Accurate action potential inference from a calcium sensor protein through biophysical modeling

David S Greenberg *[1], Damian J Wallace[1], Kay-Michael Voit[1], Silvia Wuertenberger[1], Uwe Czubayko[1], Arne Monsees[1], Takashi Handa[1], Joshua T Vogelstein[2], Reinhard Seifert[3], Yvonne Groemping[1], and Jason ND Kerr†[1]

[1]Department of Brain and Behavior Organization, Research Institute CAESAR, Bonn, Germany
[2]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA
[3]Molecular Sensory Systems, Research Institute CAESAR, Bonn, Germany

Multiphoton imaging of genetically encoded calcium indicators is routinely used to report activity from populations of spatially resolved neurons *in vivo*. However, since the relationship between fluorescence and action potentials (APs) is nonlinear and varies over neurons, quantitatively inferring AP discharge is problematic. To address this we developed a biophysical model of calcium binding kinetics for the indicator GCaMP6s that accurately describes AP-evoked fluorescence changes *in vivo*. The model's physical interpretation allowed the same parameters to describe GCaMP6s binding kinetics for both *in vitro* binding assays and *in vivo* imaging. Using this model, we developed an algorithm to infer APs from fluorescence and measured its accuracy with cell-attached electrical recordings. This approach consistently inferred more accurate AP counts and times than alternative methods for firing rates from 0 to >20 Hz, while requiring less training data. These results demonstrate the utility of quantitative, biophysically grounded models for complex biological data.

Introduction

Linking animal behavior to the underlying neuronal activity requires accurately measuring action potentials (APs) from large populations of neurons simultaneously. Extracellular electrical recordings [65, 84, 117] are widely used to monitor APs in large neuronal populations during behavior [118, 130, 136] but cannot unambiguously assign activity to individual neurons or spatial locations [12, 55]. Alternatively, AP-evoked calcium influx can be detected optically using multiphoton excitation [30] of fluorescent calcium sensors [129]. While this approach can be used to record from both active and silent neurons in anesthetized [70, 120], awake [32, 51] and freely moving animals [114, 143], quantitative readout of neuronal activity requires fluorescence signals to be converted into APs using an inference procedure. For synthetic sensors with a single calcium binding site [129, 57], fluorescence increases from successive APs combine linearly [61, 70, 138] so APs can be inferred by deconvolution [51, 53, 71, 70, 105, 113, 132, 138]. Genetically encoded calcium indicators (GECIs, reviewed in [64]), on the other hand, can label neurons with a specific cell type or projection target and track neuronal populations over days with single cell resolution. In particular GCaMP sensors, derived from calmodulin and green fluorescent protein [90], have been repeatedly improved [1, 122, 123, 127] and GCaMP6s approaches the sensitivity of synthetic sensors [18]. However, inferring APs from GCaMP fluorescence remains challenging and unreliable due to AP responses that and are complex, nonlinear and variable over neurons [1, 18, 64]. This complexity, nonlinearity and variability can be qualitatively explained by cooperativity across GCaMP's 4 calmodulin-derived binding sites, slow and multiphasic kinetics of the indicator, dependence of AP-evoked fluorescence amplitude and shape on total GCaMP concentration [64] and out-of-focus signals from background structures that can increase baseline fluorescence differently for each neuron. However, incorporating these effects into a quantitative description of GCaMP-expressing neurons requires many unknown parameters to be determined. Instead, AP inference methods have been proposed based on phenomenological modeling [29], deep learning with many free parameters but no interpretable model [126] and thresholding [32, 56]. While some of these approaches outperform linear deconvolution on GCaMP data, they produce inaccurate results for many neurons.

Here we describe a sequential binding model (SBM) linking APs to GCaMP6s fluorescence based on biophysical principles. This model quantitatively describes GCaMP6s fluorescence *in vivo* and *in vitro* and allows for more accurate AP inference than previous methods. Given an AP sequence, the SBM describes the binding kinetics that determine concentrations over time for free calcium, calcium bound to endogenous buffers and GCaMP6s binding states with 0 to 4 calcium ions. Solving the resulting differential equations provides GCaMP6s binding state concentrations at each measurement time, which we use to directly predict fluorescence. We fit the SBM to a library of combined optical/electrical recordings from neurons in mouse visual cortex, showing it can capture the AP-fluorescence relationship despite nonlinearity and variability over neurons. We also used the SBM to describe calcium-GCaMP6s interactions *in vitro* by applying the same biophysical framework to fluorescence spectroscopy, isothermal titration calorimetry and stopped-flow fluorescence experiments. Globally fitting the SBM to data from all three

*david.greenberg@caesar.de
†jason.kerr@caesar.de

in vitro binding assays yielded parameters that accurately predicted AP-evoked fluorescence *in vivo*. To use the SBM for AP inference, we developed a GPU-based sequential Monte Carlo algorithm [49] and evaluated accuracy on held-out data. SBM-based AP inference outperformed previous algorithms [29, 32, 105, 126, 132], producing more accurate firing rates, AP counts and AP times, while modeling the effect of AP discharge on fluorescence in interpretable biophysical terms.

Results

AP-evoked GCaMP6s fluorescence *in vivo*

To characterize how AP discharge sequences give rise to GCaMP6s fluorescence changes, we imaged neuronal populations expressing virally transfected GCaMP6s in L2/3 mouse visual cortex while juxtasomally recording one neuron's spontaneous APs (Figure 1a, n = 26 neurons, 10 animals, 15464 APs, 9.4 hours total, mean firing rates 0.008 to 12.5 Hz, median 0.16 Hz). Based on previous observations in L2/3 mouse visual cortex [93], we identified 4 recordings as putative interneurons from the electrophysiological data based on higher mean spontaneous firing rates, AP waveforms with larger afterhyperpolarizations (AHPs) and shorter peak-to-AHP latencies than in pyramidal neurons. We extracted fluorescence from each neuron's somatic cytosol using a semi-automated algorithm (Figure 1 - figure supplement 1, methods).

These recordings exhibited a wide variety of AP discharge patterns, evoking a diverse range of fluorescence signals (Figure 1b). In pyramidal neurons, isolated single APs and brief AP bursts evoked transient fluorescence increases (Figure 1b, upper, peak latency 173 ± 70 ms), but during periods of prolonged AP discharge fluorescence did not return to baseline between bursts (Figure 1b, middle and lower). Isolated single APs were sufficient to increase fluorescence in pyramidal neurons [18], but response shape exhibited considerable variability across neurons (average 1-AP peak ranged from 3.0%-18.7% $\Delta F/F_0$, mean 7.6%, s.d. 4.7%, Figure 1c) that could not be explained by variation across animals or differences in expression time (Figure 1 - figure supplement 2). The fluorescence evoked by AP burst of 2 or more APs grew supralinearly with the number of APs (Figure 1d, burst duration $<$ 200 ms), with the peak fluorescence evoked by 2 APs ranging from 2-7 times the 1-AP peak (Figure 1e, mean 3.5, s.d 1.2, n = 22). In interneurons, fluorescence increased during periods of more frequent AP discharge (Figure 1b, lower) but single APs did not evoke discernible fluorescence changes (maximum peak amplitude 0.2% $\Delta F/F_0$, Figure 1 - figure supplement 3). These results show that accurately inferring APs from GCaMP6s fluorescence requires methods capable of dealing with both nonlinearity and variability over neurons.

We used this dataset to test several existing AP inference algorithms: the phenomenological model-based method MLspike [29], a threshold-based approach (thr-σ) [32] and a deep-learning method [126] trained on the current data (c2s-t) or using published parameters (c2s-s). While these approaches successfully detected single APs and bursts in some cases, for a majority of neurons either less than half the APs were detected or more than half of inferred APs were false positives, for every method tested (Figure 1 - figure supplement 4). These results show that inferring AP sequences from GCaMP6s fluorescence signals remains a challenging open problem. We therefore developed a quantitative model linking AP discharge to fluorescence in GCaMP6s-expressing neurons, to better understand the relationship between the two and to provide a mathematical foundation for improved AP inference.

