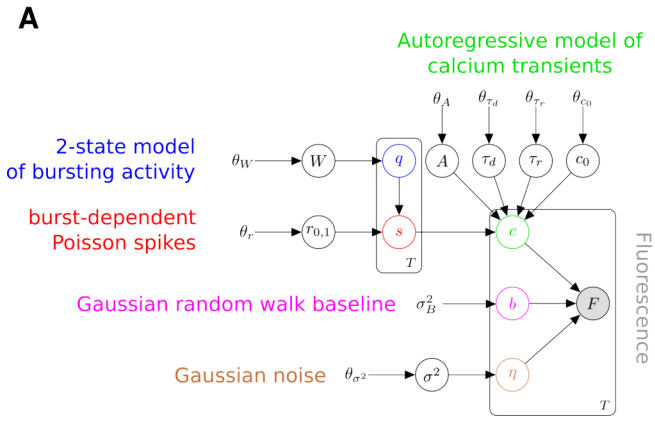
We can recast our model as a first order Markov process, where the state at time t only depends on the state at time $t-1$. This can be done by introducing a calcium vector and a spike count vector (see S2.2.3 in Ref. Pnevmatikakis et al. (2016))

$$C_t = \begin{bmatrix} c_t \\ c_{t-1} \end{bmatrix}, \quad S_t = \begin{bmatrix} s_t \\ 0 \end{bmatrix}, \tag{4}$$

where in particular the calcium vector at time t is constructed by combining the calcium levels at current and previous time. With this definition, the calcium vector C_t satisfies the first-order Markov dynamics

$$C_t = \begin{cases} [c_0 + As_1, 0] & t = 1 \\ M \cdot C_{t-1} + AS_t & t > 1 \end{cases}, \quad M = \begin{bmatrix} \gamma_1 & \gamma_2 \\ 1 & 0 \end{bmatrix}. \tag{5}$$

In this model the calcium dynamics is a deterministic process given the sequence of spikes $s_{1:T}$.



B

parameter	description	prior
$A^{(max)}$	Peak response	truncated normal
τ_r, τ_d	Rise and decay constants	truncated normal
c_0	Initial condition of the autoregressive model	normal
r_0, r_1	Baseline and burst firing rates	beta
$w_{0\to1}, w_{1\to0}$	Transition rates of the 2-state bursting process	gamma
σ^2	noise level	inverse gamma

Figure 1: **Generative model of fluorescence time series. (A)** Graphical representation of the generative model described in the main text. White circles denote unknown variables, grey circles denote measurements and bare variables are fixed prior hyperparameters. Plates denote groups of variables. **(B)** List of parameters and corresponding priors.

We reparameterize the autoregressive model in terms of phenomenological parameters: peak response ($A^{(max)}$), rise time (time to peak response τ_r) and decay time (τ_d) of unitary fluorescence response. Thanks to previous characterization of GCaMP probes (Chen et al., 2013), we can more easily design prior distributions for such phenomenological parameters rather than the autoregressive model parameters γ's, A and $\gamma_{1,2}$. $A^{(max)}$, τ_r and τ_d, referred to as kernel parameters in the upcoming sections, can be derived from $\gamma_{1,2}$ and A as (see Section Reparameterization for a derivation)

$$A^{(max)} = A \cdot g_A, \qquad g_A \equiv \left(\frac{g_+}{g_-}\right)^{\frac{g_+}{g_-}} \left(1 - \frac{g_+}{g_-}\right) \left(e^{g_+} - e^{g_-}\right)^{-1} \tag{6}$$

$$\tau_r = \frac{\log\left(\frac{g_+}{g_-}\right)}{g_- - g_+}, \qquad g_\pm = \log\left(\frac{\gamma_1 \pm \sqrt{\gamma_1 + 4\gamma_2}}{2}\right) \tag{7}$$

$$\tau_d = -\frac{1}{g_+}, \tag{8}$$

Finally, the fluorescence baseline B_t is described by a Gaussian random walk with normally distributed initial condition

$$\begin{cases} B_t \sim \mathcal{N}(0,1) & t = 1 \\ B_t \sim \mathcal{N}(B_{t-1}, \sigma_B^2 \Delta) & t > 1 \end{cases} \tag{9}$$

where Δ is the sampling period of the time series.

In the language of state-space models, the latent state of our model is the combination of the bursting state q_t, the spike count s_t, the calcium vector C_t and the baseline b_t, whereas the fluorescence F_t, defined in Eq. (1) is our observation. The time-independent parameters of our model are the firing rate constants $r_{0,1}$, the transition rates of the 2-state bursting process $W_{0\to1}, W_{1\to0}$, the kernel parameters of the calcium indicators (peak amplitude, rise and decay constants), the initial calcium level c_0 and the fluorescence noise σ. To simplify the notation we will denote the latent space as $X = \{q_t, s_t, C_t, b_t\}$ and the combination of time-independent parameters as θ.

2.2 State-space model formulation

The joint probability of the latent state trajectory $X_{1:T}$ and the fluorescence observations $F_{1:T}$ conditional to the time-independent parameters θ can be expressed as

$$P(X_{1:T}, F_{1:T}|\theta) = \mu^\theta(X_1) \cdot \prod_{t=2}^{T} f_t^\theta(X_t|X_t - 1) g_t^\theta(F_t|X_t) \tag{10}$$

where $f_t^\theta(X_t|X_{t-1})$ is the transition probability of the latent state, $g_t^\theta(F_t|X_t)$ is the probability of the observed fluorescence conditional to the latent state at time t and $\mu^\theta(X_1)$ is the probability distribution of the initial latent state. The latent state transition probability can be expressed in terms of calcium level, firing state and baseline

transitions and the Poisson probability of spike counts, namely

$$f_t^\theta(X_t|X_{t-1}) = \overbrace{\delta^{(2)}(C_t - M \cdot C_{t-1} - AS_t)}^{\text{deterministic calcium}} \cdot \overbrace{W_{q_{t-1}q_t}}^{\text{firing state}} \cdot \overbrace{\frac{(r_{q_t}\Delta)^{s_t}}{s_t!} e^{-r_{q_t}\Delta}}^{\text{Poisson spikes}} \cdot \overbrace{(2\pi\Delta\sigma_b^2)^{-1/2} \exp\left(-\frac{1}{2\Delta\sigma_b^2}(b_t - b_{t-1})^2\right)}^{\text{baseline}}. \tag{11}$$

By assuming the fluorescence noise to be normally distributed, we have

$$g_t^\theta(F_t|X_t) = (2\pi\sigma^2)^{-1/2} \cdot \exp\left[-\frac{1}{2\sigma^2}(F_t - c_t - b_t)^2\right]. \tag{12}$$

To infer latent states and time-independent parameters from fluorescence observations, we need to compute the posterior probability

$$P(X_{1:T}, \theta|F_t) = \frac{P(\theta) \cdot P(X_{1:T}, F_{1:T}|\theta)}{P(F_{1:T})}, \tag{13}$$

where $P(\theta)$ denotes the prior probability on model parameters and $P(F_{1:T})$ is the normalization factor of the posterior distribution, also known as marginal likelihood. This distribution encodes all the information about the statistics of the latent state trajectory and the model parameters. We can use it to compute point estimates but also to quantify uncertainties. The posterior distribution for general state-space models is not analytically tractable. However, we can use Monte Carlo methods to generate random samples from the posterior distribution and use them to obtain unbiased approximations of any posterior average.

2.3 Sequential Monte Carlo

In this work, we use Monte Carlo methods to approximate the posterior distribution in Eq. (13). There are two critical issues when applying these methods to state-space models. First the high-dimensionality of the posterior distribution and, second, the joint inference of state variables and constant parameters. In the general setting of a state-space model we need to estimate state variables at each time point. Consequently, the support of the target posterior distribution is a high-dimensional space. The high dimensionality of state-space models is an issue when applying standard sampling methods such as Markov Chain Monte Carlo, as their performance rapidly decreases with increasing dimensionality. Sequential Monte Carlo (SMC) methods address this issue by providing efficient sampling strategies from the latent space. The typical approach is to construct a sequential approximation of the posterior distribution where observations are accounted for iteratively. SMCs solve the problem of estimating dynamical variables at fixed model parameters (filtering).

The second issue, related to the joint estimation of parameters and dynamic variables, was addressed by Andrieu et al. (2010) with the introduction of the particle Gibbs algorithm. This algorithm alternates the sampling of model parameters and state variables as in the Gibbs sampler, with the difference that state variables are sampled from an SMC-based transition kernel that leaves the filtering distribution $P(X_{1:T}|F_{1:T}, \theta)$ invariant. In this work, we employ a version of the particle Gibbs algorithm developed by Lindsten et al. (2014), named particle Gibbs with ancestor sampling (PGAS), with better mixing properties (see Algorithm 2 and the Methods section).

To carry out inference of time-independent and dynamical variables for our model, we employed Algorithm 1, which alternates the two steps mentioned above: (1) run the PGAS transition kernel to sample a new trajectory of the state variables, and (2) sample model parameters from their full conditional distributions when analytically available, otherwise from a Metropolis-Hastings kernel.

Algorithm 1 Gibbs sampler

1: Set $\theta^{(1)}$ and $X_{1:T}^{(1)}$
2: **for** $n > 1$ **do**
3: draw $X_{1:T}^{(n)} \sim \mathcal{K}_{\theta^{(n-1)}}^N(X^{(n-1)}, \cdot)$ (PGAS kernel)
4: draw $\theta^{(n)} \sim P(\theta|X_{1:T}^{(n)}, F_{1:T})$

2.4 Validation and performance of PGBAR on simulated time-series fluorescence data

To test the performance of our inference method, we generated latent state variables and fluorescence time series from our model and compared the spikes inferred using our sampling algorithm against the ground truth simulations. In Fig. 2A we show a fluorescence time series simulated from our model. The firing pattern displays periods of increased firing rate separated by quiet time windows. By using this trace as input to Algorithm 1, we can generate a latent state trajectory $X_{1:T} = \{q_t, s_t, C_t, b_t\}_{t=1}^T$ and a set of model parameters at each iteration. In Fig. 2B we show 1000 samples of spike counts obtained by fitting the normalized fluorescence in Fig. 2A. The average spike counts over the random

samples at each time frame (Fig. 2C) can be interpreted as the instantaneous firing rate multiplied by the sampling period. To illustrate the accuracy of our method, we calculated the spike counts within 1s time intervals for each random sample, providing the posterior distribution of the number of spikes in each time bin. As shown in Fig. 2D, the ground-truth spike counts are well within the range of the posterior. Our method allows us to infer not only spike times but also the time windows of high and low firing states ($q_t = 0, 1$ in the model) of the neuron. In particular, the probability of a burst-firing state ($q = 1$) can be obtained by averaging the firing state across the Monte Carlo samples. As shown in Fig. 2E, this probability is close to one during the ground-truth bursting periods and zero otherwise, with some degree of uncertainty at the onset and offset of the bursting period. Fig. 2F compares the ground-truth baseline and the sample average. One of the key advantages of our sampling algorithm is the joint estimation of latent states and time-independent model parameters. Figure 2G illustrates the posterior distributions associated with the fluorescence probe's peak amplitude, noise level, decay, and rise time. The ground truth parameters used to simulate the testing time series are always close to the peak of the corresponding posterior distributions, showing that our model is identifiable.

