High frequency spike inference with particle Gibbs sampling

Giovanni Diana, B. Semihcan Sermet, David A. DiGregorio

Institut Pasteur, University of Paris, CNRS UMR 3571, Synapse and Circuit Dynamics Laboratory, Paris, France.
* giovanni.diana@pasteur.fr (GD); david.digregorio@pasteur.fr (DD)

Abstract

Fluorescent calcium indicators are an 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 rates greater than 20 Hz are lacking. More importantly, little attention has been devoted to the quantification the statistical uncertainties in spike time estimation, which is essential for assigning a confidence in the inference for a particular recording. To address these challenges, we introduce an auto-regressive generative model that accounts for bursting neuronal activity and baseline fluorescence modulation, and it also applies recent sequential Monte Carlo approaches to obtain joint posterior distributions of static and dynamic model parameters. We show that our inference method is competitive with state-of-the-art algorithms by analysing the CASCADE benchmark datasets. We also show that spike time intervals as short as five milliseconds can be inferred from fluorescence transients recorded using a state-of-the-art genetically encoded indicator. 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 all model parameters and it provides a 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[1–4], linear deconvolution[5–8], finite rate of innovation[9, 10], independent component analysis[11], non model-based signal processing[12], supervised learning[13–17], constrained non-negative matrix factorization[18–20], active set methods[21, 22], convex and non-convex optimization methods[23–27], interior point method[28]. Model-based approaches allow to frame the problem of spike inference in a Bayesian context and use maximum-a-posteriori estimates[29–32].

Optimization methods provide a single estimate of spike times by maximizing a cost function defined by the underlying model and constraints. This approach does not provide information about the statistical uncertainty associated to our estimates. To address this issue, previous works have proposed Bayesian inference methods[33–44] which, as opposed to optimization techniques, give access to the full probability distribution of the unknowns given the data. However, the statistical 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 static model parameters (e.g. kinetic constants) and dynamic variables equally. Instead, they require additional optimization procedures to calibrate model parameters, thus neglecting how the uncertainty about model parameters propagates to spike times.

A common approach in spike inference is to use Markov models in discrete-time (e.g. autoregressive models[45]) to describe the link between spikes and fluorescence. These models are known in the statistical literature as "state-space models" and they are used in time series analysis to describe the probabilistic dependence between data and unobserved variables (latent state). 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 the observations (time steps), the analysis of long time series requires efficient strategies to sample from high-dimensional spaces. A major breakthrough in the analysis of state-space models has been the introduction of sequential Monte Carlo methods[46]. 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 the joint distribution of static 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 fluorescence time series. Our generative model accounts for periods of high firing rates between periods of baseline (lower) firing

rate. By quantifying the performance of our method (PGBAR) on the CASCADE benchmark dataset[13] we have shown that our approach is competitive with existing techniques. Finally we tested PGBAR on in-vitro recordings of cerebellar granule cells using ultrafast GCaMP8f calcium indicator, showing that our method allows to detect spikes reliably even for high firing rates ($\sim 200Hz$).

2 Results

2.1 The model

In this section we introduce a generative model of fluorescence time series with an underlying spiking process that accounts for periods of increased firing rate, separated by periods of baseline (lower) firing rate. The normalized fluorescence trace $F_{1:T}$ is described 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 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 4.4 for a derivation). To be able to employ statistical inference methodologies developed for state-space models (particle Gibbs), we need to recast this model in the form of 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. [20])

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

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

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

Note that the calcium level trace is deterministic given the spike counts $s_{1:T}$.

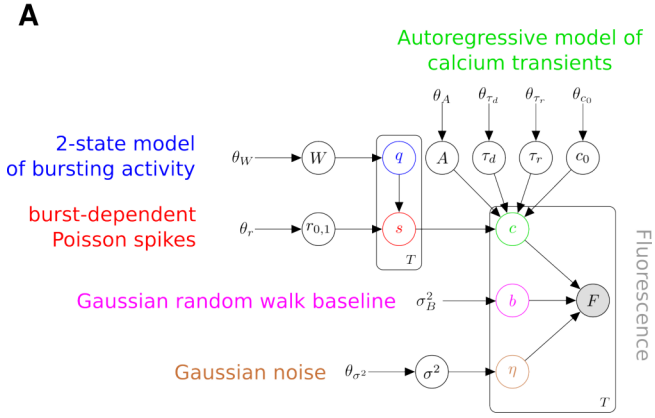
Bayesian inference requires the design of prior distributions on model parameters. However, it would be difficult to assign priors directly to $\gamma_{1,2}$ and A as they are not directly measurable. Instead, we will reparameterize the model using peak response ($A^{(max)}$), rise time (time to peak response, τ_r) and decay time (τ_d) of unitary fluorescence response, for which empirical estimates have been reported in previous works[49]. $A^{(max)}$, 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 4.5 for a derivation)

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

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

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

A



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.

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 static 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 notatation we will denote the latent space as $X = \{q_t, s_t, C_t, b_t\}$ and the combination of static 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 static 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 and 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

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

In order to infer latent states and static 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. Unfortunately, the posterior distribution for general state-space models is not analytically tractable. However we can use Monte Carlo methods, which rely on the ability to generate random samples from the posterior distribution, to obtain unbiased approximations of any statistical average with respect to the target distribution.

2.3 Sequential Monte Carlo

In this work we use sequential Monte Carlo methods to approximate the posterior distribution in Eq. (13). Suppose to have generated N samples $X^{(n)}$, $n = 1, \cdots, N$ from a target distribution $P(X)$. For any random variable $V(X)$, the empirical average of V over the Monte Carlo samples

$$\langle V \rangle \approx \frac{1}{N} \sum_n V(X^{(n)}) \tag{14}$$

provides un unbiased estimator of the expectation of V with respect to $P(X)$. There are two critical issues that arise when applying such method to time series models: first, the high dimensionality of the latent space and, second, the joint inference of model parameters and latent state trajectory. In a state-space model the number of unknowns typically scales linearly with number of observations. This is a problem for standard sampling methods, such as Markov Chain Monte Carlo, which suffer from the so-called "curse of dimensionality" as their performance rapidly decreases at high dimensions. Sequential Monte Carlo (SMC) methods allow to address the dimensionality issue by providing efficient strategies to sample from the latent space. The typical approach is to construct a sequential approximation of the posterior distribution where observations are accounted iteratively. SMCs provide a solution to the so-called "filtering" problem of estimating latent states at fixed model parameters.

In a highly influential work, Andrieu et al[47] introduced the particle Gibbs algorithm as a method to sample both static and dynamic variables in a state-space model. This algorithm alternates the sampling of model parameters and latent trajectories as in the Gibbs sampler, with the difference that latent states are sampled from a 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[48], named particle Gibbs with ancestor sampling (PGAS), with better mixing properties (see Algorithm 2 and the Methods section).

To carry out inference of static and dynamical variables for our model, we employed Algorithm 1 which alternates the two steps mentioned above: first we sample a new latent state trajectory by running the PGAS transition kernel, then we sample new model parameters according to their full conditional distributions (when available analytically) or a Metropolis-Hastings kernel.

Algorithm 1 Gibbs sampler

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

2.4 Validation and performance of PGBAR on simulated 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 patter 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. In order 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 state ($q_t = 0, 1$ in the model) of the neuron. In particular, the probability of 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 shows the comparison between 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 static model parameters. Figure 2G illustrates the posterior distributions associated to peak amplitude, noise level, decay and rise time of the fluorescence probe. The ground truth parameters used to simulate the testing time series are always close to the peak of the corresponding posterior distributions, showing the identifiability of our model.