A sequential binding model for GCaMP6s-expressing neurons

To quantitatively link APs to GCaMP6s fluorescence, we constructed a biophysical model based on the chain of causal effects through which APs cause fluorescence changes. First, the AP depolarizes the somatic membrane and calcium ions enter the neuron through ion channels [5, 17, 40, 60, 107, 124] (reviewed in [15, 62]). This temporary increase in cytosolic calcium concentration is further shaped by physiological processes, including buffering by endogenous proteins [60, 80, 111, 116] and extrusion [60, 80, 110, 111, 115, 116]. At the same time, some calcium ions bind reversibly to GCaMP6s, which contains a calmodulin-derived protein domain with 4 binding sites [90]. Calcium binding to GCaMP6s increases its fluorescence [18], with proposed mechanisms based on conformational changes that alter fluorophore protonation [133, 2, 59, 122, 7]. We reasoned that since the individual parts of this system have been extensively studied and quantified, by modeling them together we could quantitatively predict fluorescence from APs and infer APs from fluorescence (Figure 2a). Furthermore, our combined optical/electrical recordings could be used to fit model parameters, assess the quality of model fits to data and evaluate the accuracy of model-based AP inference.

We therefore developed a sequential binding model (SBM) of GCaMP6s-expressing neurons (Figure 2b). The core idea of the SBM is to model each individual binding step as calcium binds the four sites of GCaMP6s:

$$\begin{aligned} 4\mathrm{Ca}^{2+} + \quad \mathrm{GCaMP6s} &\rightleftharpoons \\ 3\mathrm{Ca}^{2+} + \mathrm{CaGCaMP6s} &\rightleftharpoons \\ 2\mathrm{Ca}^{2+} + \mathrm{Ca_2GCaMP6s} &\rightleftharpoons \\ \mathrm{Ca}^{2+} + \mathrm{Ca_3GCaMP6s} &\rightleftharpoons \\ \mathrm{Ca_4GCaMP6s} & \end{aligned} \tag{1}$$

These four reversible binding reactions involve free calcium and five possible GCaMP6s binding states, with 0 to 4 ions bound. We modeled these reactions using standard mass action kinetics [87, 137]: each forward rate is proportional to the concentrations of free calcium and the previous binding state, while each backward rate is proportional to the concentration of the next binding state:

$$r_j^+ = k_j^+[\mathrm{Ca}^{2+}][\mathrm{Ca}_{j-1}\mathrm{GCaMP6s}] \tag{2}$$

$$r_j^- = k_j^-[\mathrm{Ca}_j\mathrm{GCaMP6s}] \tag{3}$$

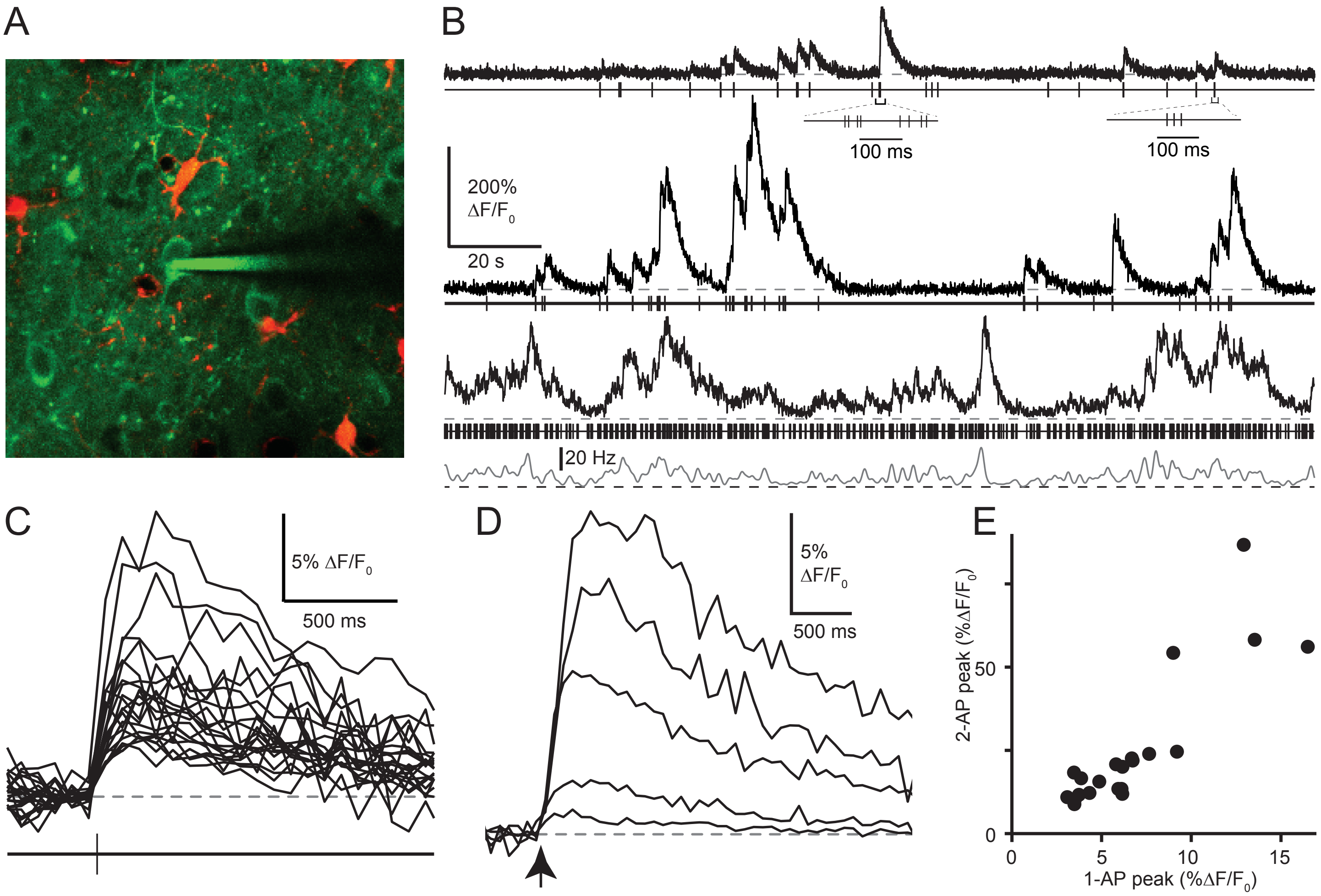

Figure 1: **(A)** Two-photon image of neuronal population expressing GCaMP6s (green) in L2/3 of mouse visual cortex, with astrocytes stained using sulforhodamine 101 (red) and juxtasomal electrical recording of APs in a single neuron. **(B)** Simultaneous electrical recording of AP times and imaging of neuronal fluorescence in 3 different neurons. For the third, an interneuron, firing rate was calculated by Gaussian filtering of the AP sequence with $\sigma = 500$ ms. Dashed lines indicate 0% $\Delta F/F_0$ and 0 Hz. **(C)** Average fluorescence evoked by isolated single APs (no APs in the preceding 5.5 s) in 22 pyramidal neurons. **(D)** Average fluorescence evoked by single APs and bursts of 2-5 APs in a pyramidal neuron. Bursts were defined as groups of APs spanning < 200 ms with no other APs 5.5 s before the first AP in the burst. **(E)** Comparison of peak amplitude evoked by 1 vs. 2 APs; same neurons as in (C). Peak amplitude was calculated as the maximum of the average fluorescence response to 1 or 2 APs for each neuron.
Figure 1–Figure supplement 1. Removal of extraneous signals with a feature extraction algorithm.
Figure 1–Figure supplement 2. Peak AP-evoked GCaMP6s fluorescence as a function of expression time.
Figure 1–Figure supplement 3. Single APs do not evoke fluorescence increases in interneurons.
Figure 1–Figure supplement 4. Accuracy of existing AP inference methods.