To quantify the importance of having two firing states on the accuracy of the inference, we compared the performance of our method against a variant with only one global Poisson firing rate. We simulated fluorescence traces at different signal-to-noise (SNR) levels (1.1, 2 and 10), defined as the ratio between peak response $A^{(max)}$ and the fluorescence noise parameter σ, and the burst firing frequency parameter, r_1 (5Hz, 10Hz, 20Hz, 50Hz). Then, we used our algorithm and its non-bursting variant to infer spikes from the fluorescence trace. To quantify the inference performance, we calculated the correlation between ground truth and estimated spikes (downsampled at 7.5 Hz for consistency with other analyses in this work), the average absolute error and the bias (average error) per time point (see Methods). The top panel of Figure 3A shows two example traces with stimulation times generated by a 5 and 50Hz Poisson distribution of spike times. For the 5Hz firing trace, the bursting model did not improve the inference accuracy compared to the variant with a single firing rate(Fig. 3A, middle and bottom). The two analyses produced comparable correlations, errors and biases (Fig. 3C-E). In contrast, the single firing rate model induced a systematic underestimate of the number of spikes for the 50 Hz trace, the original model captures the ground-truth spikes reliably.

The lower performance of the non-bursting version of the model is due to the bias induced by forcing a single firing rate across the time series. While for all conditions of noise and frequency, inference using the bursting model gives unbiased spike counts (Fig. 3E), in the case of the non-bursting model, the single Poisson firing rate leads necessarily to an underestimation of the spike count during bursting time windows and an overestimation during low activity windows. At increasing noise levels and firing frequency, the performance difference between bursting and non-bursting versions of our algorithm become more pronounced (Fig. 3D), with a clear advantage of our original bursting model in increasing correlation with ground truth and reducing error.

2.5 Validation of PGBAR on the CASCADE benchmark data and comparison to previous methods

To test our method on experimental data we analyzed neuronal recordings from the CASCADE benchmark dataset (Rupprecht et al., 2021), which allowed us to quantify the performance of our algorithm on different calcium indicators. We compared our method to CASCADE (Rupprecht et al., 2021), MLSpike (Deneux et al., 2016), Peeling (Lütcke et al., 2013), CaImAn (Giovannucci et al., 2019), Suite2p (Pachitariu et al., 2017) and JewellWitten (Jewell et al., 2020) by using their previously benchmarked performance on the same datasets obtained from extensive parameter optimization (Rupprecht et al., 2021) (Fig. 4C). As a metric for inference performance, we used the Pearson's correlation coefficient between ground-truth spikes and predicted spikes after filtering with a Gaussian kernel with 200 ms bandwidth. This metric allowed us to directly compare PGBAR with previous analyses Rupprecht et al. (2021). The correlation obtained with PGBAR averaged across cells and datasets is 0.75. As for the other methods, we observed a large variability of performance across recordings, however we did not find a particular condition where our method performed systematically better or worse across indicators (Fig. 4A) and standardized noise level (Fig. 4B), which is a noise index introduced by Rupprecht et al. (2021) to account for different sampling frequencies (ratio between the standard deviation of the normalized fluorescence and the square root of the sampling frequency). In particular, we did not observe a statistically significant difference between mean correlation coefficient across standardized noise levels in the range 0.5-3.5 (Fig. 4B), although lower noise levels seem to be associated to a larger range of correlation.

The performance range of PGBAR across the CASCADE database was comparable to previous methodologies. Among model-based approaches, our method slightly underperformed MLSpike (PGBAR median is within 6% from MLSpike, Mann-Whitney test p-value 0.0043), likely due to its description of the non-linear properties of the calcium indicator. The top performance is achieved by the supervised CASCADE method, however, it is a supervised method (i.e. needs training on ground truth data) and does not provide posterior distributions of spike times and model parameters.

The previous analysis provides an overview of the general performance of PGBAR in comparison with previous methods. To illustrate the advantages of PGBAR in estimating statistical uncertainties, we now take a closer look at a representative recording in the CASCADE dataset (Fig. 5A, CASCADE dataset 9 Chen et al. (2013)), a GCaMP6f fluorescence time series from a pyramidal neuron in the mouse visual cortex. The comparison between ground-truth

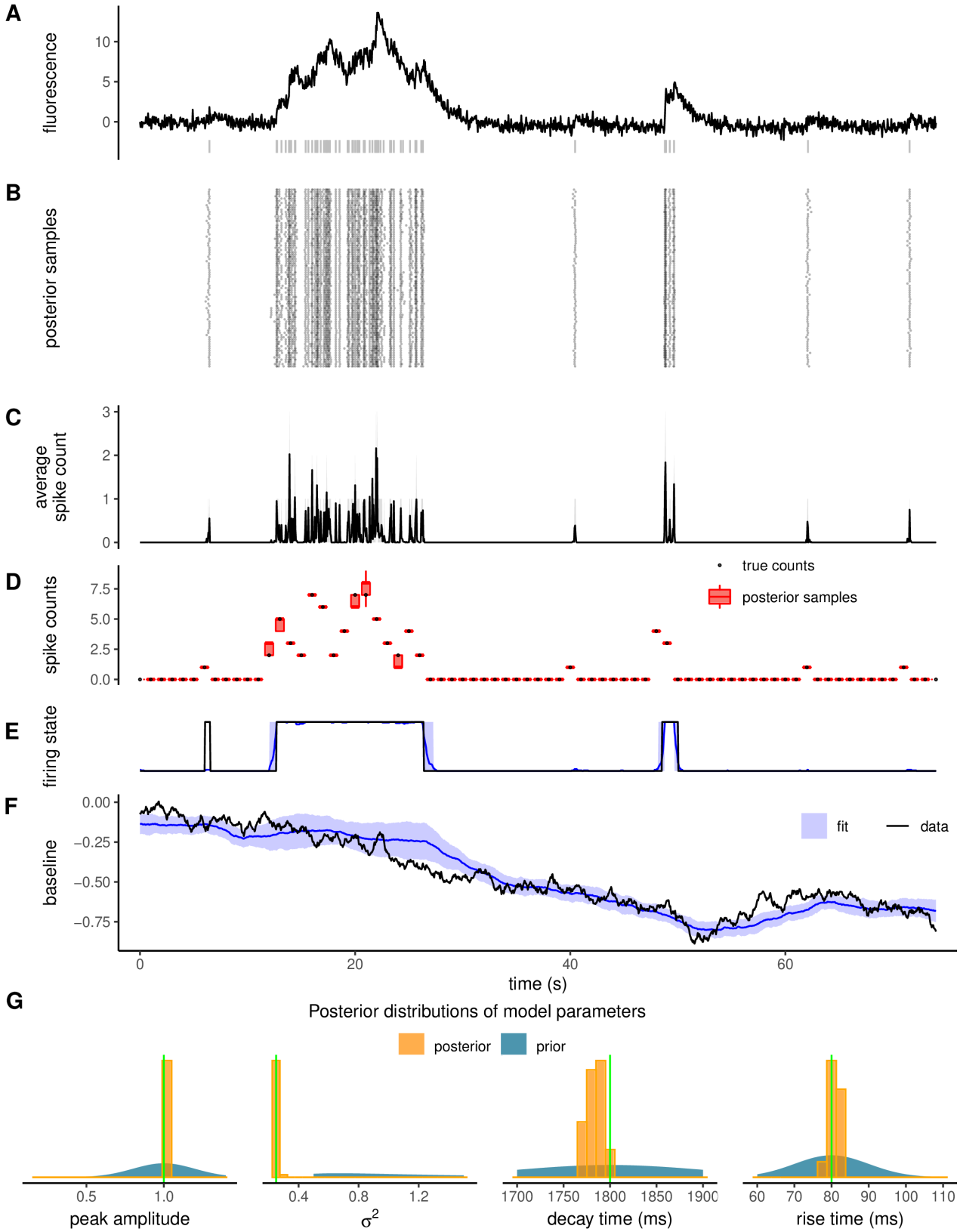


Figure 2: **Validation of the spike inference approach with simulated data. (A)** Example trajectory simulated from the model (solid, black) with ground-truth spike times shown underneath (grey vertical lines). **(B)** Raster plot representing spike times for a thousand Monte Carlo samples. **(C)** Average spike counts over the Monte Carlo samples at each time frame. **(D)** Comparison between ground-truth counts over 1s bins (black dots, from the example trace in **A**) and the corresponding posterior distributions (red boxes). **(E-F)** Comparison of ground-truth firing state and baseline (solid, black) to estimated ones (blue). Shading indicates one standard deviation from posterior averages. **(G)** Posterior distributions of peak response upon a single spike, decay time, rise time, and noise level compared with true value (vertical lines in green).