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 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 at 5Hz and 50Hz. For the 5Hz firing trace the bursting model did not improve the inference accuracy compared to the variant with 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 in the number of spikes for the 50 Hz trace, whereas the original model captures reliably the ground-truth spikes.

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 level and firing frequency, the performance difference between bursting and non-bursting versions of our algorithm becomes 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

In order to test our method on experimental data we analyzed neuronal recordings from the CASCADE benchmark dataset[13], which allowed us to quantify the performance of our algorithm on different calcium indicators. In Figure 4 we illustrate the application of our method on GCaMP6f fluorescence data from a pyramidal neuron in the mouse visual cortex (CASCADE dataset 9[49]). The comparison between ground-truth spikes and the ones inferred using PGBAR in Fig. 4A 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 model constraints (for instance the fixed peak response along the recording) or to experimental errors in estimating the ground-truth.

The estimated bursting pattern shown in Fig. 4A captures the overall periods of increased neuronal activity. The posterior distributions of model parameters are shown in Fig. 4B. Some of these distributions (burst firing rate and the rise time) shift significantly from their corresponding priors. This mismatch between prior information and data can arise when using the prior to penalize certain regions. Figure 5 summarizes our analysis of the CASCADE datasets. To quantify the performance or PGBAR and allow direct comparison with previous analyses in Ref. [13] we used the Pearson's correlation coefficient between ground-truth spikes and predicted spikes both filtered with a Gaussian kernel with 200 ms bandwith. We did not find a particular condition where our method performed better or worse by analysing different calcium indicators (Fig. 5A). The overall correlation averaged across cells and datasets is 0.75, however, consistently with previous studies, we observed a large variability of performance across cells and indicators. In Ref. [13], Rupprecht et al introduced the notion of standardized noise level as the ratio between the standard deviation of the normalized fluorescence and the square root of the sampling frequency. As shown in Fig. 5B, we found that the performance of PGBAR is robust across standardized noise levels.

We compared the performance of our method to CASCADE[13], MLSpike[29], Peeling[50], CaImAn[51], Suite2p[52] and JewellWitten[23] by using their previously benchmarked performance on the same datasets obtained from extensive parameter optimization[13] (Fig. 5C). The performance distribution of PGBAR across the available recordings in the CASCADE database was comparable to previous methodologies. Among model-based approaches, our method was slightly underperformed with respect to MLSpike, likely due to its use of a more accurate description of the calcium indicator. The top performance is achieved by the supervised CASCADE method, however it cannot be used to obtain statistical uncertainties.

Figure 6A shows two simulated fluorescence traces with two spikes 20ms apart at low (1.4) and high (3.4) SNR levels. By running our algorithm on these simulated data we extracted the posterior distribution of total number of spikes (Fig. 6B), showing the correct identification of two spikes at both SNR levels. The posterior ISI distribution conditional to 2 spikes (Fig. 6C) showed a narrow spike distribution (SD $\approx$ 1ms) centered on the ground truth interval of 20 ms for low noise, while the higher noise trace resulted in a spike interval distribution that was broader (SD $\approx$ 1.6 ms) and whose mode ws shifted from the ground truth by -1.6 ms.

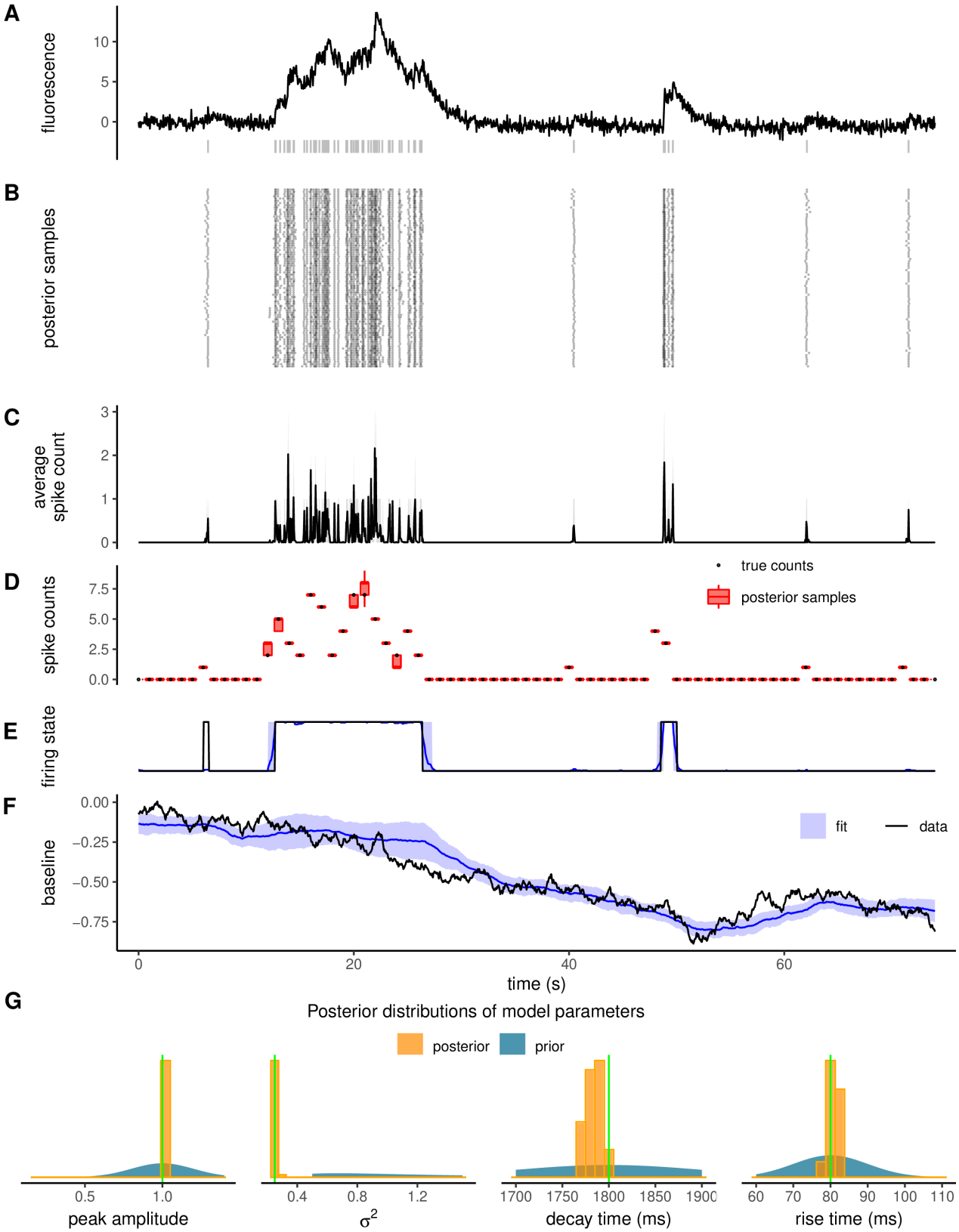


Figure 2: **Validation of the spike inference approach with simulated data. (A)** Example trajectory simulated from the model (solid, black) with ground-truth spike times shown underneath (grey vertical lines). **(B)** Raster plot representing spike times for a thousand Monte Carlo samples. **(C)** Average 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 single spike, decay time, rise time and noise level compared with true value (vertical lines in green).