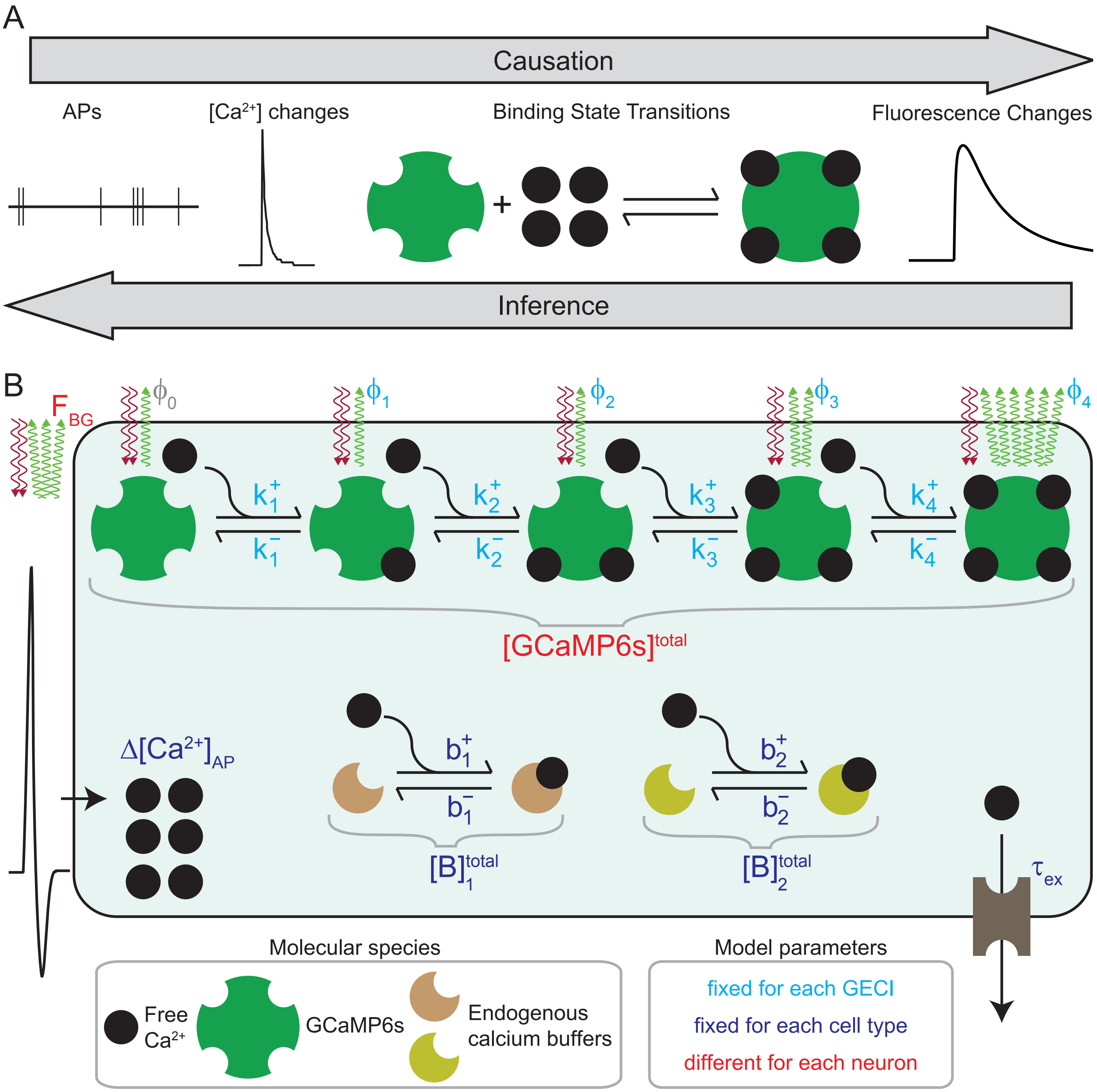

Figure 2: **(A)** Conceptual illustration showing biophysical modeling used to quantitatively link fluorescence and APs. AP discharge, calcium kinetics, calcium-dependent binding state transitions and the resulting fluorescence changes are linked by a causal biophysical model (left to right) and inference procedure (right to left). **(B)** Diagram showing the sequential binding model (SBM). Each AP causes a calcium influx that increases cystosolic free calcium concentration by $\Delta[\mathrm{Ca}^{2+}]_{\mathrm{AP}}$. Binding of calcium to GCaMP6s then proceeds with on-rates k_j^+ and off-rates k_j^-, and binding to endogenous buffers with on-rates b_j^+ and off-rates b_j^-. $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ denotes the total GECI concentration in the neuron and the total concentration of each endogenous buffer is $[\mathrm{B}]_\ell^{\mathrm{total}}$. The calcium dependence of GCaMP6s fluorescence is described by brightness values ϕ_j for each binding state $\mathrm{Ca}_j\mathrm{GCaMP6s}$. The total fluorescence predicted by the SBM is calculated (eq. 6-7) based on the current GCaMP6s binding state concentrations, the brightness ϕ_j of each binding state and the background fluorescence F_{BG}.

Figure 2–Figure supplement 1. Global rate equation describing mass action kinetics and extrusion.

The constants of proportionality are given by the kinetic rate constants k_j^+ (on-rates) and k_j^- (off-rates) for each binding step, which together describe calcium's interactions with GCaMP6s. The SBM includes a different total indicator concentration $[\text{GCaMP6s}]^{\text{total}}$ for each neuron to incorporate varying levels of indicator protein expression [64].

In order to link APs and fluorescence, however, the SBM must describe not only the interaction between calcium and GCaMP6s but also the effects of AP discharge, endogenous buffering and extrusion on calcium concentration. We modeled AP discharge as an increase in free calcium concentration by $\Delta[\text{Ca}^{2+}]_{\text{AP}}$, and binding to endogenous buffers with two additional reactions:

$$\text{Ca}^{2+} + \text{B}_1 \rightleftharpoons \text{CaB}_1 \tag{4}$$

$$\text{Ca}^{2+} + \text{B}_2 \rightleftharpoons \text{CaB}_2$$

where B_1 and B_2 are endogenous buffers with on-rates b_1^+ and b_2^+, off-rates b_1^- and b_2^- and total concentrations $[\text{B}]_1^{\text{total}}$ and $[\text{B}]_2^{\text{total}}$. Extrusion was included as the reaction $\text{Ca}^{2+} \rightarrow \emptyset$ involving only calcium ions, with the rate given by:

$$r_{\text{ex}} = \left([\text{Ca}^{2+}] - [\text{Ca}^{2+}]_{\text{rest}}\right)/\tau_{\text{ex}} \tag{5}$$

where τ_{ex} is the extrusion time constant and $[\text{Ca}^{2+}]_{\text{rest}}$ is the neuron's resting calcium concentration. Because GCaMP6s, the endogenous buffers and the extrusion process compete for calcium ions, the SBM's calcium dynamics are shaped by all three together.

The SBM thus contains 7 reactions and 10 molecular species. Given the current concentrations of these 10 species, the rates of change of all concentrations are given by a global rate equation (Figure 2 - figure supplement 1). The rate equation can therefore be used to calculate the evolution of all molecular species' concentrations over time, which the SBM then uses to predict the neuron's fluorescence at the time of each measurement. We modeled the total fluorescence F_{cyt} arising from GCaMP6s located in the neuron's somatic cytosol by adding up the fluorescence arising from each binding state of the indicator:

$$F_{\text{cyt}} = \sum_j \phi_j [\text{Ca}_j\text{GCaMP6s}] \tag{6}$$

where ϕ_j is the brightness of Ca_jGCaMP6s. For example, when each ϕ_j is greater than the previous value ϕ_{j-1}, GCaMP6s fluorescence increases for every calcium ion bound. Alternatively, if ϕ_j is the same for all binding states with less than four calcium ions bound, only the final binding step increases fluorescence. In this way the SBM model class can describe spectroscopically silent steps, in which binding of a calcium ion does not change the protein's fluorescence properties.

Since under *in vivo* imaging conditions neuronal fluorescence can exhibit time-varying baseline fluorescence [51, 70] as well as contamination from tissue autofluorescence [125], neuropil background signals [18, 70] and misfolded indicator protein [23, 90], the SBM also incorporated a drifting fluorescence baseline F_{BL} and a constant background fluorescence term F_{BG}. Combining these effects (see methods), the observed fluorescence predicted by the SBM is:

$$F = F_{\text{BL}} \left(\frac{F_{\text{cyt}} + F_{\text{BG}}}{F_{\text{cyt}}^{\text{eq}} + F_{\text{BG}}} \right) \tag{7}$$

where $F_{\text{cyt}}^{\text{eq}}$ is the equilibrium value of F_{cyt} when no APs are discharged and $[\text{Ca}^{2+}] = [\text{Ca}^{2+}]_{\text{rest}}$. Since rescaling the SBM's concentration units along with F_{BG} does not change the predicted fluorescence, we fixed $[\text{Ca}^{2+}]_{\text{rest}}$ to 50 nM in accordance with previous measurements *in vitro* [60, 80, 116] and *in vivo* [142]. Consequently, all concentration values in the SBM can alternatively be interpreted as multiples of the resting calcium concentration $[\text{Ca}^{2+}]_{\text{rest}}$.