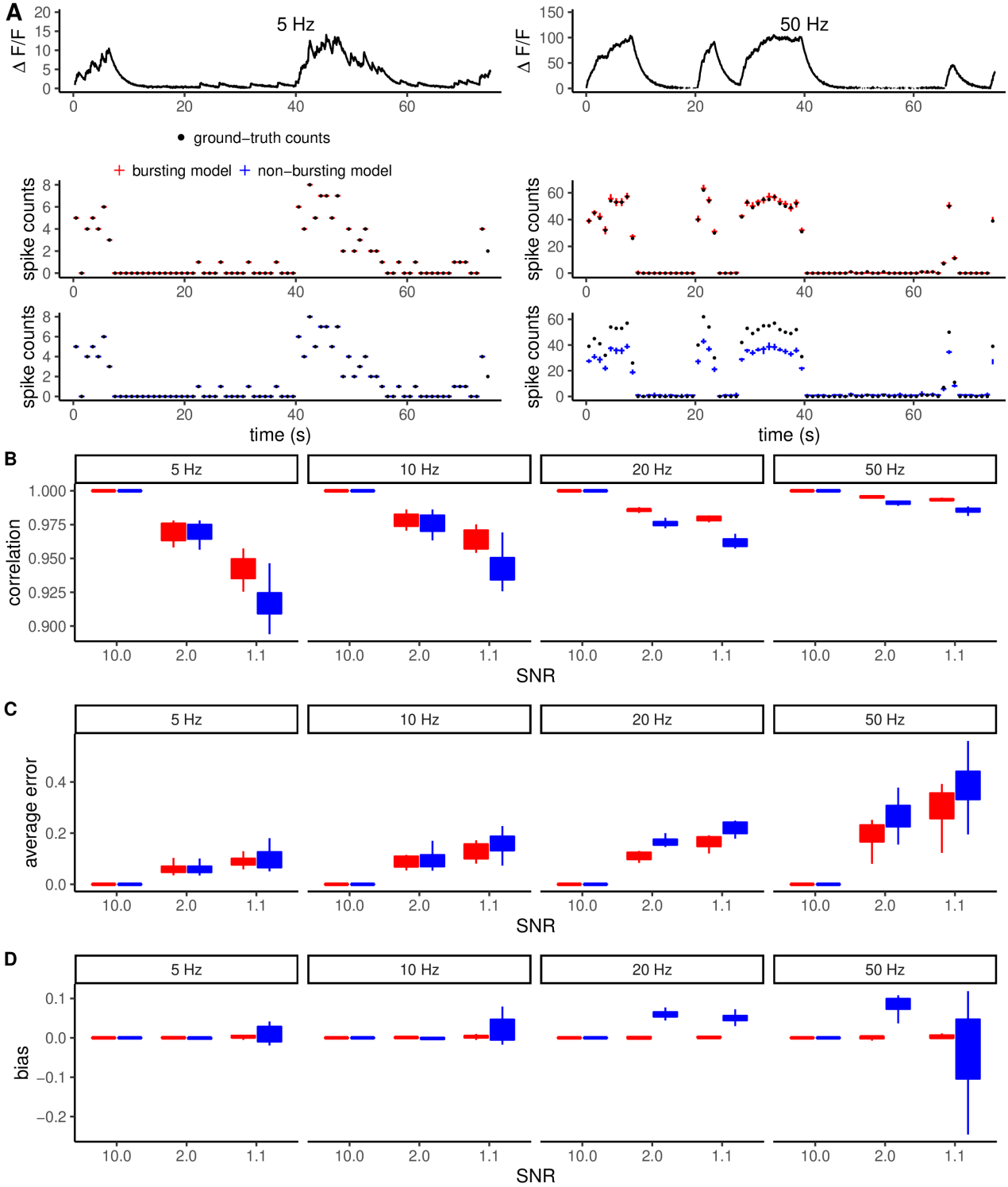


Figure 3: **Dependency of inference performance on noise and firing frequency and the bias of non-bursting models. (A)** Example fluorescence time series simulated at 5Hz and 50Hz bursting frequencies (top). The analysis of these traces using the bursting and non-bursting variant of the model highlights the large bias generated by the non-bursting model at high frequency (bottom). **(B-D)** Quantification of correlation with true spikes, average error, and bias at different levels of SNR and frequency. At increasing firing frequency, the correlation with ground-truth spikes generally increases. This is an effect of calculating correlations at fixed temporal resolution. The average error was quantified as the sum of the absolute deviation from the true spike counts divided by the number of time steps.

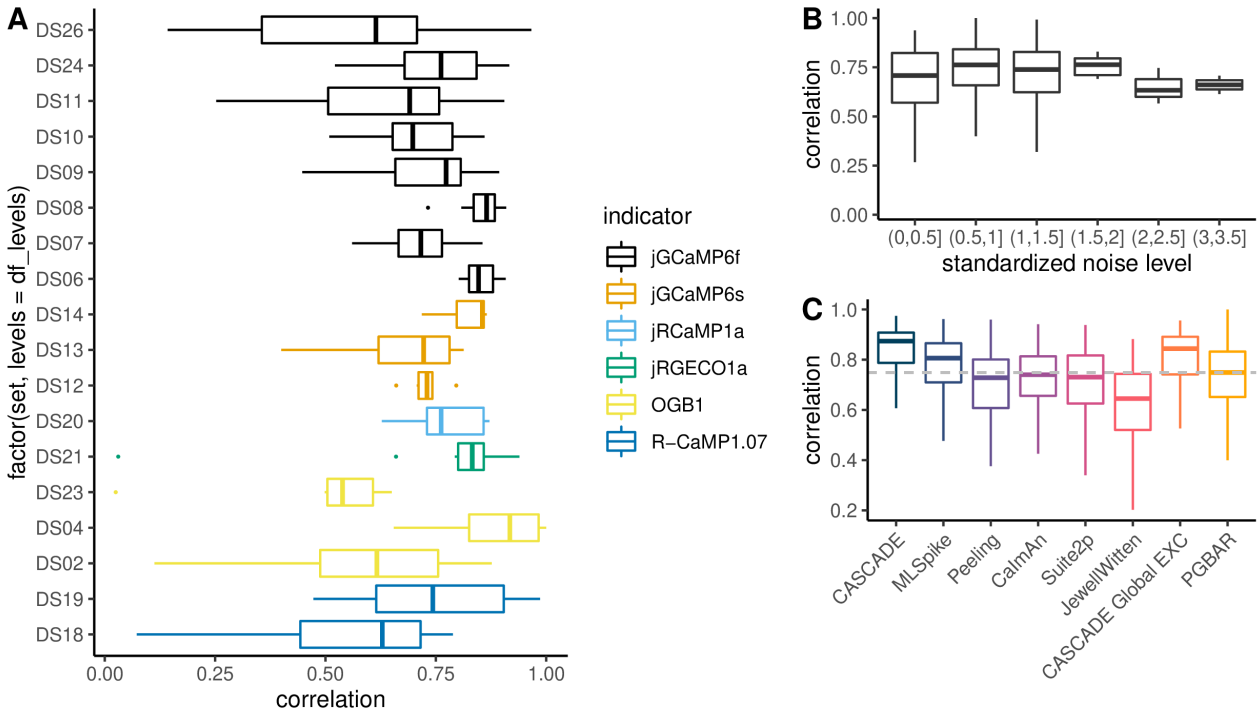


Figure 4: **Comparison of PGBAR with existing methods using analysis of CASCADE benchmark data. (A)** Correlation between estimated and ground-truth firing rates filtered with a 200ms bandwidth Gaussian kernel (CASCADE dataset). The color code represents the different calcium indicators employed in each dataset. **(B)** Correlation with ground-truth spikes as a function of the standardized noise level (Rupprecht et al., 2021). **(C)** Comparison with existing methods. Correlation averaged across datasets and neurons.

spikes and those inferred using PGBAR in Fig. 5A shows differences outside the 1st-3rd interquartile range in 30% of the 1s time intervals. This statistical discrepancy between the posterior distributions and the ground truth can be attributed to the limitations of the autoregressive model to reproduce the biophysical properties of GCaMP6. In spite of such discrepancy, the estimated bursting pattern shown in Fig. 5A (lower panel) captures faithfully the overall periods of increased neuronal activity. The posterior distributions of model parameters are shown in Fig. 5B. Some distributions (burst firing rate and the rise time) shift significantly from their corresponding priors. The mismatch between prior and posterior also highlights the shortcomings of the model. In this particular case, the posterior distribution of the burst firing rate appears shifted to larger values. This result could be associated to the use of Gaussian noise in the model. Indeed, non Gaussian fluorescence fluctuations might be erroneously interpreted by the sampler as true spikes, pushing the burst firing rate to larger values. This example illustrates that prior distributions can be used to partially compensate this bias by penalizing unrealistic parameter configurations.

2.6 Validation and performance of PGBAR on simulations of short-interval stimuli

We tested the accuracy of PGBAR in resolving pairs of stimuli with inter-spike intervals (ISI) below 10ms by performing model simulations of different signal-to-noise levels. Figure 6A shows two simulated fluorescence traces with two spikes 10ms apart at low (1.4) and high (3.4) SNR levels. By running our algorithm on these simulated data PGBAR was accurately detecting 2 spikes with over 95% confidence across all SNR values. Next we focus on the temporal accuracy of the ISI estimation. Increasing SNR or sampling frequency has the effect of narrowing the ISI posterior distribution around the ground-truth value (Fig. 6B). When examining multiple stimulus intervals, SNRs, and sampling frequencies we show that high sampling frequeuncies enable more reliable extraction of spike intervals. Figure 6C shows that the posterior probability of the ISI falls within 3-6ms of the ground-truth ISI for low (1kHz) and high (3kHz) sampling frequencies, respectively (averaged over 30 trials). When this probability is above 0.5, the ground-truth value is equal to the mode of the posterior ISI distribution (using 3ms binsize), therefore we can use this probability as a metric to quantify the temporal accuracy of our inference. For ground-truth values above 3ms, the contours of constant probability (iso-probability) show that increasing the temporal separation between stimuli has no effect on the inference accuracy. On the contrary, increasing the sampling frequency shifts the iso-probability contours towards lower SNR, indicating that for high sampling rates, lower SNRs are required to reach the same accuracy. Figure 6D shows the posterior ISI distributions obtained from ten independent simulations with ground-truth ISI of 5ms at different SNR levels. At low SNR the posterior distributions have larger variance while at higher SNR levels they