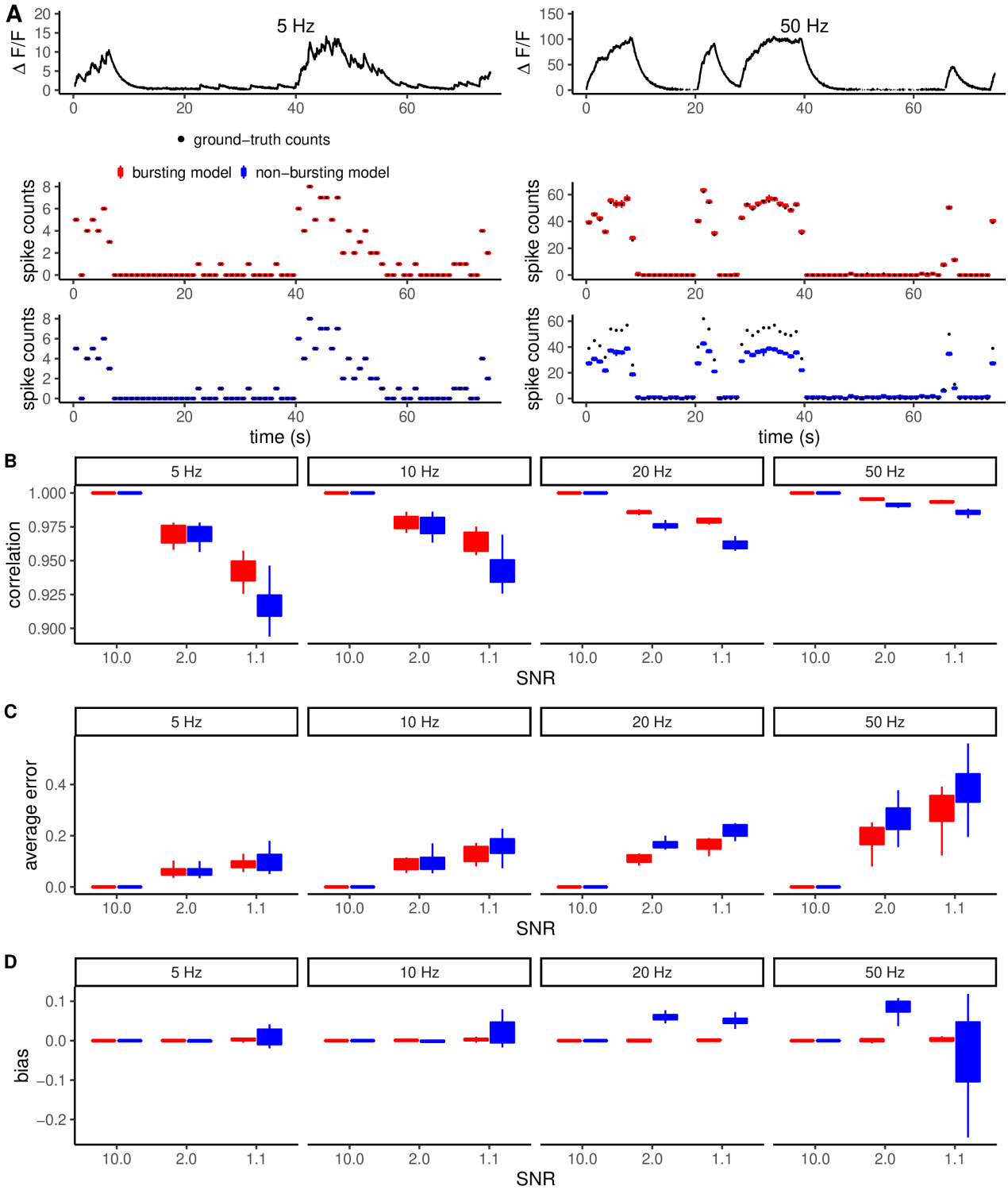


Figure 3: **Dependency of inference performance on noise and firing frequency and the bias of non-bursting models. (A)** Example fluorescence time series simulated at 5Hz and 50Hz bursting frequencies (top). The analysis of these traces using bursting and non-bursting variant of the model highlights 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.

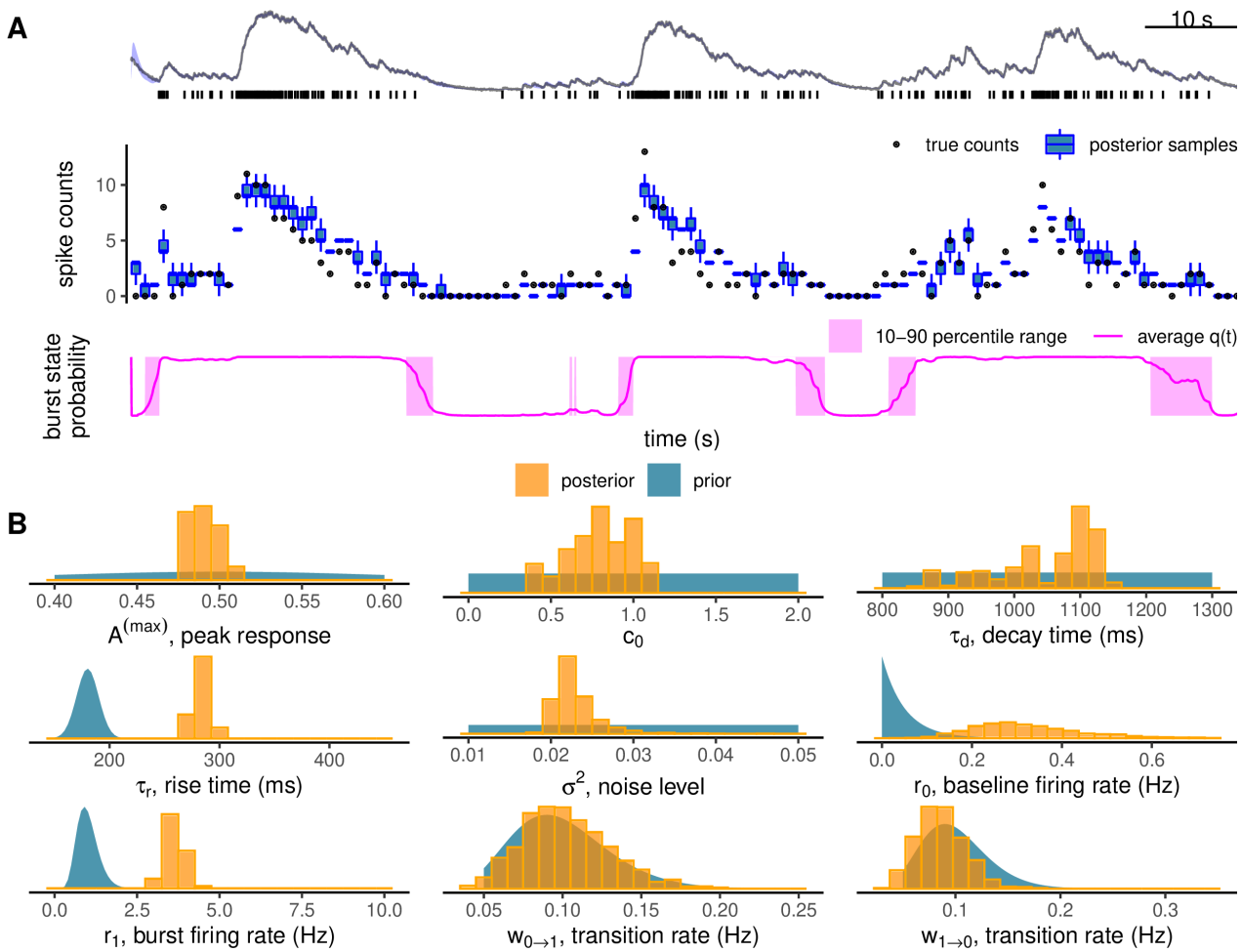


Figure 4: **Analysis of GCaMP6f recordings from the CASCADE dataset. (A)** example $\Delta F/F$ from the CASCADE dataset (#DS09, GCaMP6f, mouse visual cortex) with ground-truth spikes shown underneath fluorescence (top), comparison of spike counts within 1s time intervals (middle) and burst probability (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 ($w_{0\to1}, w_{1\to0}$).