Overall, the SBM contains 22 different free parameters (Table 1): 12 GECI-specific parameters describing the calcium-binding and fluorescence properties of GCaMP6s, 8 cell type-specific parameters describing calcium influx, buffering and extrusion and two remaining parameters, F_{BG} and $[\text{GCaMP6s}]^{\text{total}}$, that can vary for each neuron. Having designed the SBM to quantitatively link APs to fluorescence based on established biophysical principles, we next fit its parameters to our *in vivo* GCaMP6s dataset.

Fitting the SBM to combined optical / electrical recordings *in vivo*

We fit the SBM to our *in vivo* dataset by adjusting all its parameters to give the best possible predictions of fluorescence from AP sequences (Figure 3a-d). For this purpose, we first estimated F_{BL} by interpolating fluorescence values from silent periods without APs (Figure 3 - figure supplement 1, see methods; interneurons lacked silent periods and were excluded from fitting, but included when testing AP inference below). We next initialized the SBM parameters to default starting values, using estimates from experimental studies where available (e.g. calcium influx per AP and extrusion rate, full details in methods). To predict fluorescence from a given AP sequence, we started at the beginning of the recording and moved forward in time, incrementing free calcium by $\Delta[\text{Ca}^{2+}]_{\text{AP}}$ at each AP while solving the rate equation to determine the concentration over time for all molecular species (Figure 3c, time step 10 ms). We then added up the contributions of all 5 GCaMP6s binding states (eq. 6-7) to generate a prediction of the neuron's fluorescence (Figure 3b, orange) and compared the result to observed fluorescence values (Figure 3b, black). While calculating fluorescence predictions AP sequences in this way, we then adjusted the SBM parameters to minimized the mismatch between predicted and observed fluorescence values across all pyramidal neurons' recordings ($n = 22$), using iterative optimization of with multiple initializations (Figure 3 - figure supplement 2, methods). This fitting procedure resulted in parameters (Table 1) that closely predicted the measured fluorescence signals (Figure 3a-d). Simplified versions of the SBM (Figure 3 - figure supplement 3) without endogenous buffers or variation across neurons in F_{BG} or $[\text{GCaMP6s}]^{\text{total}}$ predicted fluorescence signals less accurately for most neurons (Figure 3 - figure supplement 4, $p < 0.005$, sign tests). Using more than 2 buffers, more elaborate extrusion mechanisms or a shorter time step did not improve fit quality ($p > 0.05$).

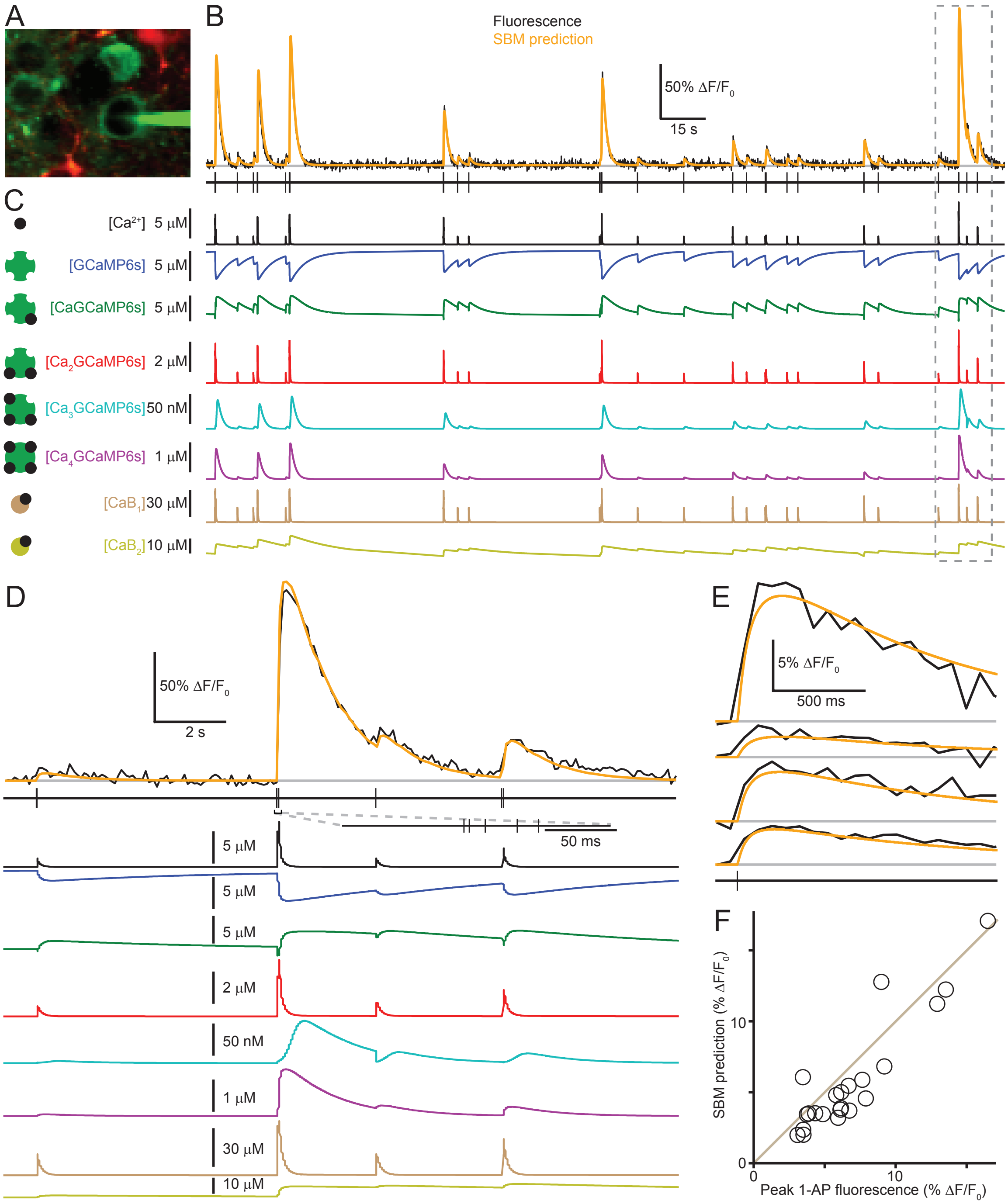

Figure 3: **A** Two-photon image of neuronal population expressing GCaMP6s (green) in L2/3 of mouse visual cortex, with astrocytes stained using sulforhodamine 101 (red) and juxtasomal electrical recording of APs in a pyramidal neuron. **(B)** GCaMP6s fluorescence (black, upper) and simultaneous electrical recording of APs (lower) from the neuron shown in (A). The baseline fluorescence values used for model fitting are shown in gray, and the SBM's prediction after globally fitting all model parameters to the full dataset is shown in orange. **(C)** Time-varying concentrations of free calcium, GCaMP6s binding states and calcium-bound endogenous buffers calculated by the SBM to generate the prediction in (B). **(D)** Expansion of the period indicated by the dashed box in (B-C). **(E)** Average 1-AP fluorescence increase in four pyramidal neurons (black) and SBM fits (orange). **(F)** Peak fluorescence increase following single APs compared to SBM-fit values for all pyramidal neurons.

Figure 3–Figure supplement 1. Estimation of drifting baseline fluorescence with known AP times.
Figure 3–Figure supplement 2. Convergence of SBM parameters while fitting *in vivo* data.
Figure 3–Figure supplement 3. Comparison of SBM fits to data with increasing model complexity.
Figure 3–Figure supplement 4. Comparison of SBM variants over neurons.
Figure 3–Figure supplement 5. SBM simulations of fluorescence responses to AP discharge.
Figure 3–Figure supplement 6. Free calcium after AP discharge depends on $[GCaMP6s]^{total}$ in SBM simulations.
Figure 3–Figure supplement 7. Response nonlinearity depends on $[GCaMP6s]^{total}$ in SBM simulations.
Figure 3–Figure supplement 8. Endogenous buffers shape AP-evoked fluorescence in SBM simulations.
Figure 3–Figure supplement 9. Variation of SBM fluorescence predictions over multiple parameter sets.