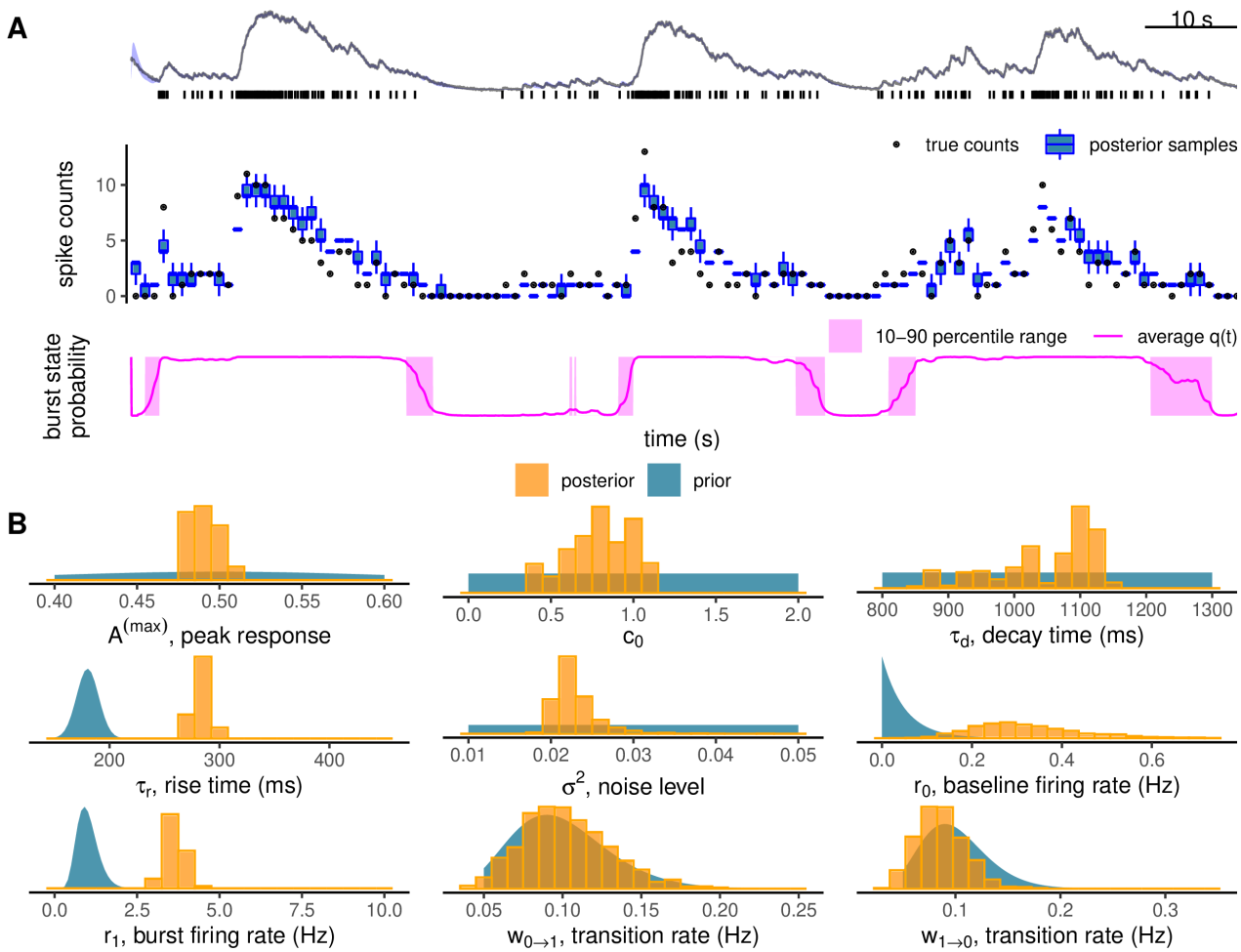


Figure 5: **Analysis of GCaMP6f recordings from the CASCADE dataset. (A)** example $\Delta F/F$ from the CASCADE dataset (#DS09, GCaMP6f, mouse visual cortex) with ground-truth spikes shown underneath fluorescence (top), comparison of spike counts within 1s time intervals (middle) and burst probability (bottom). Shading denotes uncertainty within one standard deviation. **(B)** Comparison between posterior distributions of the model parameters (histograms) and priors (continuous densities): maximal calcium response to single spikes (A_{max}), initial calcium level (c_0), decay and rise time, noise level (σ^2), bursting (r_1) and baseline (r_0) firing rates, transition rates between firing states ($w_{0\to1}, w_{1\to0}$).

shrink around the ground-truth value. This analysis illustrates the variability expected when analyzing multiple trials of the same neuron. We have shown that PGBAR can provide accurate estimates of inter-spike intervals down to 5 ms but the trade-off between temporal accuracy, SNR and sampling frequency must be considered.

2.7 PGBAR spike inference from fluorescence transients recorded using the fast calcium indicator GCaMP8f

We tested our approach on the fast calcium indicator GCaMP8f by performing high-speed (≈2.8kHz) 2-photon linescan calcium imaging of cerebellar granule cells *in vitro*. We used adeno-associated viruses (AAV) to express GCaMP8f in Crus I of the cerebellum (Fig. 7A). Fluorescence was recorded for both granule cell somata and synaptic boutons while ground truth spikes were evoked by extracellular stimulation of granule cell axons in the molecular layer (Fig. 7B).

First, we obtained prior distributions of amplitudes, rise and decay times of single action potential-evoked GCaMP8f fluorescence responses (Fig. 7C) to constrain the inference procedure. Next, we recorded fluorescence responses in a granule cell soma in response to a 20 Hz Poisson stimulation protocol; both a single trial and several trials are shown. (Fig. 7E). We used PGBAR to analyze independently each trial in Figure 7E. By generating thousands of posterior samples of spike time patterns, we obtained the spike probability for all time frames and trials (Fig. 7F). The spike patterns obtained using our method are very similar across trials, showing that PGBAR can reliably detect single-trial action potential-evoked GCaMP8f fluorescence transients.

To illustrate the temporal accuracy of PGBAR we focused on a part of the stimulus train with a short 5 ms interval between two spikes (Fig. 7G). Despite the relatively low SNR ($A^{(max)}/\sigma \approx 2.4$), we could reliably identify the two spikes in each trial. The ground-truth ISI is well within the posterior distribution of each trial, with posterior modes symmetrically distributed around it (Fig. 7H). To better understand the variability of the posterior modes, we simulated 10 fluorescence traces with two spikes separated by 5 ms and sampling frequency matching our recordings. By analysing these simulations with PGBAR we obtained a distribution of ISI posterior modes that is statistically compatible with the real data (Fig. 7I, Mann-Whitney test p-value = 0.97), which provides further evidence that our results are statistically unbiased and that PGBAR can be used to infer spike times for intervals down to 5 ms.

Genetically encoded calcium indicators have been used to monitor neuronal activity in somata and boutons. We examine the accuracy of of PGBAR to infer spike times in 2.8 kHz linescans. Fluorescence transients were evoked using a 20Hz Poisson stimulation protocol containing 29 stimulation events. Each trial was analyzed independently using PGBAR. The distribution of the posterior modes of the total number of spikes across experiments and trials is centered around the ground-truth value for both somas and boutons with a relative error of 15% (Fig. 8A). The time to nearest spike detection, averaged over all stimulation time points, is 0.43 ms (±0.10 ms, SEM) and 1.39 ms (±0.11 ms, SEM) in boutons and somas respectively, highlighting a significant delay of 0.95 ms (±0.15 ms SEM, Mann-Whitney test p-value = 4.8×10^{-7}) from the average time of detection in boutons (Fig. 8B). To examine the temporal accuracy of our approach we calculated the Pearson's correlation between detected spike times and stimulation times over different time-scales. We binned detected spike times and the Poisson stimulation train using bin sizes ranging from 2 ms to 50 ms and calculated their correlation. If we consider as reference correlation the average across the CASCADE dataset (0.75), this analysis shows that the correlation is larger than 0.75 up to 10ms bin size, confirming that our method provides accurate inference of short intervals.

3 Discussion

Fluorescence indicators provide essential tools for monitoring the activity of neuronal populations in model organisms. However, the extraction of underlying firing patterns from fluorescence time series is challenging due to low signal-to-noise ratio, incomplete knowledge of the indicator dynamics, complex firing statistics, and unknown fluorescence modulation. In spite of the proliferation of methodologies developed to address this issue, limited attention has been devoted to the estimation of the statistical uncertainties associated to spike inference. The vast majority of the spike detection algorithms are indeed based on optimization techniques, providing only point estimates of the detected spikes. Quantifying statistical uncertainties is key to compare firing patterns across neurons (Diana et al., 2019) and establishing their causal relationships. The work of Pnevmatikakis et al. (2013) and Vogelstein et al. (2009), addressed this issue by using Monte Carlo methods to approximate the full posterior distributions of spiking patterns. Building upon their work, here we improve the generative model used to infer spiking patterns from fluorescence time series and describe efficient Monte Carlo strategies to infer spike times and their statistical uncertainties.

Bursting dynamics and baseline modulation. Neural activity patterns are not always well described by a simple Poisson spiking process. The PGBAR is the first Monte Carlo method (Greenberg et al., 2018; Pnevmatikakis et al., 2013; Vogelstein et al., 2009) to perform statistical inference based on a non-homogeneous Poisson firing and baseline modulation model. PGBAR uses a two-state process to enable transitions between low and high firing rates. This feature is used to mimic the alternation between periods of low baseline firing and bursting activity transients where the firing rate is significantly increased. We have shown that not taking into account bursting activity in the model used for inference can lead to biased results, especially at low SNR levels and high firing frequency. By explicitly

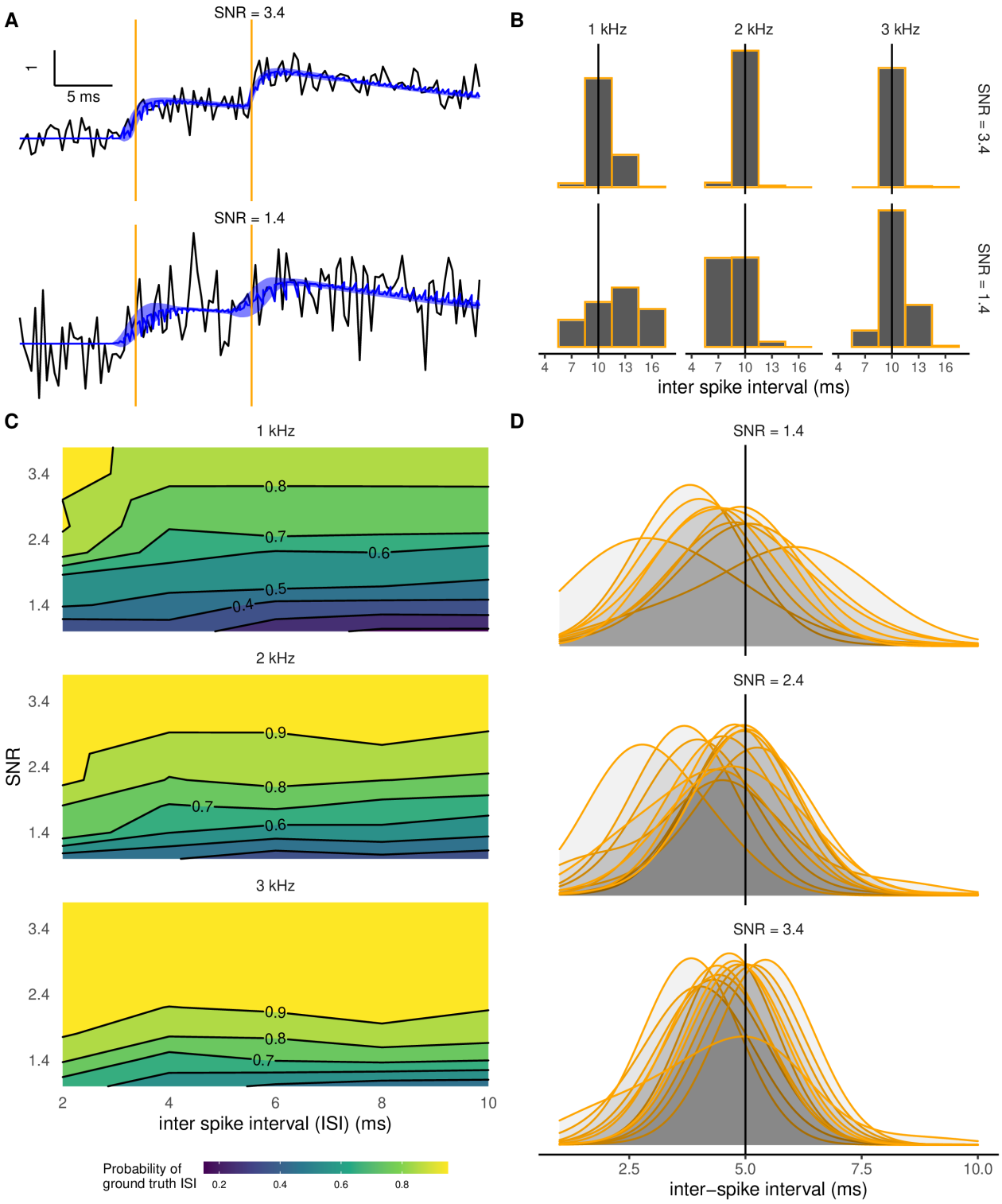


Figure 6: **Sensitivity of spike detection to sampling frequency and SNR level. (A)** Examples of simulated fluorescence traces with two spikes separated by 10ms (vertical lines) at low SNR (1.4, bottom) and high SNR (3.4, top). Shaded bands display denoised fits (calcium fluorescence plus baseline) within one standard deviation. (**B**) Posterior distributions of the inter-spike interval (ISI). Increasing SNR (from bottom to top) and sampling frequency (left to right) has the effect of reducing the ISI posterior variance, bringing the maximum-a-posteriori estimate (MAP) closer to the ground-truth. **(C)** Posterior probability of the ISI to be within an interval of 3ms centered around the ground truth ISI as a function of SNR and ground-truth ISI with sampling frequency of 1, 2 and 3kHz. **(D)** Trial-to-trial variability of ISI posterior distributions. We analyzed 12 simulated fluorescence traces with sampling frequency of 3kHz and two spikes separated by 5ms. Density plots have been smoothed with 1ms bandwith. For all simulations we used $\tau_r = 3.7$ms and $\tau_d = 40$ms.