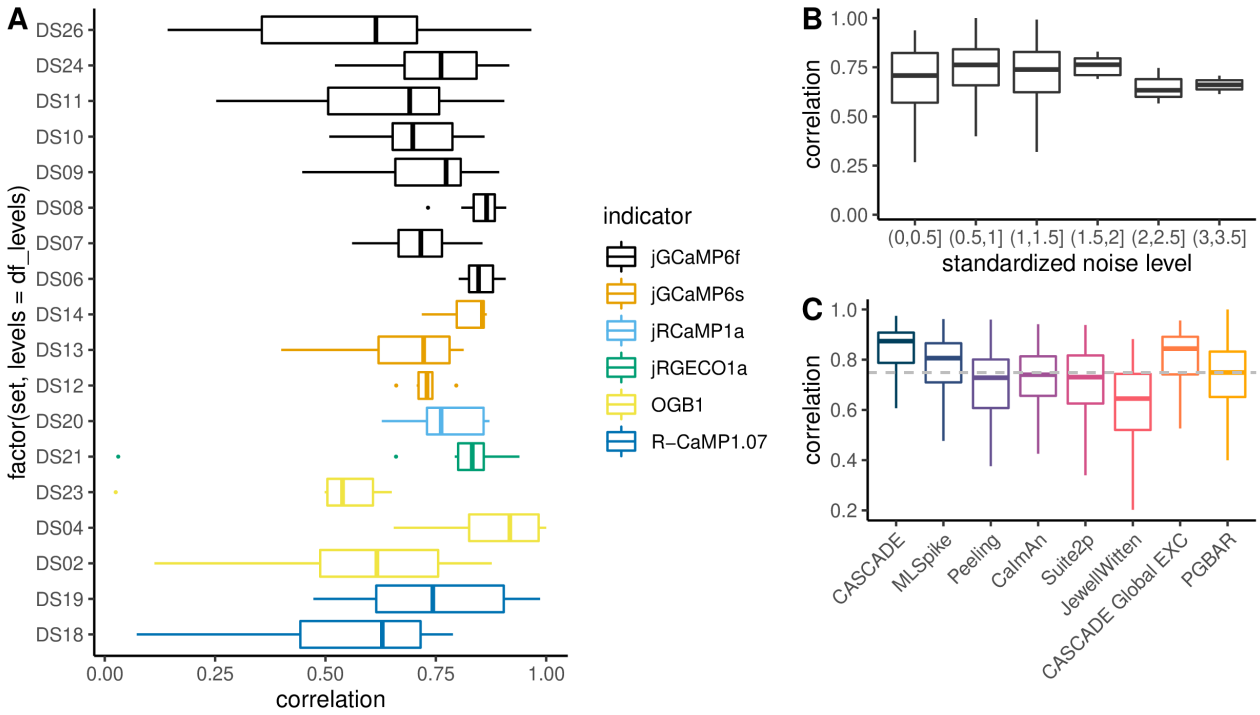


Figure 5: **Analysis of CASCADE benchmark data and comparison with existing methods. (A)** Correlation between estimated and ground-truth firing rates from CASCADE datasets. The color code represents the different calcium indicators employed in each dataset.**(B)** Correlation with ground-truth spikes as a function of the standardized noise level[13]. **(C-D)** Comparison with existing methods. Correlation averaged across datasets and neurons (**C**).

To examine the spike discrimination performance for multiple intervals (3-20 ms) and noise levels (SNR 1-3.8), we calculated the ISI posterior probability evaluated at the ground-truth ISI and displayed in the contour plot in Fig. 6D as a function of ISI and SNR. When this probability is higher than 0.5, the peak of the posterior coincides with the ground truth value. The SNR level at which the ground-truth probability is larger than 0.5 depends weakly on the ISI. Figure 6E 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 large variance while at higher SNR levels they shrink around the ground-truth value. This analysis reflects the degree of variability expected when analyzing multiple recordings of the same neuron.

2.6 PGBAR spike inference from fluoresence transients recorded using the fast calcium indicator GCaMP8f

We tested our approach on the fast calcium indicator GCaMP8f by performing high-speed (≈3.3kHz) 2-photon linescan calcium imaging of cerebellar granule cells *in vitro*. We used adeno-associated viruses (AAV) to express GCaMP8f in Lobe X of the cerebellum (Fig. 7A). Fluorescence was recorded at granule cell somata and ground truth spikes were evoked by extracellular stimulation of granule cell axons in the molecular layer (Fig. 7B).

In order to constrain our inference, we quantified the amplitude and kinetics (rise and decay times) of single AP-evoked GCaMP8f fluorescence transients. Analyzing these recordings using our approach allowed us to build prior distributions on kinetic parameters of unitary fluorescence responses (Fig. 7C). Next, we recorded granule cell activity in response to a 20 Hz Poisson stimulation protocols (Fig. 7E). Figure 7F shows the average spike count obtained from PGBAR by analyzing each trial independently. The spike patterns obtained using our method are very similar across trials, showing that PGBAR can reliably detect single trial action potentials.

To illustrate the temporal accuracy of PGBAR we focused on the short interval between the first two spikes (Fig. 7G). In spite of the relatively low SNR ($A^{(max)}/\sigma \approx 2.4$), we could reliably identify the two spikes in each single trial. Figure 7H shows the posterior distribution of the inter spike interval obtained by analysing each trial independently. The ground-truth inter spike interval of 5.3 ms is well within each posterior distribution obtained from single trials. In addition, the posterior modes across trials are distributed symmetrically around the ground-truth ISI of 5.3 ms with a trial-averaged standard deviation of 2.5 ms, highlighting the unbiaseness of our analysis.