Table 1: SBM parameters fit *in vivo*

Symbol	Description	Best-fit value	Range of fit values##	Units
$\Delta[\mathrm{Ca}^{2+}]_{\mathrm{AP}}$	1-AP calcium influx	20.2	2.3 - 32.6	µM
τ_{ex}	Extrusion time constant	7.5	5.5 - 64.9	ms
$[\mathrm{Ca}^{2+}]_{\mathrm{rest}}$	Resting free calcium concentration	50**		nM
k_1^+	First GCaMP6s on-rate	2.5	1.6 - 52.1	$\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$
k_2^+	Second GCaMP6s on-rate	16.9	1.2 - 20.2	$\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$
k_3^+	Third GCaMP6s on-rate	1.1	0.5 - 636	$\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$
k_4^+	Fourth GCaMP6s on-rate	1069	54 - 1766	$\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$
k_1^-	First GCaMP6s off-rate	0.1	0.1 - 961	s^{-1}
k_2^-	Second GCaMP6s off-rate	205	2 - 984	s^{-1}
k_3^-	Third GCaMP6s off-rate	11.8	2.9 - 515	s^{-1}
k_4^-	Fourth GCaMP6s off-rate	5.8	1.4 - 35.1	s^{-1}
ϕ_1/ϕ_0	CaGCaMP6s/GCaMP6s brightness ratio*	1.0	1.0 - 2.6	***
ϕ_2/ϕ_0	$\mathrm{Ca_2}$GCaMP6s/GCaMP6s brightness ratio*	1.0	1.0 - 42.9	***
ϕ_3/ϕ_0	$\mathrm{Ca_3}$GCaMP6s/GCaMP6s brightness ratio*	1.0	1.0 - 81.0	***
ϕ_4/ϕ_0	$\mathrm{Ca_4}$GCaMP6s/GCaMP6s brightness ratio*	81.0	33.1 - 81.0	***
$[\mathrm{GCaMP6s}]^{\mathrm{total}}$	Total GCaMP6s concentration	7.4 ****		µM
$F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}}$	Background/cytosolic fluorescence at rest	2.5 ****		***
b_1^-/b_1^+	Fast buffer dissociation constant	3.4	2.2 - 100	µM
$1/b_1^-$	Fast buffer time constant	$< 1.0^{\#}$		ms
$[B_1]^{\mathrm{total}}$	Fast buffer concentration	64	0.0 - 154	µM
b_2^-/b_2^+	Slow buffer dissociation constant	0.59	0.1 - 3.6	µM
$1/b_2^-$	Slow buffer time constant	17	12 - 30	s
$[B_2]^{\mathrm{total}}$	Slow buffer concentration	119	23 - 129	µM

*For two photon excitation at 920 nm *in vivo*

**Fixed to this value based on previous reports, not fit to data

***Unitless ratio

****Maximum likelihood estimate for the mode of a log-normal fit to parameter's distribution over neurons

#Model fitting procedures results in the lowest allowed value of 1 ms for the fast buffer time constant.

Range over multiple SBM parameter sets fit while excluding each neuron's data in turn.

We next examined how individual aspects of the SBM contributed to predicted fluorescence changes using simulations in which some model parameters were perturbed or set to zero. Changing $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ affected the simulated 1-AP fluorescence peak height, 2-AP peak height and peak timing (Figure 3 - figure supplement 5), as well as the simulated peak free calcium concentration after AP discharge (Figure 3 - figure supplement 6). This is consistent with previous observations that synthetic indicators at high concentrations contribute significantly to Ca^{2+} buffering [60, 80, 116], as well as a theoretical study predicting that GECI expression levels influence AP response amplitude and shape [64]. These simulations also reproduced the nonlinearity of GCaMP6s's AP responses, with 2 APs evoking 4 to 7 times the fluorescence increases observed for 1 AP depending on the value of $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ (Figure 3 - figure supplement 7a-b), so that nonlinearity can vary over neurons as observed *in vivo* (Figure 1e). Simulations omitting one of the endogenous buffers (Figure 3 - figure supplement 8) showed that fluorescence responses to 1 or 2 APs were shaped by a fast endogenous buffer (time constant $\tilde{1}$ ms), but the return to baseline after long bursts was shaped by a slow buffer (> 10 s). Overall, by incorporating variation over neurons in $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG} together with endogenous buffering and extrusion properties that were fixed for all neurons, the SBM was able to capture variability in the rising and falling phases of observed AP responses (Figure 3e). This allowed the SBM to closely predict the peak fluorescence evoked by single APs over the entire range of peak amplitudes observed (Figure 3f, r = 0.91, n = 22 neurons).

While the SBM was able to predict fluorescence changes from AP sequences, this does not guarantee that a unique set of best-fitting parameter values can be unambiguously identified from the available data. On the contrary, given the SBM's complexity its parameters may not be fully identifiable from our data, and multiple SBM parameter sets might predict the observed fluorescence signals equally well from the AP sequences. This uncertainty in parameter values is common in detailed biophysical [95] and biochemical models with many parameters [45, 54, 108]. Therefore, to test how tightly SBM parameters can be constrained given our *in vivo* dataset, we repeatedly fit the SBM while excluding one neuron at a time to generate multiple SBM parameter sets, reasoning that if the parameters are identifiable given our *in vivo* dataset then their optimal values should not be significantly altered when removing a single neuron's data. Comparing the resulting parameter sets showed that most parameters ranged over about 1 order of magnitude, suggesting that they cannot be precisely determined given the available data (Table 1). Nonetheless, when tested on the same AP sequence these different SBM parameter sets produced nearly identical predictions of fluorescence (Figure 3 - figure supplement 9a). The concentrations over time of all molecular species were also highly correlated over SBM parameter sets (minimum correlation 0.98), although the absolute scale of the concentrations differed (Figure 3 - figure supplement 9b). These results show that fitting the SBM to *in vivo* data unambiguously identifies the quantitative mapping from APs to fluorescence, but multiple SBM parameter sets are consistent with the same mapping given our *in vivo* data.

SBM analysis of *in vitro* binding assays

The SBM describes the interaction of calcium and GCaMP6s as a sequence of protein-ligand binding reactions, allowing quantitative data analysis through the biophysical framework of mass action kinetics. To test whether this framework could describe the same reaction sequence under controlled conditions, we carried out *in vitro* binding assays to measure GCaMP6s-calcium interactions outside of neurons. This allowed us to study these reactions beyond the physiological concentration ranges for GCaMP6s and calcium and using a more diverse set of measurement techniques. We used fluorescence spectroscopy to characterize GCaMP6s's excitation spectrum at equilibrium (Figure 4a), and how the spectrum changes as calcium concentration is increased. We used isothermal titration calorimetry (ITC) to detect the heat absorbed or released as calcium was titrated into a GCaMP6s solution (Figure 4b), providing readout of calcium binding that does not depend on fluorescence changes. We used stopped-flow fluorimetry to measure fluorescence changes over time after GCaMP6s was rapidly mixed with calcium or a calcium-chelator (Figure 4c). We globally fit the SBM to all data simultaneously ($n = 3$ spectra, 8 ITC experiments, 19 stopped-flow traces, 4 pools of purified GCaMP6s) to determine kinetic and equilibrium binding properties for GCaMP6s, adjusting all model parameters to achieve the best possible prediction of spectra, heat and stopped-flow fluorescence data (Figure 4d-g, Figure 4 - figure supplement 1, details in methods). This global fit resulted in SBM parameters (Table 2) that closely predicted data from all three binding assays. Similarly to SBM fits on *in vivo* fluorescence data, we found that a range of values for each SBM parameter were consistent with *in vitro* measurements, possibly due to spectroscopically silent binding steps or reactions too fast to resolve using the methods employed. Nonetheless, these results show that the SBM provides a quantitative account of calcium-GCaMP6s interactions in all three binding assays.