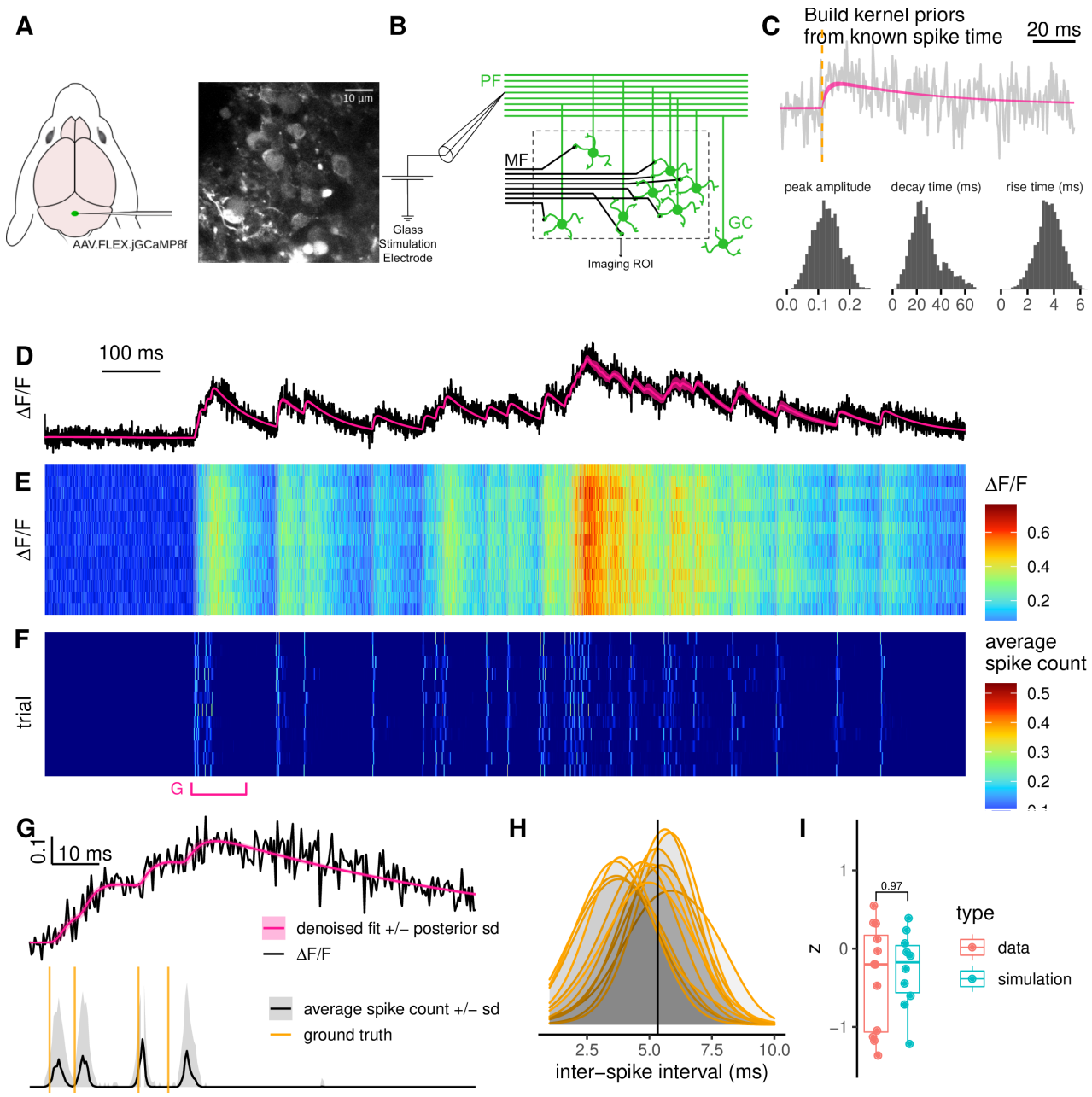


Figure 7: **High-speed 2-photon linescan calcium imaging.** **(A)** GCaMP8f virus injection in the cerebellar vermis. **(B)** Induction of action potentials to cerebellar granule cells by direct stimulation of the parallel fibers. **(C)** Single pulse stimulation to extract kinetic model parameters of GCaMP8f indicator. **(D-E)** High-speed (3kHz) 2-photon linescan calcium imaging of granule cell somata. **(D)** Representative fluorescence time series from a single trial and **(E)** heatmap showing fluorescence transients evoked using the same Poisson train across trials. Single-trial fluorescence (D) and denoised fit (calcium level plus baseline). **(F)** Spike detection for each trial. **(G)** 100 ms time window highlighting the first four stimulation-induced action potentials. Normalized fluorescence and denoised fit (top), average spike count (bottom). Orange vertical lines denote stimulation time points. **(H)** Comparison of the posterior distributions of the interval between the first two detected spikes across experimental trials. The solid vertical line at 5.3 ms denotes the time interval between the first two stimulations. **(I)** Comparison of the posterior modes of the inter-spike interval across trials between real data and simulations.

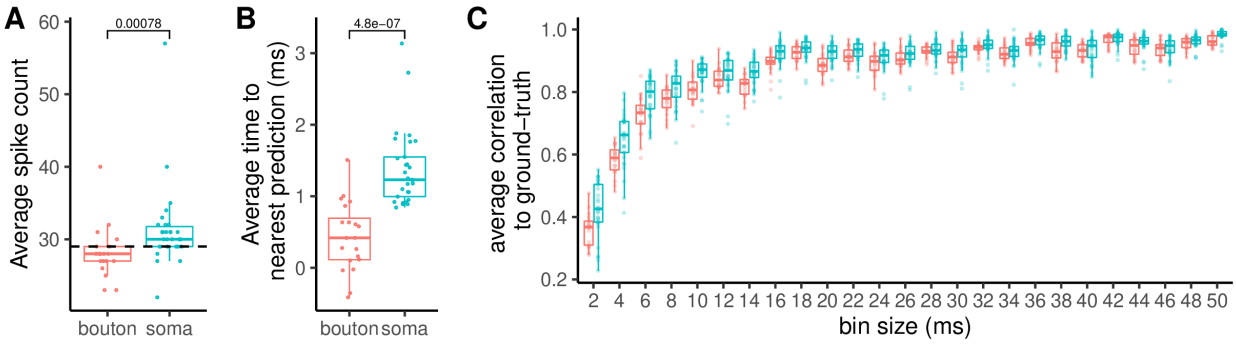


Figure 8: **Temporal resolution analysis across somatic and bouton recordings. (A)** Average spike count estimated from 6 somata and 4 boutons (5 trials each). **(B)** Time from stimulus time predicted spike averaged across all 29 stimulations per trial. The comparison between somas and boutons shows that somatic transients are delayed by 1 ms. **(C)** Correlation between predicted spikes and stimulation events across time scales.

modeling the bursting dynamics, PGBAR produced unbiased results at all levels of noise and frequency when tested on simulated data (Figure 3).

Although the model used in Pnevmatikakis et al. (2013) did not account for fluorescence baseline modulation, they acknowledged that it would be important to account for baseline fluctuations in *in vivo* recordings. PGBAR uses a Gaussian random walk, as MLSpike Deneux et al. (2016)), to describe slow changes in baseline fluorescence across the recording. While this is a simple Markov model of fluorescence baseline it introduces additional noise. To avoid this effect, it is possible to employ alternative baseline models, such as the integrated random walk.

Joint estimation of time-independent parameters and dynamic variables. The estimation of time-independent model parameters is a well known issue in spike detection algorithms, requiring ad-hoc calibration procedures and manual settings. In particular, unknown firing rate and peak amplitude (in response to a single action potential) can lead to unidentifiable parameters. PGBAR employs a fully Bayesian approach where model parameters and dynamic variables are treated equally. This enables users to constrain the detection of action potential by controlling mean and variance of phenomenological parameters (e.g. rise and decay constants, firing rates, bursting frequencies) by constraining their prior distributions. We employed for the first time state-of-the-art particle Gibbs algorithms to infer spikes from noisy fluorescence. This is a key novelty compared to previous SMC-based methods (Greenberg et al., 2018; Vogelstein et al., 2009), allowing for a joint estimation of time-independent parameters and dynamic variables.

Comparison with benchmark datasets. The proliferation of spike inference methodologies led to the development of community-based initiatives (Berens et al., 2018; Theis et al., 2016) to rank the performance of available methods. We applied our approach on the CASCADE dataset (Rupprecht et al., 2021) which provides a curated database of neuronal recordings from mouse and zebrafish using different calcium indicators. The performance of PGBAR, measured by the correlation coefficient with ground-truth spikes, is within one quartile from the correlation distributions of existing unsupervised approaches. In addition, it provides information about the statistical uncertainty associated to spike detection that is not currently possible with state-of-the-art techniques.

PGBAR detection of short high-frequency bursts using an ultrafast calcium indicator PGBAR employs a second-order autoregressive process to link spiking activity to fluorescence. This simple model accounts for the basic qualitative aspects of calcium transients, and it is well-suited for linear indicators. For this reason, we have tested the performance of PGBAR on the ultrafast GCaMP8f (Zhang et al., 2021) that exhibits improved linearity in comparison to previous calcium probes. We showed that the combination of PGBAR with the GCaMP8f enables the detection of inter-spike intervals of 5 ms with an accuracy of 2.5 ms from single trials, thus offering a statistical tool for estimating high-frequency neural activity patterns.