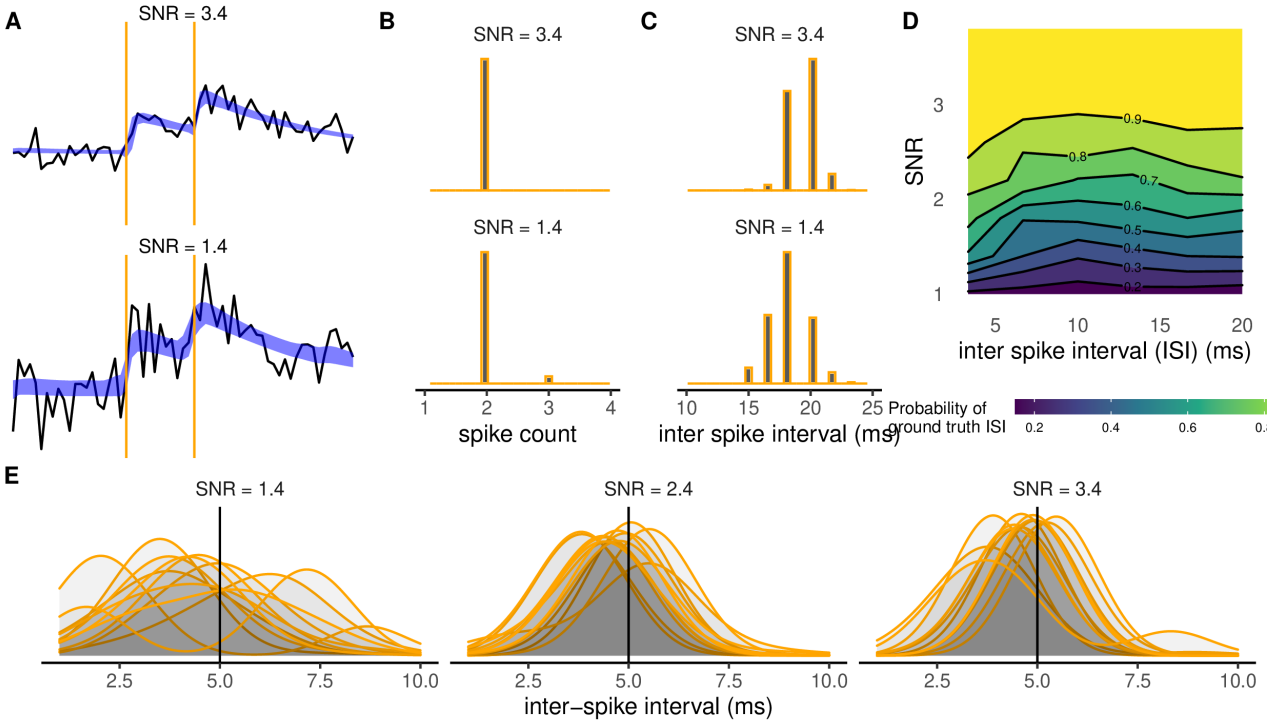


Figure 6: **Sensitivity of spike detection to sampling frequency and SNR level. (A)** Examples of simulated fluorescence traces with two spikes separated by 20ms (vertical lines) at low (1.4) and high (3.4) SNR. Here the sampling frequency was set to 600Hz. Ribbon lines display denoised fits (calcium level plus baseline) within one standard deviation. (**B**) Posterior distribution of the total number of detected spikes. (**C**) Posterior distribution of the inter-spike interval (ISI). **(D)** Raster plot of the probability of the true ISI at different SNR levels and ISI. **(E)** Comparison of ISI posterior distributions generated from the analysis of 12 fluorescence traces with two spikes 5ms apart and sampling frequency of 3kHz. Density plots have been smoothed with 1ms bandwith. In all simulated traces we used $\tau_r = 3.7$ms and $\tau_d = 40$ms.

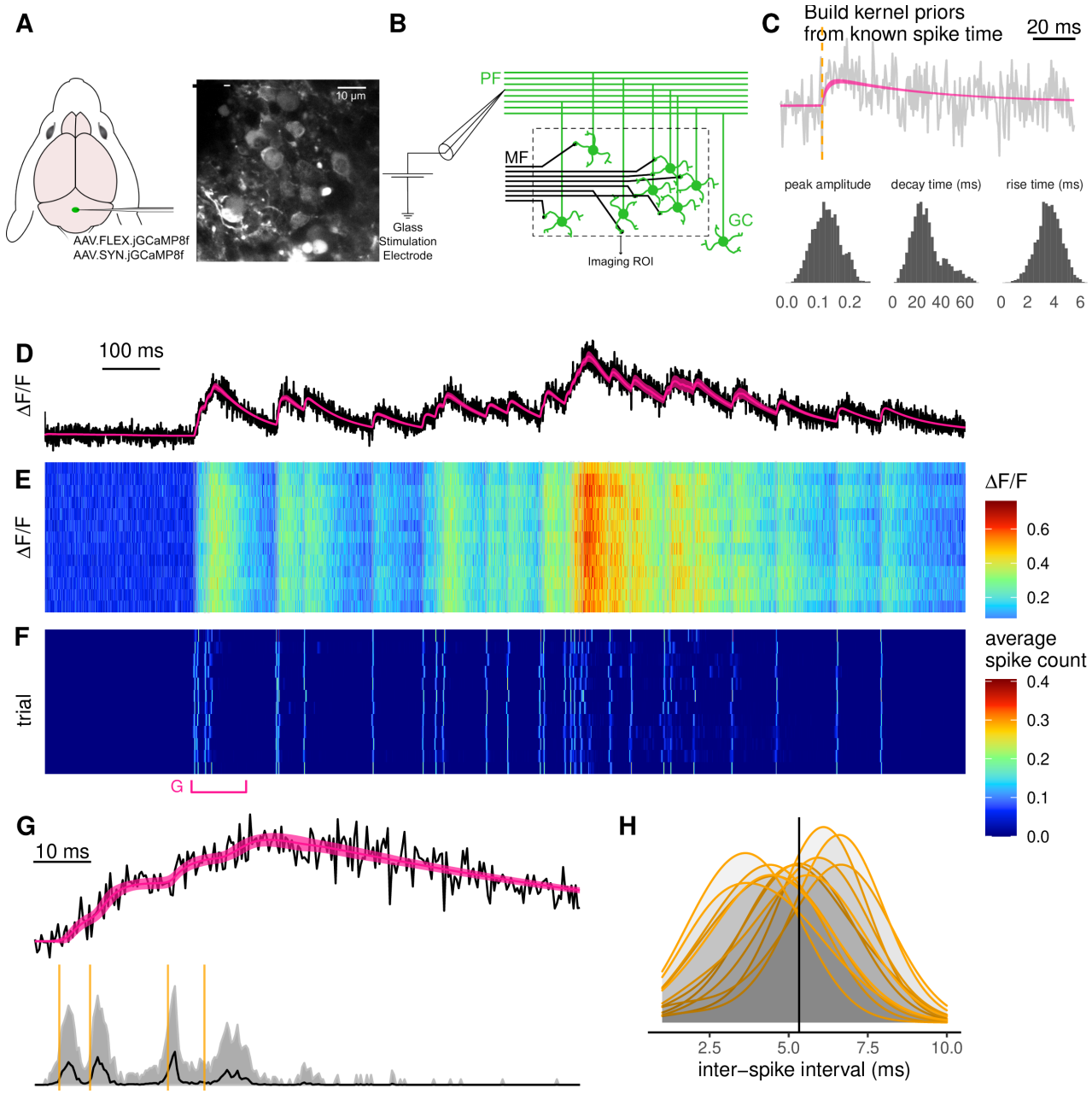


Figure 7: **High-speed 2-photon linescan calcium imaging.** **(A)** GCaMP8f virus injection in the cerebellar vermis. **(B)** Induction of action potentials to cerebellar granule cells by direct stimulation of the parallel fibers. **(C)** Use of single pulse stimulation to extract kinetic model parameters of GCaMP8f indicator. **(D-E)** High-speed (3kHz) 2-photon linescan calcium imaging of granule cell bodies. 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 stimulation.

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[53] and establish their causal relationships. The works from Pnevmatikakis et al[34] and Vogelstein et al[37], 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 link spiking patterns to fluorescence time series and provide 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. Unlike existing model-based approaches, including Monte Carlo methods[34, 37, 40], our statistical inference is based on a model that accounts for non homogeneous Poisson firing statistics and baseline modulation. In particular, 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 esplicitely 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 by Pnevmatikakis et al[34] did not account for fluorescence baseline modulation, they highlighted its importance for in vivo recordings. PGBAR uses a Gaussian random walk (analogously to MLSpike[29]) to describe slow changes in baseline fluorescence across the recording. This is the simplest Markov model of fluorescence baseline but it generates noisy baselines. To avoid this effect, it is possible to employ alternative baseline models, such as the integrated random walk, which allows to reduce the additional noise introduced by the baseline stochastic process.