We next examined whether the SBM rate constants obtained from *in vitro* binding assays could accurately predict AP-evoked fluorescence changes *in vivo*, as compared to parameters determined exclusively from *in vivo* recordings. To test this, we fixed the *in vitro*-derived rate constants and fit all remaining parameters to *in vivo* fluorescence signals. This resulted in fluorescence predictions similar to those arising from *in vivo*-derived rate constants (Figure 4h). Since this procedure fit the SBM to the same data with fewer free parameters, larger errors would be expected than when fitting all parameters on *in vivo* data alone. However, if the rate constants obtained *in vitro* can accurately describe calcium binding in neurons, then the fitting errors for *in vitro*- vs. *in vivo*-derived rate constants should be approximately equal, and should fall on a line with a slope of one. Comparing each neuron's r.m.s. error value for the *in vitro*- vs. *in vivo*-derived rate constants in this way, we observed that *in vitro*-derived rate constants produced only slightly higher error values (Figure 4i, r.m.s. error 5.9 ± 3.0 vs. $6.3 \pm 3.2\%$ $\Delta F/F_0$), and that the errors fell on a line with a slope close to one (1.06, 95% confidence interval 1.02 - 1.09). Fits using *in vitro*-derived rate constants also predicted *in vivo* fluorescence more accurately than simplified versions of the SBM that omitted endogenous buffers or variation in F_{BG} or $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ over neurons but were fit to *in vivo* data alone (Figure 3 - figure supplement 4). These results show that *in vitro*-derived rate constants lie within the range of SBM parameter sets capable of accurately describing GCaMP6s fluorescence *in vivo*, and that a single set of rate constants can accurately describe both scenarios when the different *in vitro* and cytosolic environments are taken into account. In contrast, previous phenomenological fitting approaches based on rise times and Hill exponents have required different parameter values *in vivo* and *in vitro* [61].

AP inference with the SBM

We next applied the SBM to the problem of inferring AP times and neuron-specific parameters ($[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG}) from fluorescence data alone. To identify the AP-sequence most consistent with the fluorescence data, we developed a sequential Monte Carlo algorithm [49] (SMC, or particle filtering, reviewed in [35]). SMC, a data-driven simulation technique that has also been used with synthetic indicators [131], generates many simulations or particles whose predictions are compared to observed data (Figure 5a). For the SBM each particle consisted of a simulated AP sequence along with time courses for free calcium, GCaMP6s binding states and baseline fluorescence (Figure 5b). The algorithm was advanced forward in time by randomly extending each particle's AP sequence and baseline fluorescence while solving the rate equation to update all molecular species' concentrations (Figure 2 - figure supplement 1), using *in vivo*-derived rate constants as these provided slightly tighter fits to the data (Figure 4i). At each fluorescence measurement the particles' predictions were compared to the observed fluorescence value (Figure 5a) to calculate probability weights (Figure 5c, methods), and particles were then eliminated, retained or copied multiple times with probabilities determined by their weights. Particles whose predictions did not match the data were more likely to be eliminated, so those with incorrect AP sequences disappeared after a small number of measurements. After the SMC algorithm traversed all the fluorescence measurements, a fixed-lag smoother [73] combined the particles and weights at each time point to compute the probability of spiking over time given the fluorescence data. As the number of particles increases, SMC methods converge to unbiased Bayesian estimators of the hidden state variables (APs, binding states and baseline) and of the data's likelihood given model parameters [27].

Developing an SMC algorithm for the SBM required several new computational techniques to improve the speed and

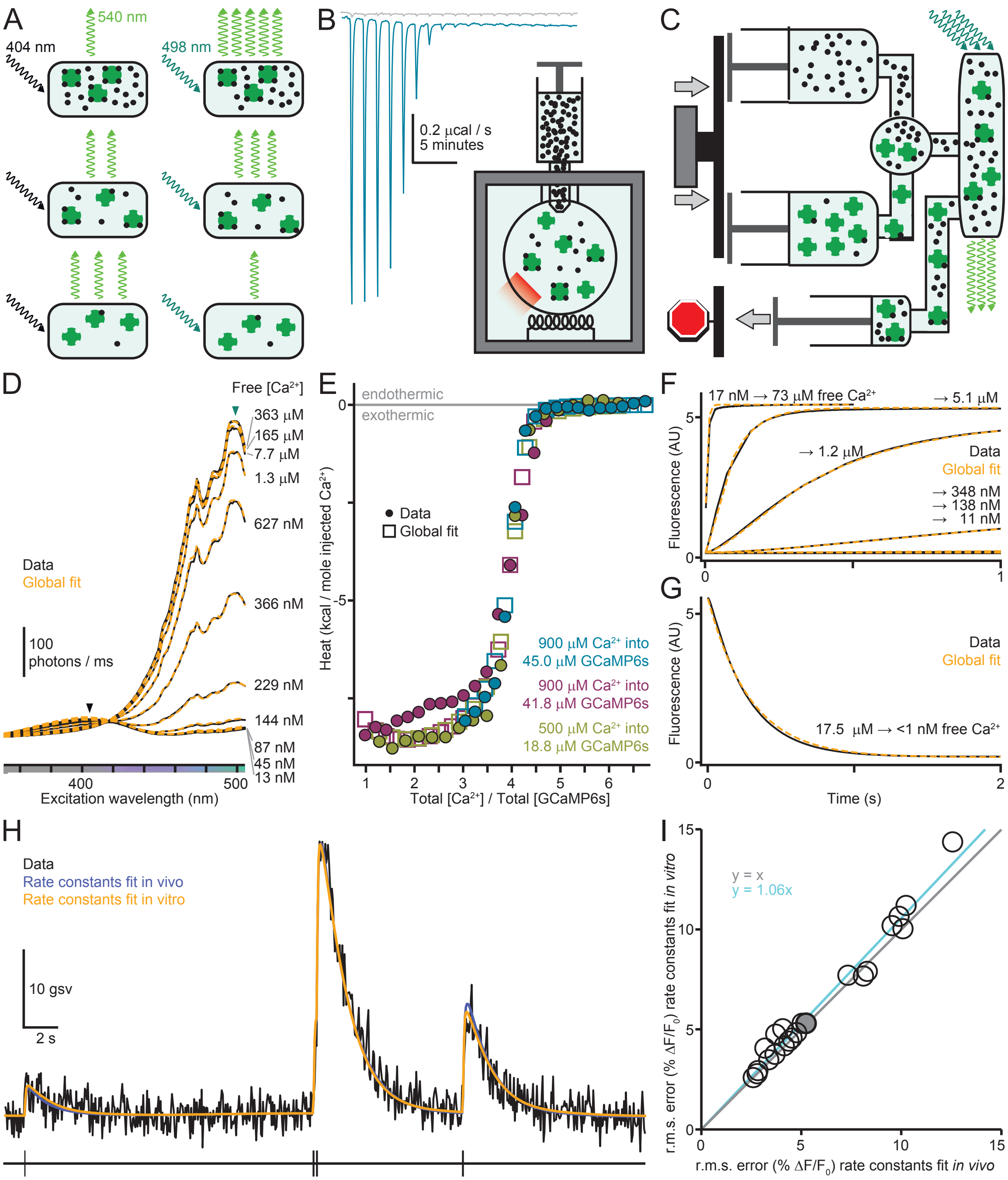

Figure 4: **(A)** Fluorescence spectroscopy of GCaMP6s; fluorescence showed the largest increase with [Ca^{2+}] at 498 nm, and the largest decrease at 404 nm. **(B)** Isothermal titration calorimetry experiment, with repeated injection of concentrated Ca^{2+} into GCaMP6s and measurement of the heat absorbed or released upon binding. Raw calorimetric data (turquoise curve) show a net exothermic process for each injection (negative peaks) until Ca^{2+}-saturation of GCaMP6s. Heats of dilution for Ca^{2+} (gray) were measured by injecting Ca^{2+} into low-calcium pH buffer (30 mM MOPS with 100 mM Kcl, pH 7.2). **(C)** Stopped-flow fluorescence measurements of GCaMP6s Ca^{2+}-binding kinetics. BAPTA-buffered solutions of GCaMP6s and calcium are rapidly mixed and fluorescence is then measured over time. **(D)** Fluorescence excitation spectra of GCaMP6s (black, emission wavelength 540 nm) and global SBM fit (orange) for a range of calcium concentrations and 1.6 mM BAPTA. Arrows indicate 404 and 498 nm. **(E)** Integrated peak heats (circles) for 3 ITC experiments and global fit (squares) vs. the ratio of total Ca^{2+} and GCaMP6s after each injection. **(F)** Kinetics of Ca^{2+} binding to GCaMP6s measured using a stopped-flow device. 0.9 µM GCaMP6s in 37 µM BAPTA was mixed with 1.7 mM BAPTA and total Ca^{2+} ranging from 0-1.9 mM. **(G)** Kinetics of calcium release from GCaMP6s. 0.9 µM GCaMP6s with 21.1 µM total Ca^{2+} was mixed with 8 mM BAPTA. **(H)** Electrically detected APs (upper) and simultaneously recorded GCaMP6s fluorescence (black, lower) from a L2/3 mouse visual cortical pyramidal neuron, with model fit to *in vivo* data alone (blue) and model with rate constants fit to *in vitro* data and other parameters fit *in vivo* (orange). **(I)** Root-mean-square error for SBM with rate constants fit *in vitro* vs. *in vivo* (n = 22 pyramidal neurons). Unity line is shown in gray and linear regression in cyan.