PGBAR limitations and future perspectives. Although full Bayesian inference is known to be computationally expensive, SMC algorithms are highly parallelizable. Posterior distributions are represented by particles that are simultaneously propagated through time. In particular, GPU parallelization of SMC methods is an active field of research in computational statistics. Future advances in this directions might dramatically boost these methods and offer tools for online processing of fluorescence time series.

Many commonly used indicators exhibit a nonlinear fluorescence response to increases in intracellular calcium (Chen et al., 2013). The autoregressive model used in this study does not account for that nonlinearity. This work provides a statistical framework generalizable to more specific biophysical models of calcium indicators to account for non-linear effects. In addition, these models depend on rate constants that are difficult to measure directly. Not being able to constrain model parameters is an issue for spike inference as different parameter combinations might describe equally well the observed fluorescence by adjustments of the inferred spike pattern. For this reason, when

model parameters are not identifiable, the problem of spike inference from fluorescence time series is ill-defined. Our Bayesian framework offers a systematic approach to address this issue by exploiting current and future data on the kinetics of calcium indicators through informative priors. Thanks to our joint sampling strategy of model parameters and spike patterns, our Monte Carlo method can keep the inference within biophysically relevant regions.

4 Materials and Methods

4.1 Particle Gibbs with ancestor sampling

The PGAS step used in Algorithm 1 to sample latent state trajectories was introduced in Lindsten et al. (2014) to improve the performance of the original particle Gibbs sampler (Andrieu et al., 2010). We refer to their original works for details on convergence and mixing properties of the method. The PGAS step is a SMC algorithm that generate a new latent state trajectory starting from a reference trajectory and the model parameters. We initialize the algorithm with a set of N latent states (particles) at time $t = 1$,

$$X_1^{(i)} = \{q_1^{(i)}, s_1^{(i)}, C_1^{(i)}, b_1^{(i)}\} \quad i = 1, \cdots, N \tag{14}$$

where the first $N-1$ are sampled from a proposal distribution equal to the probability of the initial state $\mu^\theta(X_1)$

$$\mu^\theta(X_1) = \rho_1(X_1) = \frac{(r_{q_1}\Delta)^{s_1}}{s_1!} e^{-r_{q_1}\Delta} \cdot \frac{e^{-b_1^2/2}}{\sqrt{2\pi}} \delta^2(C_1 - \hat{C}_1) \tag{15}$$

where bursting and baseline firing states $q_1 = 0, 1$ have equal probability and the calcium vector is constrained by the initial condition of Eq. (5), $\hat{C}_1 \equiv (c_0 + AS_1, 0)$ and the model parameter c_0. The last particle is constrained by the reference trajectory X'. To conclude the initialization stage, we assign importance weights $w_1^{(i)}$ to all particles

$$w_1^{(i)} = \mu^\theta(X_1^{(i)}) g^\theta(F_1|X_1^{(i)})/\rho_1(X^{(i)}), \quad i = 1, \cdots, N. \tag{16}$$

Next, for $t > 1$, we evolve the particle system through time by assigning ancestor particles $\{\tilde{X}_{t-1}^{(i)}\}_{i=1}^N$, propagate these to time t to get a new set of particles $\{X_{1:t}^{(i)}\}_{i=1}^N$ and assigning weights $w_t^{(i)}$. For the first $N-1$ particles the ancestors are obtained by multinomial resampling from the particle system at time $t-1$ with probability proportional to the importance weights $w_{t-1}^{(i)}$. The ancestor J of the last particle is drawn from the distribution

$$\mathbb{P}(J = i) = \frac{w_{t-1}^{(i)} f_t^\theta(X_t'|X_{t-1}^{(i)})}{\sum_{k=1}^N w_{t-1}^{(k)} f_t^\theta(X_t'|X_{t-1}^{(k)})} \tag{17}$$

exploiting the fact that we know its state at time t.

In order to evolve the particle system through time we need to set a proposal distribution $\rho_t(X_t|X_{t-1})$ to sample new latent states. This proposal is then taken into account in the reweighting stage. Although the choice of the proposal distribution is arbitrary, it can be shown that the conditional distribution $P(X_t|X_{t-1}, F_t)$ reduces the variance of the importance weights. We can express this optimal proposal as

$$P(X_t|F_t, X_{t-1}) = \frac{P(X_t, F_t|X_{t-1})}{\int_{X_t} P(X_t, F_t|X_{t-1})} = \frac{f_t^\theta(X_t|X_{t-1}) g_t^\theta(F_t|X_t)}{Z^\theta(X_{t-1}, F_t)} \tag{18}$$

where $Z^\theta(X_{t-1}, F_t)$ is the normalization factor as a function of the latent state at time $t-1$ and current fluorescence F_t. We can now use the expressions of f_t^θ and g_t^θ in Eqs. (11,12) for our model to compute the optimal proposal distribution. To do so we decompose $P(X_t|X_{t-1}, F_t)$ as the product

$$P(X_t = \{q_t, s_t, C_t, b_t\}|X_{t-1}, F_t) = P(q_t, s_t|X_{t-1}, F_t) \cdot P(C_t|C_{t-1}, s_t) \cdot P(b_t|q_t, s_t, C_t, X_{t-1}, F_t) \tag{19}$$

where we used the fact that C_t is deterministic and only depends on C_{t-1} and the spike count at time t. The idea is to use this chain decomposition to sample first the firing state q_t and the spike count s_t, then calculate C_t from its deterministic evolution and finally sample the baseline b_t from its distribution conditional to the other variables. The first term $P(q_t, s_t|X_{t-1})$ can be obtained by integrating the product $f_t^\theta(X_t|X_{t-1}) g_t^\theta(F_t|X_t)$ over b_t and C_t and then normalizing the result. The integration over b_t and C_t leads to

$$\begin{aligned}
\int db_t dC_t\, f_t^\theta(X_t|X_{t-1}) g_t^\theta(F_t|X_t) &= \int db_t dC_t \delta^{(2)}(C_t - M \cdot C_{t-1} - AS_t) \cdot W_{q_{t-1}q_t} \frac{(r_{q_t}\Delta)^{s_t}}{s_t!} e^{-r_{q_t}\Delta} \cdot \\
&\cdot (2\pi\Delta\sigma_b^2)^{-1/2} \exp\left(-\frac{1}{2\Delta\sigma_b^2}(b_t - b_{t-1})^2\right) \cdot (2\pi\sigma^2)^{-1/2} \cdot \exp\left[-\frac{1}{2\sigma^2}(F_t - c_t - b_t)^2\right] \\
&= W_{q_{t-1}q_t} \frac{(r_{q_t}\Delta)^{s_t}}{s_t!} e^{-r_{q_t}\Delta} \cdot I(b_{t-1}, F_t - c_t, \Delta\sigma_b^2, \sigma^2)
\end{aligned} \tag{20}$$

where we introduced the function $I(y_1, y_2, \sigma_1^2, \sigma_2^2)$ as the integral

$$I(y_1, y_2, \sigma_1^2, \sigma_2^2) = (2\pi\sigma_1^2)^{-1/2}(2\pi\sigma_2^2)^{-1/2} \int dx e^{-\frac{(x-y_1)^2}{2\sigma_1^2} - \frac{(x-y_2)^2}{2\sigma_2^2}} = \frac{\exp\left[-\frac{1}{2}\frac{(y_1-y_2)^2}{\sigma_1^2+\sigma_2^2}\right]}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \tag{21}$$

The normalization factor $Z^\theta(X_{t-1}, F_t)$ is obtained by taking the sum of Eq. (20) over firing state and spike count:

$$Z^\theta(X_{t-1}, F_t) = \sum_{q' \in \{0,1\}} \sum_{s'=0}^{\infty} W_{q_{t-1}q'} \frac{(r_{q'}\Delta)^{s'}}{s'!} e^{-r_{q'}\Delta} \cdot I(b_{t-1}, F_t - c_t, \Delta\sigma_b^2, \sigma^2). \tag{22}$$

To draw a combination of q_t and s_t from this distribution we applied a cutoff to the number of spikes per time step $S^{(max)} = 20$ and constructed a probability matrix of size $2 \times S^{(max)}$ for all combinations of firing state and spike count.

To obtain the full conditional distribution of b_t we consider again the product $f_t^\theta(X_t|X_{t-1})g_t^\theta(F_t|X_t)$ and by keeping only terms in b_t and normalizing we obtained a Gaussian distribution with mean μ_{prop} and variance σ_{prop} given by

$$\mu_{prop} = \frac{b_{t-1}\sigma^2 + (F_t - c_t)\sigma_b^2\Delta}{\sigma^2 + \sigma_b^2\Delta} \tag{23}$$

$$\sigma_{prop}^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_b^2\Delta}\right)^{-1}. \tag{24}$$

The final step is to reweight all particles according using the importance weight

$$w_t^{(i)} = f_t^\theta(X_t^{(i)}|\tilde{X}_{t-1}^{(i)})g_t^\theta(F_t|X_t^{(i)})/\rho_t(X_t^{(i)}|\tilde{X}_{t-1}^{(i)}) \tag{25}$$

However, due to the form of the optimal proposal in Eq. (18) the importance weights reduce to the normalization factor calculated in Eq. (22)

$$w^{(i)} = Z^\theta(X_{t-1}^{(i)}, F_t) \tag{26}$$

Algorithm 2 PGAS kernel

Input: reference trajectory $X'_{1:T}$, and model parameters θ

Draw $X_1^{(i)}$ from the poposal distribution ρ_1 for $i = 1, \cdots, N-1$
Set $X_1^{(N)} = X'_1$
Set importance weights $w_1^{(n)} = \mu^\theta(X_1^{(n)})g^\theta(F_1|X_1^{(n)})/\rho_1(X^{(n)})$ for $n = 1, \cdots, N$
for t in $2{:}T$ **do**
  // Resampling and ancestor sampling
  Resample $N-1$ particles $\{\tilde{X}_{1:t-1}^{(i)}\}_{i=1}^{N-1}$ with probabilities proportional to the importance weights $\{w_{t-1}^{(i)}\}_{i=1}^{N}$
  Draw J with probability $\mathbb{P}(J = i) \propto w_{t-1}^{(i)} f_t^\theta(X'_t|X_{t-1}^{(i)})$ and set $\tilde{X}_{1:t-1}^{(N)} = X_{1:t-1}^{(J)}$
  // Particle propagation
  Draw $X_t^{(i)}$ from the proposal distribution $\rho_t(X_t|\tilde{X}_{t-1}^{(i)})$ for $i = 1, \cdots, N-1$
  Set $X_t^{(N)} = X'_t$
  Set $X_{1:t}^{(i)} = (\tilde{X}_{1:t-1}^{(i)}, X_t^{(i)})$ for $i = 1, \cdots, N$
  // Weighting
  Set $w_t^{(i)} = f_t^\theta(X_t^{(i)}|\tilde{X}_{t-1}^{(i)})g_t^\theta(F_t|X_t^{(i)})/\rho_t(X_t^{(i)}|\tilde{X}_{t-1}^{(i)})$ for $i = 1, \cdots, N$
Draw k with $\mathbb{P}(k = i) \propto w_T^{(i)}$

Output: $X_{1:T}^{(k)}$

4.2 Prior distributions

As discussed in the text, we use a reparameterization of the autoregressive model in terms of the maximal amplitude $A^{(max)}$, rise and decay times τ_r and τ_d, for which it is easier to design realistic prior distributions based on previous empirical estimates of the kinetics of calcium indicators. We have used truncated normal priors for the maximal amplitude, the initial condition of the autoregressive model c_0, rise and decay time. In order to calculate the full conditional distributions on bursting/baseline firing rates $r_{0,1}$ and the transition matrix parameters $w_{q\to q'}$ we have used a gamma distribution, whereas for the noise level σ^2 an inverse gamma distribution.