Joint estimation of static parameters and dynamic variables. The estimation of static 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 setting their corresponding 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[37, 40], allowing for a joint estimation of static parameters and dynamic variables.

Comparison with benchmark datasets. The proliferation of spike inference methodologies led to the development of community-based initiatives[14, 54] to rank the performance of available methods. We applied our approach on the CASCADE dataset[13] which provides a curated database of neuronal recordings from mouse and zebrafish using different calcium indicators. The performance of PGBAR is comparable to existing unsupervised approaches. In addition, it provides information about the statistical uncertainty associated to spike detection that is not available through 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[55] with 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 might boost dramatically these methods and offering tools for online processing of fluorescence time series.

Many commonly used indicators are nonlinear in the peak amplitude[49]. This type of behavior is not accounted by the autoregressive model employed in this study, which assumes that action potentials generate an invariant increase in fluorescence. This work provides a statistical framework generalizable to more specific biophysical models

of calcium indicators to account for non-linear effects. Accurate descriptions of the calcium probe are likely to increase the number of parameters, which might introduce issues of model identifiability. Our Bayesian framework offers a systematic approach to address this issues by integrating current and future data on the kinetics of calcium indicators into informative priors constraining the Monte Carlo sampler within biophysically relevant parametric 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 by Lindsten et al in Ref. [48] to improve the performance of the original particle Gibbs sampler[47]. 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 $t = 1$,

$$X_1^{(i)} = \{q_1^{(i)}, s_1^{(i)}, C_1^{(i)}, b_1^{(i)}\} \quad i = 1, \cdots, N \tag{15}$$

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

$$\mu^\theta(X_1) = \rho_1(X_1) = \frac{(r_{q_1}\Delta)^{s_1}}{s_1!} e^{-r_{q_1}\Delta} \cdot \frac{e^{-b_1^2/2}}{\sqrt{2\pi}} \delta^2(C_1 - \hat{C}_1) \tag{16}$$

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

$$w_1^{(i)} = \mu^\theta(X_1^{(i)}) g^\theta(F_1|X_1^{(i)})/\rho_1(X^{(i)}), \quad i = 1, \cdots, N. \tag{17}$$

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

$$\mathbb{P}(J = i) = \frac{w_{t-1}^{(i)} f_t^\theta(X_t'|X_{t-1}^{(i)})}{\sum_{k=1}^N w_{t-1}^{(k)} f_t^\theta(X_t'|X_{t-1}^{(k)})} \tag{18}$$

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

$$P(X_t|F_t, X_{t-1}) = \frac{P(X_t, F_t|X_{t-1})}{\int_{X_t} P(X_t, F_t|X_{t-1})} = \frac{f_t^\theta(X_t|X_{t-1}) g_t^\theta(F_t|X_t)}{Z^\theta(X_{t-1}, F_t)} \tag{19}$$

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

$$P(X_t = \{q_t, s_t, C_t, b_t\}|X_{t-1}, F_t) = P(q_t, s_t|X_{t-1}, F_t) \cdot P(C_t|C_{t-1}, s_t) \cdot P(b_t|q_t, s_t, C_t, X_{t-1}, F_t) \tag{20}$$

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

$$\begin{aligned}
\int db_t dC_t\, f_t^\theta(X_t|X_{t-1}) g_t^\theta(F_t|X_t) &= \int db_t dC_t \delta^{(2)}(C_t - M \cdot C_{t-1} - AS_t) \cdot W_{q_{t-1}q_t} \frac{(r_{q_t}\Delta)^{s_t}}{s_t!} e^{-r_{q_t}\Delta} \cdot \\
&\cdot (2\pi\Delta\sigma_b^2)^{-1/2} \exp\left(-\frac{1}{2\Delta\sigma_b^2}(b_t - b_{t-1})^2\right) \cdot (2\pi\sigma^2)^{-1/2} \cdot \exp\left[-\frac{1}{2\sigma^2}(F_t - c_t - b_t)^2\right] \\
&= W_{q_{t-1}q_t} \frac{(r_{q_t}\Delta)^{s_t}}{s_t!} e^{-r_{q_t}\Delta} \cdot I(b_{t-1}, F_t - c_t, \Delta\sigma_b^2, \sigma^2)
\end{aligned} \tag{21}$$

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

$$I(y_1, y_2, \sigma_1^2, \sigma_2^2) = (2\pi\sigma_1^2)^{-1/2}(2\pi\sigma_2^2)^{-1/2} \int dx e^{-\frac{(x-y_1)^2}{2\sigma_1^2} - \frac{(x-y_2)^2}{2\sigma_2^2}} = \frac{\exp\left[-\frac{1}{2}\frac{(y_1-y_2)^2}{\sigma_1^2+\sigma_2^2}\right]}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \tag{22}$$

The normalization factor $Z^\theta(X_{t-1}, F_t)$ is obtained by taking the sum of Eq. (21) over firing state and spike count:

$$Z^\theta(X_{t-1}, F_t) = \sum_{q' \in \{0,1\}} \sum_{s'=0}^{\infty} W_{q_{t-1}q'} \frac{(r_{q'}\Delta)^{s'}}{s'!} e^{-r_{q'}\Delta} \cdot I(b_{t-1}, F_t - c_t, \Delta\sigma_b^2, \sigma^2). \tag{23}$$

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

$$\mu_{prop} = \frac{b_{t-1}\sigma^2 + (F_t - c_t)\sigma_b^2\Delta}{\sigma^2 + \sigma_b^2\Delta} \tag{24}$$

$$\sigma_{prop}^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_b^2\Delta}\right)^{-1}. \tag{25}$$

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

$$w_t^{(i)} = f_t^\theta(X_t^{(i)}|\tilde{X}_{t-1}^{(i)})g_t^\theta(F_t|X_t^{(i)})/\rho_t(X_t^{(i)}|\tilde{X}_{t-1}^{(i)}) \tag{26}$$

However, due to the form of the optimal proposal in Eq. (19) the importance weights reduce to the normalization factor calculated in Eq. (23)

$$w^{(i)} = Z^\theta(X_{t-1}^{(i)}, F_t) \tag{27}$$

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 static parameters

In Algorithm 1, after a new latent state trajectory is sampled from the PGAS kernel, we draw static 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

$$P(\theta) \cdot P_\theta(X_{1:T}, F_{1:T}|\theta) \tag{28}$$

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

$$P(r_q|\cdots) \propto r_q^{\alpha-1} e^{-\beta r_q - r_q \Delta T} r_q^{\sum_{t:q_t=q} s_t} \tag{29}$$

therefore the full conditional is a gamma distribution with updated parameters

$$\alpha' = \alpha + \sum_{t:q_t=q} s_t \tag{30}$$

$$\beta' = \beta + \Delta T \tag{31}$$

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

$$P(w_{q\to q'}|\cdots) \propto w_{q\to q'}^{\alpha-1} e^{-\beta w_{q\to q'}} w_{q\to q'}^{N_{qq'}} (1 - \Delta w_{q\to q'})^{N_{qq}} \approx w_{q\to q'}^{\alpha+N_{qq'}-1} e^{-(\beta+\Delta N_{qq}) w_{q\to q'}} \tag{32}$$