Figure 4–Figure supplement 1. Contributions of each GCaMP6s binding state to *in vitro* data fit.

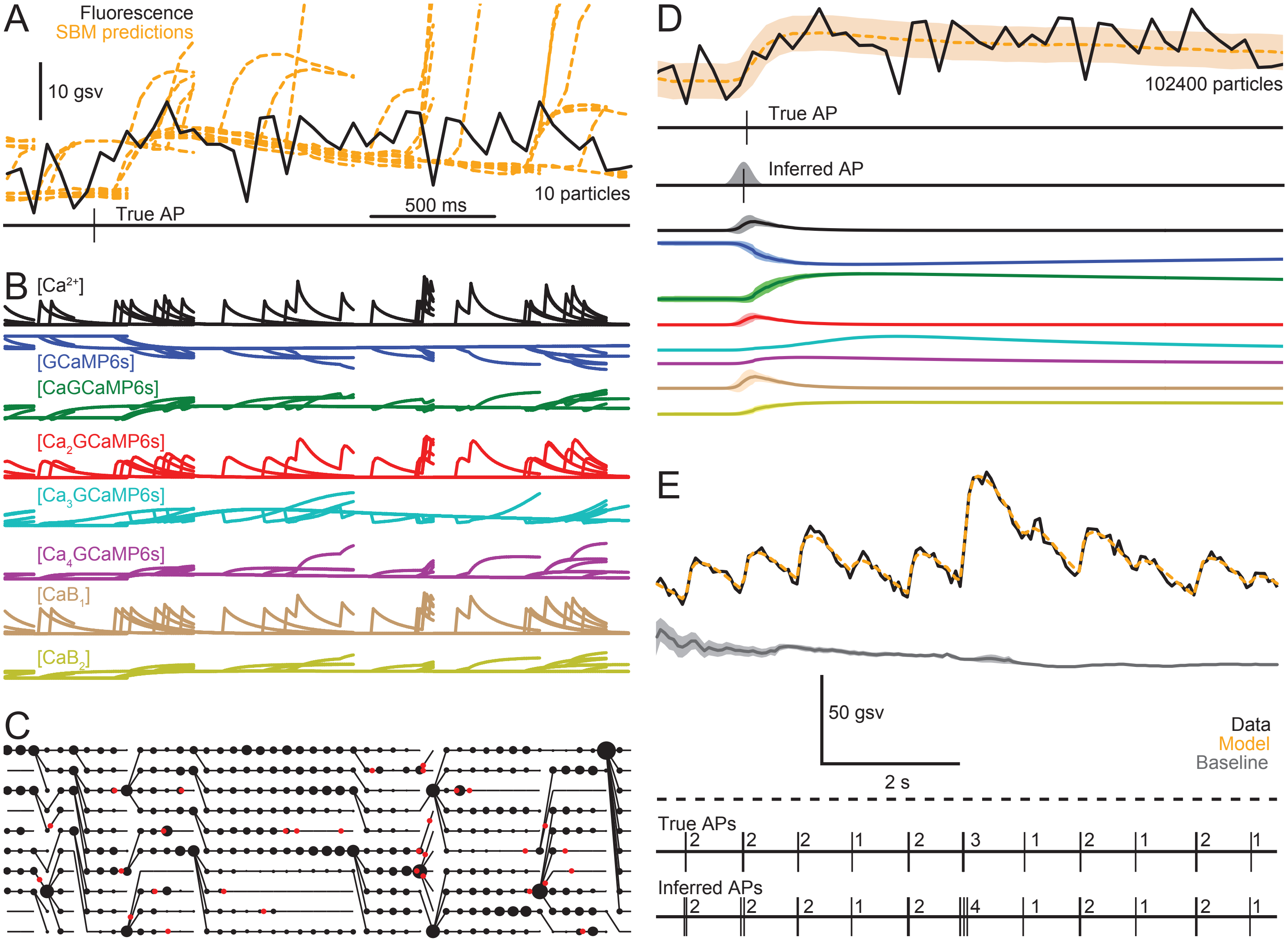

Figure 5: **(A)** GCaMP6s fluorescence (black, upper) from a L2/3 mouse visual cortical pyramidal neuron, with an electrically recorded AP (lower) and predicted fluorescence from ten SBM simulations used in a sequential Monte Carlo (SMC) algorithm (orange). **(B)** Time-varying concentrations of free calcium, GCaMP6s binding states and calcium-bound endogenous buffers for the SBM simulations in (A). Scalebar in (A): 4 µM Ca^{2+}, 10 µM GCaMP6s, 10 µM CaGCaMP6s, 2.5 µM Ca_2GCaMP6s, 20 nM Ca_3GCaMP6s, 750 nM Ca_4GCaMP6s, 40 µM CaB_1 and 15 µM CaB_2. **(C)** Probability weights (areas of black dots) for the SMC filtering distribution, calculated using the simulations and data in (A). Black lines indicate particle ancestry for the SMC algorithm; red dots indicate AP discharge times in the SBM simulations. **(D)** Inferred AP time with temporal uncertainty (upper) from an SMC algorithm with 102400 particles; same data as in (A-C). Means and standard deviations (lower) are shown for the SMC smoothing distributions of all molecular species' concentrations given fluorescence data in (A) and 102400 particles, along with the posterior mean and standard deviation of de-noised fluorescence (orange). Data units are scaled as in (A-B). **(E)** SMC/SBM inference of AP times from 9 seconds of data during which fluorescence never decays to baseline (dashed line indicates 0 fluorescence).

Figure 5–Figure supplement 1. Numerical integration of the SBM rate equation with a custom ODE solver.

Figure 5–Figure supplement 2. Sampling and resampling techniques for SMC variance reduction.

Figure 5–Figure supplement 3. Effect of the fixed-lag smoother delay on inferred AP discharge probability using the SMC algorithm with the SBM.

Figure 5–Figure supplement 4. Accuracy and speed of SMC/SBM-based AP inference as a function of particle count.

Figure 5–Figure supplement 5. Identification of per-neuron parameters from fluorescence data alone.

Figure 5–Figure supplement 6. Fitting a single AP sequence to the posterior probability of AP discharge given fluorescence data inferred by the SMC algorithm.

Table 2: SBM parameters fit *in vitro*

Symbol	Description	Best-fit value	Range of fit values*	Units
k_1^+	First GCaMP6s on-rate	2.4	2.3 - 12752	$\mu M^{-1}\,s^{-1}$
k_2^+	Second GCaMP6s on-rate	263	2.9 - 413	$\mu M^{-1}\,s^{-1}$
k_3^+	Third GCaMP6s on-rate	31	2.5 - 44	$\mu M^{-1}\,s^{-1}$
k_4^+	Fourth GCaMP6s on-rate	30	30 - 43	$\mu M^{-1}\,s^{-1}$
k_1^-	First GCaMP6s off-rate	2.2	0.3 - 7093	s^{-1}
k_2^-	Second GCaMP6s off-rate	390	1.5 - 532	s^{-1}
k_3^-	Third GCaMP6s off-rate	2.2	2.0 - 19	s^{-1}
k_4^-	Fourth GCaMP6s off-rate	3.7	3.4 - 4.3	s^{-1}

*Range of parameter values fit from multiple initializations that converged with a normalized r.m.s. residual <10% higher than the lowest r.m.s. residual over all initializations (17 / 50 initializations satisfied this criterion).

reliability of AP inference. To deal with the computational burden of solving the SBM's global rate equation (Figure 2 - figure supplement 1) separately for each particle, we implemented a custom ordinary differential equation solver without loops or branch points to maximize the efficiency of GPU-based multithreaded computation (Figure 5 - figure supplement 1, methods). This solver uses a backward (implicit) Euler method, allowing time steps 1000 times longer than for a forward (explicit) solver while maintaining accuracy and numerical stability. We also designed new techniques for randomly extending particles' AP sequences (the SMC sampler) and selecting particles at each fluorescence observation (resampler) specifically for SMC-based AP inference (Figure 5 - figure supplement 2, details in methods). Most importantly, our SMC algorithm analyzed the fluorescence data twice, with the results of the first round used to design sampling and resampling distributions for the second round. Finally, we adjusted the fixed-lag smoother delay, which determines how far the algorithm looks into future fluorescence data when inferring APs at each time point (see methods). Observing the standard tradeoff [73] between incorrect results at short delays and inconsistent results at long delays (Figure 5 - figure supplement 3), we set this value to 500 ms. These techniques greatly increased the reliability of the SMC algorithm's output, but over 10^5 particles were still required for consistently accurate inference (Figure 5d, Figure 5 - figure supplement 4).