4.3 Sampling rules for time-independent parameters

In Algorithm 1, after a new latent state trajectory is sampled from the PGAS kernel, we draw time-independent model parameters from the conditional distribution $P(\theta_i|X_{1:T}, F_{1:T})$. We use a mixed approach where the parameters $r_{0,1}$, $w_{q\to q'}$ and σ^2 are sampled from their full conditional distribution, which can be obtained analytically by using gamma priors, while kernel parameters, $A^{(max)}$ and $\tau_{r,d}$, are sampled using the Metropolis-Hastings acceptance rule.

The full conditional distribution of a given parameter can be obtained from the joint probability of model parameters, latent state and fluorescence trajectories

$$P(\theta) \cdot P_\theta(X_{1:T}, F_{1:T}|\theta) \tag{27}$$

where $P(\theta)$ is the prior distribution. We will now calculate the full conditionals for firing rates r_q, transition parameters $w_{q\to q'}$ and noise variance σ^2. For simplicity we will use the same symbols for shape, α, and rate, β of all prior distributions, although they differ numerically for each parameter. By combining the expressions in Eqs.(10), (11) and (12) with the gamma prior gamma(α, β) and by keeping only terms proportional to $r_{0,1}$ we obtain

$$P(r_q|\cdots) \propto r_q^{\alpha-1} e^{-\beta r_q - r_q \Delta T} r_q^{\sum_{t:q_t=q} s_t} \tag{28}$$

therefore the full conditional is a gamma distribution with updated parameters

$$\alpha' = \alpha + \sum_{t:q_t=q} s_t \tag{29}$$

$$\beta' = \beta + \Delta T \tag{30}$$

By applying the same method to the transition rates $w_{q\to q'}$ we have

$$P(w_{q\to q'}|\cdots) \propto w_{q\to q'}^{\alpha-1} e^{-\beta w_{q\to q'}} w_{q\to q'}^{N_{qq'}} (1 - \Delta w_{q\to q'})^{N_{qq}} \approx w_{q\to q'}^{\alpha+N_{qq'}-1} e^{-(\beta+\Delta N_{qq}) w_{q\to q'}} \tag{31}$$

where the approximation holds when the transition rate between firing states is much slower than the sampling frequency ($w_{q\to q'} \ll \Delta^{-1}$). Therefore the full conditional is again a gamma distribution with parameters

$$\alpha' = \alpha + N_{qq'} \tag{32}$$

$$\beta' = \beta + \Delta N_{qq} \tag{33}$$

For the noise variance parameter we used an inverse gamma prior and by applying the same method we can compute the full conditional as

$$P(\sigma^2|\cdots) \propto (\sigma^2)^{-\alpha-1-T/2} \exp\left(-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}\sum_t (F_t - c_t - b_t)^2\right) \tag{34}$$

therefore the updated shape and rate of the inverse gamma are

$$\alpha' = \alpha + T/2 \tag{35}$$

$$\beta' = \beta + \frac{1}{2}\sum_t (F_t - c_t - b_t)^2 \tag{36}$$

4.4 Response kernel

The response to a single spike can be obtained by writing Eq. (3) in the form of a first order Markov process in terms of the new variables $C_t \equiv [c_t, c_{t-1}]$ and $S_t \equiv [s_t, 0]$ so that

$$C_t = M \cdot C_{t-1} + AS_t, \quad M = \begin{bmatrix} \gamma_1 & \gamma_2 \\ 1 & 0 \end{bmatrix} \tag{37}$$

with the initial condition $C_1 = [c_0 + AS_1, 0]$. We can now write the solution at time t as

$$C_t = M^{t-1} C_1 + A \sum_{k=2}^{t} M^{t-k} S_k. \tag{38}$$

If $s_t = \delta_{t,1}$ and $c_0 = 0$ then Eq. (38) simplifies to

$$C_t = AM^{t-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{39}$$

By introducing eigenvectors and eigenvalues of M

$$\gamma_\pm \equiv \frac{\gamma_1 \pm \sqrt{\gamma_1^2 + 4\gamma_2}}{2}, \quad e_\pm \equiv \begin{bmatrix} \gamma_\pm \\ 1 \end{bmatrix} \tag{40}$$

we can express Eq. (39) as

$$C_t = AM^{t-1}\left(\frac{e_+ - e_-}{\gamma_+ - \gamma_-}\right) = A\left(\frac{\gamma_+^{t-1}e_+ - \gamma_-^{t-1}e_-}{\gamma_+ - \gamma_-}\right) \tag{41}$$

therefore we have

$$c_t = A\left(\frac{\gamma_+^t - \gamma_-^t}{\gamma_+ - \gamma_-}\right). \tag{42}$$

By setting the time derivative of c_t to zero we obtain the time to reach the maximal response τ_r as

$$\tau_r = \frac{\log\left(\frac{g_+}{g_-}\right)}{g_- - g_+}, \quad g_\pm = \log\gamma_\pm \tag{43}$$

whereas if we take the long time limit of Eq. (39) we obtain

$$c_t = A\gamma_+^t[1 - (\gamma_-/\gamma_+)^t] \approx Ae^{t/\tau_d}, \quad \tau_d = -\frac{1}{g_+} \tag{44}$$

4.5 Reparameterization

To reparameterize the autoregressive model in terms of kinetic parameters we need to find the inverse map $\tau_{r,d} \to \gamma_{1,2}$ to obtain the original autoregressive parameters given rise and decay times. By combining the expressions of τ_r and τ_d in terms of $g_\pm$ we have

$$\frac{\tau_r}{\tau_d} = \frac{\log\frac{g_-}{g_+}}{\frac{g_-}{g_+} - 1} \tag{45}$$

which shows that the ratio g_-/g_+ can be expressed as

$$\frac{g_-}{g_+} = f^{-1}\left(\frac{\tau_r}{\tau_d}\right), \quad f(x) = \frac{\log(x)}{x-1} \tag{46}$$

where the inverse function $f^{-1}(x)$ can be determined by numerical interpolation in the range [0,1]. To obtain the original autoregressive parameteres $\gamma_{1,2}$ we first obtain $g_\pm$ as

$$g_+ = -\frac{1}{\tau_d} \tag{47}$$

$$g_- = g_+ \cdot f^{-1}\left(\frac{\tau_r}{\tau_d}\right) \tag{48}$$

and then

$$\gamma_1 = e^{g_+} + e^{g_-} \tag{49}$$

$$\gamma_2 = -e^{g_+ + g_-} \tag{50}$$

4.6 Experimental methods

GCaMP8f virus injection in the Crus I lobule of the cerebellum. Virus injection was targeted to Crus I (6.00 mm posterior to the Bregma, 3.00 mm lateral to the midline; vertical depth of 0.50 mm from pial surface) under deep isofluorane anesthesia. 100 nl of adeno-associated virus encoding GCaMP8f (AAV-DJ-CAG-FLEx-jGCaMP8f, Janelia Research Campus) was injected with Nanoject III (Drummond Scientific) using thin glass pipette (diameter 30 μm). Injection was made when the C57BL/6J-Gabra6tm2(cre)Wwis mice were 7 weeks old. After injections the mice were returned to their home cage for 6 weeks to allow time for expression.

Slice Preparation. Acute coronal slices (200 μm) of the Crus I lobule of the cerebellum was prepared from adult C57BL/6J-Gabra6tm2(cre)Wwis mice, aged 97 and 102 days. Following transcardial perfusion with an ice-cold solution containing (in mM): 2.5 KCl, 0.5 CaCl2, 4 MgCl2, 1.25 NaH2PO4, 24 NaHCO3, 25 glucose, 230 sucrose, and 0.5 ascorbic acid, the brains were removed and placed in the same solution. The solution was bubbled with 95% O2 and 5% CO2. Slices were cut from the dissected lateral cerebellum using a vibratome (Leica VT1200S), and incubated at room temperature for 30 minutes in a solution containing (in mM): 85 NaCl, 2.5 KCl, 0.5 CaCl2, 4 MgCl2, 1.25

NaH2PO4, 24 NaHCO3, 25 glucose, 75 sucrose and 0.5 ascorbic acid. Slices were then transferred to an external recording solution containing (in mM): 125 NaCl, 2.5 KCl, 1.5 CaCl2, 1.5 MgCl2, 1.25 NaH2PO4, 24 NaHCO3, 25 glucose and 0.5 ascorbic acid, and maintained at room temperature for up to 6 hr.

Cellular Imaging. The brain slices containing GCaMP8f expressing cells were identified with a 4x objective lens (Olympus UplanFI 4x, 0.13 NA) using very brief illumination with 470 nm light to excite GcaMP8f fluorescence. GCs were identified using infrared Dodt-gradient contrast and a QlClick digital CCD camera (QImaging, Surrey, BC, Canada) mounted on an Ultima multiphoton microscopy system (Bruker Nano Surfaces Division, Middleton, WI, USA) that was mounted on an Olympus BX61W1 microscope, equipped with a water-immersion objective (Olympus 60x, 1.1 NA). Two-photon excitation was performed with a Ti-sapphire laser (Spectraphysics). To visualize GCs expressing GcaMP8f, two-photon excitation was performed at 920 nm. Infrared Dodt-gradient contrast was used to position the stimulation pipette in molecular layer targeting PF to directly activate GC bodies. Linescan imaging of GC bodies was performed by scanning through the whole cell membrane marked by freehand linescan mode (Prairie View). Total laser illumination per sweep lasted 2000 ms. Fluorescence was detected using both proximal epifluorescence and substage photomultiplier tube gallium arsenide phosphide (H7422PA-40 SEL, Hamamatsu).

5 Code and Data

In this work we employed the publicly available CASCADE dataset in Rupprecht et al. (2021). Linescan imaging data and the source code code of PGBAR are available through the GitHub repository https://github.com/giovannidiana/pgbar.