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

$$\alpha' = \alpha + N_{qq'} \tag{33}$$

$$\beta' = \beta + \Delta N_{qq} \tag{34}$$

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

$$P(\sigma^2|\cdots) \propto (\sigma^2)^{-\alpha-1-T/2} \exp\left(-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}\sum_t (F_t - c_t - b_t)^2\right) \tag{35}$$

therefore the updated shape and rate of the inverse gamma are

$$\alpha' = \alpha + T/2 \tag{36}$$

$$\beta' = \beta + \frac{1}{2}\sum_t (F_t - c_t - b_t)^2 \tag{37}$$

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

$$C_t = M \cdot C_{t-1} + AS_t, \quad M = \begin{bmatrix} \gamma_1 & \gamma_2 \\ 1 & 0 \end{bmatrix} \tag{38}$$

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

$$C_t = M^{t-1} C_1 + A \sum_{k=2}^{t} M^{t-k} S_k. \tag{39}$$

If $s_t = \delta_{t,1}$ and $c_0 = 0$ then Eq. (39) simplifies to

$$C_t = AM^{t-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{40}$$

By introducing eigenvectors and eigenvalues of M

$$\gamma_\pm \equiv \frac{\gamma_1 \pm \sqrt{\gamma_1^2 + 4\gamma_2}}{2}, \quad e_\pm \equiv \begin{bmatrix} \gamma_\pm \\ 1 \end{bmatrix} \tag{41}$$

we can express Eq. (40) as

$$C_t = AM^{t-1}\left(\frac{e_+ - e_-}{\gamma_+ - \gamma_-}\right) = A\left(\frac{\gamma_+^{t-1}e_+ - \gamma_-^{t-1}e_-}{\gamma_+ - \gamma_-}\right) \tag{42}$$

therefore we have

$$c_t = A\left(\frac{\gamma_+^t - \gamma_-^t}{\gamma_+ - \gamma_-}\right). \tag{43}$$

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

$$\tau_r = \frac{\log\left(\frac{g_+}{g_-}\right)}{g_- - g_+}, \quad g_\pm = \log \gamma_\pm \tag{44}$$

whereas if we take the long time limit of Eq. (40) we obtain

$$c_t = A\gamma_+^t[1 - (\gamma_-/\gamma_+)^t] \approx Ae^{t/\tau_d}, \quad \tau_d = -\frac{1}{g_+} \tag{45}$$

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

$$\frac{\tau_r}{\tau_d} = \frac{\log \frac{g_-}{g_+}}{\frac{g_-}{g_+} - 1} \tag{46}$$

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

$$\frac{g_-}{g_+} = f^{-1}\left(\frac{\tau_r}{\tau_d}\right), \quad f(x) = \frac{\log(x)}{x-1} \tag{47}$$

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

$$g_+ = -\frac{1}{\tau_d} \tag{48}$$

$$g_- = g_+ \cdot f^{-1}\left(\frac{\tau_r}{\tau_d}\right) \tag{49}$$

and then

$$\gamma_1 = e^{g_+} + e^{g_-} \tag{50}$$

$$\gamma_2 = -e^{g_+ + g_-} \tag{51}$$

4.6 Experimental methods

GcaMP8f virus injection in the cerebellar vermis. Virus injection was targeted to Lobe X (6.75 mm posterior to Bregma; 0 mm lateral to the midline; vertical depth 2.42 mm from pial surface) under deep isofluorane anesthesia. 100 nl of adeno-associated virus encoding GcaMP8f (AAV1-syn-GCaMP8f-WPRE, 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 Shank3tm1Tmb mouse is 5 months old. After injection mouse was returned to its home cage for 8 weeks to allow time for expression.

Slice Preparation. Acute coronal slices (200 μm) of cerebellar vermis were prepared from adult CB6F1 mouse, aged 118 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 cerebellar vermis using a vibratome (Leica VT1200S), and incubated at room temperature for 30 min 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 60×, 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-gradiend 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 scaning 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 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

[1] Benjamin F Grewe, Dominik Langer, Hansjörg Kasper, Björn M Kampa, and Fritjof Helmchen. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nature methods*, 7(5):399–405, 2010.

[2] Eva L Dyer, Christoph Studer, Jacob T Robinson, and Richard G Baraniuk. 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)*, pages 593–596. IEEE, 2013.

[3] Jason ND Kerr, David Greenberg, and Fritjof Helmchen. Imaging input and output of neocortical networks in vivo. *Proceedings of the National Academy of Sciences*, 102(39):14063–14068, 2005.

[4] Tingwei Quan, Xiuli Liu, Xiaohua Lv, Wei R Chen, and Shaoqun Zeng. Method to reconstruct neuronal action potential train from two-photon calcium imaging. *Journal of biomedical optics*, 15(6):066002, 2010.

[5] Terrence F Holekamp, Diwakar Turaga, and Timothy E Holy. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. *Neuron*, 57(5):661–672, 2008.

[6] Jérôme Tubiana, Sébastien Wolf, Thomas Panier, and Georges Debregeas. Blind deconvolution for spike inference from fluorescence recordings. *Journal of Neuroscience Methods*, 342:108763, 2020.

[7] Xue-Xin Wei, Ding Zhou, Andres Grosmark, Zaki Ajabi, Fraser Sparks, Pengcheng Zhou, Mark Brandon, Attila Losonczy, and Liam Paninski. A zero-inflated gamma model for deconvolved calcium imaging traces. *arXiv preprint arXiv:2006.03737*, 2020.

[8] Emre Yaksi and Rainer W Friedrich. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved ca 2+ imaging. *Nature methods*, 3(5):377–383, 2006.

[9] Jon Oñativia, Simon R Schultz, and Pier Luigi Dragotti. A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. *Journal of neural engineering*, 10(4):046017, 2013.

[10] Jon Oñativia and Pier Luigi Dragotti. Sparse sampling: theory, methods and an application in neuroscience. *Biological cybernetics*, 109(1):125–139, 2015.

[11] Eran A Mukamel, Axel Nimmerjahn, and Mark J Schnitzer. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron*, 63(6):747–760, 2009.

[12] Jilt Sebastian, Mari Ganesh Kumar, Venkata Subramanian Viraraghavan, Mriganka Sur, and Hema A Murthy. Spike estimation from fluorescence signals using high-resolution property of group delay. *IEEE Transactions on Signal Processing*, 67(11):2923–2936, 2019.

[13] Peter Rupprecht, Stefano Carta, Adrian Hoffmann, Mayumi Echizen, Antonin Blot, Alex C Kwan, Yang Dan, Sonja B Hofer, Kazuo Kitamura, Fritjof Helmchen, et al. A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nature Neuroscience*, 24(9):1324–1337, 2021.

[14] Lucas Theis, Philipp Berens, Emmanouil Froudarakis, Jacob Reimer, Miroslav Román Rosón, Tom Baden, Thomas Euler, Andreas S Tolias, and Matthias Bethge. Benchmarking spike rate inference in population calcium imaging. *Neuron*, 90(3):471–482, 2016.

[15] Jilt Sebastian, Mriganka Sur, Hema A Murthy, and Mathew Magimai-Doss. Signal-to-signal neural networks for improved spike estimation from calcium imaging data. *PLoS Computational Biology*, 17(3):e1007921, 2021.

[16] Huu Hoang, Masa-aki Sato, Shigeru Shinomoto, Shinichiro Tsutsumi, Miki Hashizume, Tomoe Ishikawa, Masanobu Kano, Yuji Ikegaya, Kazuo Kitamura, Mitsuo Kawato, et al. Improved hyperacuity estimation of spike timing from calcium imaging. *Scientific reports*, 10(1):1–16, 2020.

[17] Takuya Sasaki, Naoya Takahashi, Norio Matsuki, and Yuji Ikegaya. Fast and accurate detection of action potentials from somatic calcium fluctuations. *Journal of neurophysiology*, 100(3):1668–1676, 2008.