6 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 896051. This work was also supported by a Pasteur-Roux-Cantarini fellowship of the Institut Pasteur. We would like to thank Peter Rupprecht for his help with the analysis of the CASCADE dataset and for thorough discussions. We also thank Nicolas Chopin for suggesting the use of backward steps methods in particle Gibbs, Andrea Giovannucci, Marco Banterle and Diana Passaro for helpful discussions.

References

Ahrens MB, Orger MB, Robson DN, Li JM, Keller PJ. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nature methods. 2013; 10(5):413–420.

Andrieu C, Doucet A, Holenstein R. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(3):269–342.

Berens P, Freeman J, Deneux T, Chenkov N, McColgan T, Speiser A, Macke JH, Turaga SC, Mineault P, Rupprecht P, et al. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. PLoS computational biology. 2018; 14(5):e1006157.

Chen TW, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature. 2013; 499(7458):295–300.

Chopin N, Papaspiliopoulos O. An Introduction to Sequential Monte Carlo. Springer-Verlag; 2020.

Deneux T, Kaszas A, Szalay G, Katona G, Lakner T, Grinvald A, Rózsa B, Vanzetta I. Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. Nature communications. 2016; 7(1):1–17.

Diana G, Sainsbury TT, Meyer MP. Bayesian inference of neuronal assemblies. PLoS computational biology. 2019; 15(10):e1007481.

Dyer EL, Studer C, Robinson JT, Baraniuk RG. A robust and efficient method to recover neural events from noisy and corrupted data. In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* IEEE; 2013. p. 593–596.

Fletcher AK, Rangan S. Scalable inference for neuronal connectivity from calcium imaging. arXiv preprint arXiv:14090289. 2014; .

Friedrich J, Paninski L. Fast active set methods for online spike inference from calcium imaging. Advances In Neural Information Processing Systems. 2016; 29:1984–1992.

Friedrich J, Zhou P, Paninski L. Fast online deconvolution of calcium imaging data. PLoS computational biology. 2017; 13(3):e1005423.

Giovannucci A, Friedrich J, Gunn P, Kalfon J, Brown BL, Koay SA, Taxidis J, Najafi F, Gauthier JL, Zhou P, et al. CaImAn an open source tool for scalable calcium imaging data analysis. Elife. 2019; 8:e38173.

Greenberg DS, Wallace DJ, Voit KM, Wuertenberger S, Czubayko U, Monsees A, Handa T, Vogelstein JT, Seifert R, Groemping Y, et al. Accurate action potential inference from a calcium sensor protein through biophysical modeling. BioRxiv. 2018; p. 479055.

Grewe BF, Langer D, Kasper H, Kampa BM, Helmchen F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. Nature methods. 2010; 7(5):399–405.

Hoang H, Sato Ma, Shinomoto S, Tsutsumi S, Hashizume M, Ishikawa T, Kano M, Ikegaya Y, Kitamura K, Kawato M, et al. Improved hyperacuity estimation of spike timing from calcium imaging. Scientific reports. 2020; 10(1):1–16.

Holekamp TF, Turaga D, Holy TE. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. Neuron. 2008; 57(5):661–672.

Huys QJ, Paninski L. Smoothing of, and parameter estimation from, noisy biophysical recordings. PLoS computational biology. 2009; 5(5):e1000379.

Im DJ, Prakhya S, Yan J, Turaga S, Branson K. Importance Weighted Adversarial Variational Autoencoders for Spike Inference from calcium Imaging Data. arXiv preprint arXiv:190603214. 2019; .

Jewell S, Witten D. Exact spike train inference via ℓ_0 optimization. The annals of applied statistics. 2018; 12(4):2457.

Jewell SW, Hocking TD, Fearnhead P, Witten DM. Fast nonconvex deconvolution of calcium imaging data. Biostatistics. 2020; 21(4):709–726.

Kazemipour A, Liu J, Solarana K, Nagode DA, Kanold PO, Wu M, Babadi B. Fast and stable signal deconvolution via compressible state-space models. IEEE Transactions on Biomedical Engineering. 2017; 65(1):74–86.

Kerr JN, Greenberg D, Helmchen F. Imaging input and output of neocortical networks in vivo. Proceedings of the National Academy of Sciences. 2005; 102(39):14063–14068.

Lindsten F, Jordan MI, Schon TB. Particle Gibbs with ancestor sampling. Journal of Machine Learning Research. 2014; 15:2145–2184.

Lütcke H, Gerhard F, Zenke F, Gerstner W, Helmchen F. Inference of neuronal network spike dynamics and topology from calcium imaging data. Frontiers in neural circuits. 2013; 7:201.

Malik WQ, Schummers J, Sur M, Brown EN. Denoising two-photon calcium imaging data. PloS one. 2011; 6(6):e20490.

Mishchencko Y, Vogelstein JT, Paninski L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. The Annals of Applied Statistics. 2011; p. 1229–1261.

Mishchenko Y, Paninski L. Efficient methods for sampling spike trains in networks of coupled neurons. The Annals of Applied Statistics. 2011; p. 1893–1919.

Mukamel EA, Nimmerjahn A, Schnitzer MJ. Automated analysis of cellular signals from large-scale calcium imaging data. Neuron. 2009; 63(6):747–760.

Oñativia J, Dragotti PL. Sparse sampling: theory, methods and an application in neuroscience. Biological cybernetics. 2015; 109(1):125–139.

Oñativia J, Schultz SR, Dragotti PL. A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. Journal of neural engineering. 2013; 10(4):046017.

Pachitariu M, Stringer C, Dipoppa M, Schröder S, Rossi LF, Dalgleish H, Carandini M, Harris KD. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. BioRxiv. 2017; .

Pnevmatikakis EA, Gao Y, Soudry D, Pfau D, Lacefield C, Poskanzer K, Bruno R, Yuste R, Paninski L. A structured matrix factorization framework for large scale calcium imaging data analysis. arXiv preprint arXiv:14092903. 2014; .

Pnevmatikakis EA, Merel J, Pakman A, Paninski L. Bayesian spike inference from calcium imaging data. In: *2013 Asilomar Conference on Signals, Systems and Computers* IEEE; 2013. p. 349–353.

Pnevmatikakis EA, Soudry D, Gao Y, Machado TA, Merel J, Pfau D, Reardon T, Mu Y, Lacefield C, Yang W, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron. 2016; 89(2):285–299.

Quan T, Liu X, Lv X, Chen WR, Zeng S. Method to reconstruct neuronal action potential train from two-photon calcium imaging. Journal of biomedical optics. 2010; 15(6):066002.

Rahmati V, Kirmse K, Marković D, Holthoff K, Kiebel SJ. Inferring neuronal dynamics from calcium imaging data using biophysical models and Bayesian inference. PLoS computational biology. 2016; 12(2):e1004736.

Ranganathan GN, Koester HJ. Optical recording of neuronal spiking activity from unbiased populations of neurons with high spike detection efficiency and high temporal precision. Journal of neurophysiology. 2010; 104(3):1812–1824.

Rupprecht P, Carta S, Hoffmann A, Echizen M, Blot A, Kwan AC, Dan Y, Hofer SB, Kitamura K, Helmchen F, et al. A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. Nature Neuroscience. 2021; 24(9):1324–1337.

Sasaki T, Takahashi N, Matsuki N, Ikegaya Y. Fast and accurate detection of action potentials from somatic calcium fluctuations. Journal of neurophysiology. 2008; 100(3):1668–1676.

Sebastian J, Kumar MG, Viraraghavan VS, Sur M, Murthy HA. Spike Estimation From Fluorescence Signals Using High-Resolution Property of Group Delay. IEEE Transactions on Signal Processing. 2019; 67(11):2923–2936.

Sebastian J, Sur M, Murthy HA, Magimai-Doss M. Signal-to-signal neural networks for improved spike estimation from calcium imaging data. PLoS Computational Biology. 2021; 17(3):e1007921.

Shibue R, Komaki F. Deconvolution of calcium imaging data using marked point processes. PLoS computational biology. 2020; 16(3):e1007650.

Speiser A, Yan J, Archer E, Buesing L, Turaga SC, Macke JH. Fast amortized inference of neural activity from calcium imaging data with variational autoencoders. arXiv preprint arXiv:171101846. 2017; .

Stern M, Shea-Brown E, Witten D. Inferring the spiking rate of a population of neurons from wide-field calcium imaging. bioRxiv. 2020; .

Theis L, Berens P, Froudarakis E, Reimer J, Rosón MR, Baden T, Euler T, Tolias AS, Bethge M. Benchmarking spike rate inference in population calcium imaging. Neuron. 2016; 90(3):471–482.

Theis L, Chagas AM, Arnstein D, Schwarz C, Bethge M. Beyond GLMs: a generative mixture modeling approach to neural system identification. PLoS computational biology. 2013; 9(11):e1003356.

Tsunoda T, Oda Y, Omori T, Okada M, Inoue M, Miyakawa H, Aonishi T. Statistical Calibration Method for Physiological Ca2+ Fluorescence Signals. Australian Journal of Intelligent Information Processing Systems. 2010; 11(1).

Tsunoda T, Omori T, Miyakawa H, Okada M, Aonishi T. Estimation of intracellular calcium ion concentration by nonlinear state space modeling and expectation-maximization algorithm for parameter estimation. Journal of the Physical Society of Japan. 2010; 79(12):124801.

Tubiana J, Wolf S, Panier T, Debregeas G. Blind deconvolution for spike inference from fluorescence recordings. Journal of Neuroscience Methods. 2020; 342:108763.

Vogelstein JT, Packer AM, Machado TA, Sippy T, Babadi B, Yuste R, Paninski L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. Journal of neurophysiology. 2010; 104(6):3691–3704.

Vogelstein JT, Watson BO, Packer AM, Yuste R, Jedynak B, Paninski L. Spike inference from calcium imaging using sequential Monte Carlo methods. Biophysical journal. 2009; 97(2):636–655.

Wei XX, Zhou D, Grosmark A, Ajabi Z, Sparks F, Zhou P, Brandon M, Losonczy A, Paninski L. A zero-inflated gamma model for deconvolved calcium imaging traces. arXiv preprint arXiv:200603737. 2020; .

Yaksi E, Friedrich RW. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca 2+ imaging. Nature methods. 2006; 3(5):377–383.

Zhang Y, Rózsa M, Liang Y, Bushey D, Wei Z, Zheng J, Reep D, Broussard GJ, Tsang A, Tsegaye G, et al. Fast and sensitive GCaMP calcium indicators for imaging neural populations. BioRxiv. 2021; .

Zhou P, Resendez SL, Rodriguez-Romaguera J, Jimenez JC, Neufeld SQ, Giovannucci A, Friedrich J, Pnevmatikakis EA, Stuber GD, Hen R, et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. Elife. 2018; 7:e28728.