[18] Pengcheng Zhou, Shanna L Resendez, Jose Rodriguez-Romaguera, Jessica C Jimenez, Shay Q Neufeld, Andrea Giovannucci, Johannes Friedrich, Eftychios A Pnevmatikakis, Garret D Stuber, Rene Hen, et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife*, 7:e28728, 2018.

[19] Eftychios A Pnevmatikakis, Yuanjun Gao, Daniel Soudry, David Pfau, Clay Lacefield, Kira Poskanzer, Randy Bruno, Rafael Yuste, and Liam Paninski. A structured matrix factorization framework for large scale calcium imaging data analysis. *arXiv preprint arXiv:1409.2903*, 2014.

[20] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016.

[21] Johannes Friedrich and Liam Paninski. Fast active set methods for online spike inference from calcium imaging. *Advances In Neural Information Processing Systems*, 29:1984–1992, 2016.

[22] Johannes Friedrich, Pengcheng Zhou, and Liam Paninski. Fast online deconvolution of calcium imaging data. *PLoS computational biology*, 13(3):e1005423, 2017.

[23] Sean W Jewell, Toby Dylan Hocking, Paul Fearnhead, and Daniela M Witten. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4):709–726, 2020.

[24] Merav Stern, Eric Shea-Brown, and Daniela Witten. Inferring the spiking rate of a population of neurons from wide-field calcium imaging. *bioRxiv*, 2020.

[25] Sean Jewell and Daniela Witten. Exact spike train inference via ℓ_0 optimization. *The annals of applied statistics*, 12(4):2457, 2018.

[26] Wasim Q Malik, James Schummers, Mriganka Sur, and Emery N Brown. Denoising two-photon calcium imaging data. *PloS one*, 6(6):e20490, 2011.

[27] Gayathri N Ranganathan and Helmut J Koester. Optical recording of neuronal spiking activity from unbiased populations of neurons with high spike detection efficiency and high temporal precision. *Journal of neurophysiology*, 104(3):1812–1824, 2010.

[28] Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of neurophysiology*, 104(6):3691–3704, 2010.

[29] Thomas Deneux, Attila Kaszas, Gergely Szalay, Gergely Katona, Tamás Lakner, Amiram Grinvald, Balázs Rózsa, and Ivo Vanzetta. Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nature communications*, 7(1):1–17, 2016.

[30] Alyson K Fletcher and Sundeep Rangan. Scalable inference for neuronal connectivity from calcium imaging. *arXiv preprint arXiv:1409.0289*, 2014.

[31] Abbas Kazemipour, Ji Liu, Krystyna Solarana, Daniel A Nagode, Patrick O Kanold, Min Wu, and Behtash Babadi. Fast and stable signal deconvolution via compressible state-space models. *IEEE Transactions on Biomedical Engineering*, 65(1):74–86, 2017.

[32] Takamasa Tsunoda, Toshiaki Omori, Hiroyoshi Miyakawa, Masato Okada, and Toru Aonishi. 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*, 79(12):124801, 2010.

[33] Yuriy Mishchencko, Joshua T Vogelstein, and Liam Paninski. A bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *The Annals of Applied Statistics*, pages 1229–1261, 2011.

[34] Eftychios A Pnevmatikakis, Josh Merel, Ari Pakman, and Liam Paninski. Bayesian spike inference from calcium imaging data. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 349–353. IEEE, 2013.

[35] Yuriy Mishchenko and Liam Paninski. Efficient methods for sampling spike trains in networks of coupled neurons. *The Annals of Applied Statistics*, pages 1893–1919, 2011.

[36] Quentin JM Huys and Liam Paninski. Smoothing of, and parameter estimation from, noisy biophysical recordings. *PLoS computational biology*, 5(5):e1000379, 2009.

[37] Joshua T Vogelstein, Brendon O Watson, Adam M Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential monte carlo methods. *Biophysical journal*, 97(2):636–655, 2009.

[38] Lucas Theis, Andrè Maia Chagas, Daniel Arnstein, Cornelius Schwarz, and Matthias Bethge. Beyond glms: a generative mixture modeling approach to neural system identification. *PLoS computational biology*, 9(11):e1003356, 2013.

[39] Takamasa Tsunoda, Yoshiaki Oda, Toshiaki Omori, Masato Okada, Masashi Inoue, Hiroyoshi Miyakawa, and Toru Aonishi. Statistical calibration method for physiological ca2+ fluorescence signals. *Australian Journal of Intelligent Information Processing Systems*, 11(1), 2010.

[40] David S Greenberg, Damian J Wallace, Kay-Michael Voit, Silvia Wuertenberger, Uwe Czubayko, Arne Monsees, Takashi Handa, Joshua T Vogelstein, Reinhard Seifert, Yvonne Groemping, et al. Accurate action potential inference from a calcium sensor protein through biophysical modeling. *BioRxiv*, page 479055, 2018.

[41] Artur Speiser, Jinyao Yan, Evan Archer, Lars Buesing, Srinivas C Turaga, and Jakob H Macke. Fast amortized inference of neural activity from calcium imaging data with variational autoencoders. *arXiv preprint arXiv:1711.01846*, 2017.

[42] Daniel Jiwoong Im, Sridhama Prakhya, Jinyao Yan, Srinivas Turaga, and Kristin Branson. Importance weighted adversarial variational autoencoders for spike inference from calcium imaging data. *arXiv preprint arXiv:1906.03214*, 2019.

[43] Vahid Rahmati, Knut Kirmse, Dimitrije Marković, Knut Holthoff, and Stefan J Kiebel. Inferring neuronal dynamics from calcium imaging data using biophysical models and bayesian inference. *PLoS computational biology*, 12(2):e1004736, 2016.

[44] Ryohei Shibue and Fumiyasu Komaki. Deconvolution of calcium imaging data using marked point processes. *PLoS computational biology*, 16(3):e1007650, 2020.

[45] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413–420, 2013.

[46] N Chopin and O Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer-Verlag, 2020.

[47] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[48] Fredrik Lindsten, Michael I Jordan, and Thomas B Schon. Particle gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15:2145–2184, 2014.

[49] Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013.

[50] Henry Lütcke, Felipe Gerhard, Friedemann Zenke, Wulfram Gerstner, and Fritjof Helmchen. Inference of neuronal network spike dynamics and topology from calcium imaging data. *Frontiers in neural circuits*, 7:201, 2013.

[51] Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown, Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou, et al. Caiman an open source tool for scalable calcium imaging data analysis. *Elife*, 8:e38173, 2019.

[52] Marius Pachitariu, Carsen Stringer, Mario Dipoppa, Sylvia Schröder, L Federico Rossi, Henry Dalgleish, Matteo Carandini, and Kenneth D Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv*, 2017.

[53] Giovanni Diana, Thomas TJ Sainsbury, and Martin P Meyer. Bayesian inference of neuronal assemblies. *PLoS computational biology*, 15(10):e1007481, 2019.

[54] Philipp Berens, Jeremy Freeman, Thomas Deneux, Nikolay Chenkov, Thomas McColgan, Artur Speiser, Jakob H Macke, Srinivas C Turaga, Patrick Mineault, Peter Rupprecht, et al. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS computational biology*, 14(5):e1006157, 2018.

[55] Yan Zhang, Márton Rózsa, Yajie Liang, Daniel Bushey, Ziqiang Wei, Jihong Zheng, Daniel Reep, Gerard Joey Broussard, Arthur Tsang, Getahun Tsegaye, et al. Fast and sensitive gcamp calcium indicators for imaging neural populations. *BioRxiv*, 2021.