In addition to inferring APs and baseline drift, the algorithm also estimated $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG} for each neuron by maximizing the likelihood of fluorescence data given these parameters (Figure 5 - figure supplement 5, methods). After the SMC algorithm returned the probability of an AP every 10 ms, a subsequent processing step produced a single AP sequence without time discretization along with temporal uncertainty for each AP (Figure 5d-e, Figure 5 - figure supplement 6, methods).

When a neuron's true AP sequence is known, baseline fluorescence can be calculated directly using periods without APs (Figure 3 - figure supplement 1). However, when AP times are unknown or have been held out for testing, baseline and APs must be inferred together based on fluorescence data alone. To test whether an SMC algorithm could cope with data where true baseline fluorescence is never observed due to high firing rates, we analyzed a recording from a pyramidal neuron where rhythmic AP bursts prevented the fluorescence from returning to baseline (Figure 5e) without including periods of inactivity before or after the analyzed data. Despite the lack of fluorescence observations at baseline, the algorithm correctly identified the number of APs in nearly all bursts and inferred a drifting baseline less than the minimum fluorescence value. Having verified that our SBM/SMC algorithm can successfully infer APs in individual cases, we next tested it on our complete *in vivo* dataset.

Accuracy of AP inference techniques

We evaluated the accuracy of SBM-based AP inference along with alternative methods (MLspike, c2s-s, c2s-t and thr-σ), comparing true electrically detected APs to the APs inferred by each method (Figure 6a-c). For correlation-based analyses that do not depend on data units, we also included two linear deconvolution-based methods that do not infer APs, but rather a unitless quantity proportional to the neuron's firing rate: FOOPSI [132] and CFOOPSI [105]. We tested performance using a cross-validation procedure for all methods and accuracy measures to prevent overfitting: the SBM or other method was fit with each individual neuron excluded from the training data, and accuracy was evaluated using the held out neuron.

We first evaluated the agreement of true and inferred AP sequences by calculating their correlation over time for each algorithm and neuron. We used Gaussian smoothing (σ = 100 ms, see methods) to emphasize accuracy of AP counts as opposed to precise AP discharge times within bursts (see below for analysis of timing accuracy). Correlation between the true and inferred AP sequences (Figure 6d) was highest for the SBM (0.83 ± 0.13, n = 26 neurons), with significantly higher mean (p < 5e-6, t-tests) and median correlations (p < 0.005, rank sum tests) than other methods and higher correlation for nearly all neurons (24-26/26, p < 2e-5, sign tests). The SBM also produced the highest correlations for 5 neurons recorded in 3 animals after development of the SBM and all algorithms for model fitting and AP inference (0.77 ± 0.15, next highest c2s-t at 0.63 ± 0.12). The F1-score [29], which like correlation reaches 1 for perfect accuracy but is also sensitive to the units of algorithms' outputs (replacing every inferred AP with 2 APs does not change the correlation), was also highest for the SBM (0.65 ± 0.23, Figure 6 - figure

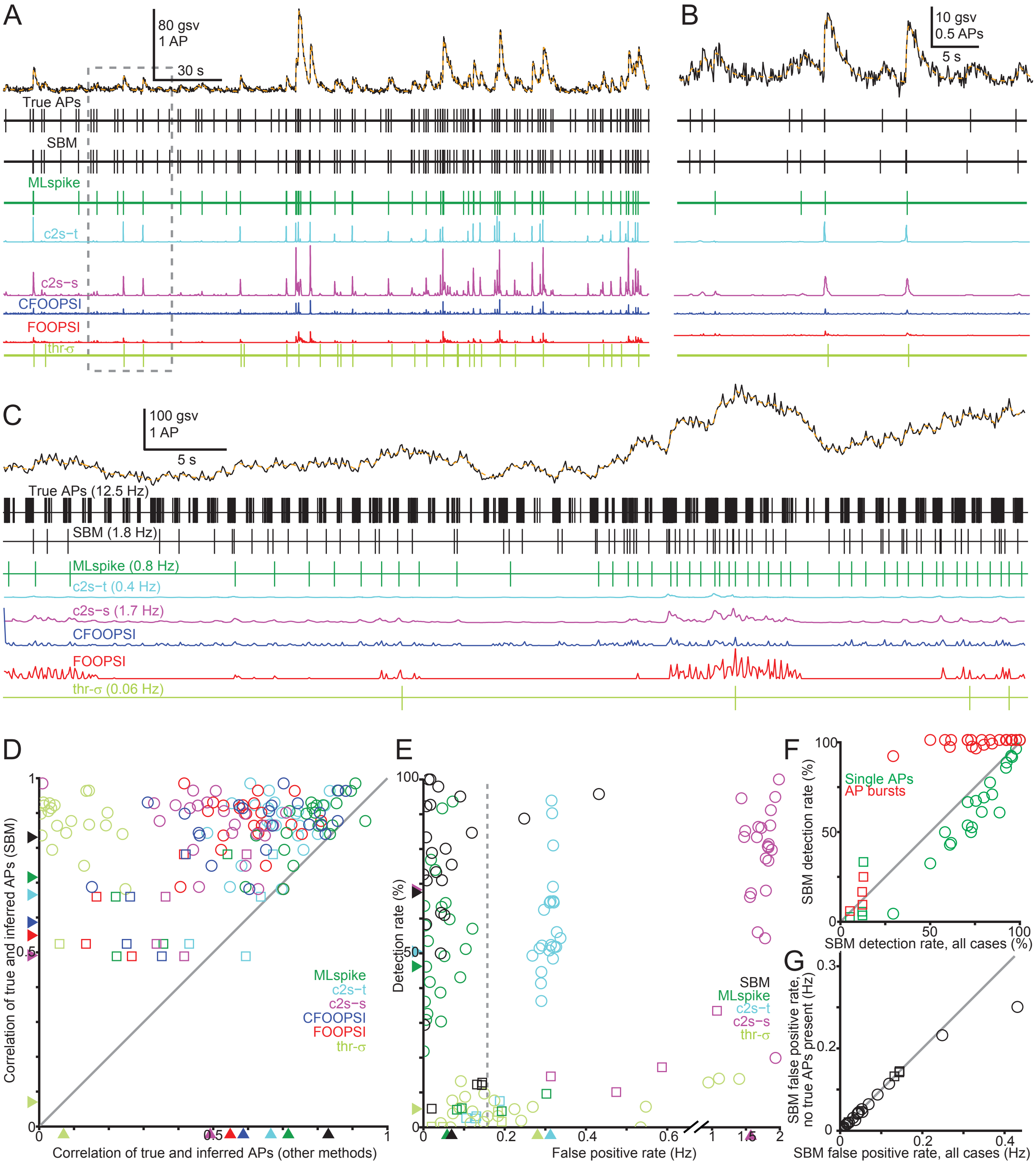

Figure 6: **(A)** Electrically detected APs and GCaMP6s fluorescence (black, upper) from a L2/3 mouse visual cortical pyramidal neuron. Orange curve shows the SBM's posterior mean for denoised fluorescence. AP inference results (lower). **(B)** Detailed view from (A). **(C)** As in (A-B) for an interneuron neuron firing at 12.5 Hz. **D** Correlation between each neuron's true and inferred AP sequences for the SBM (y-axis) compared to other methods, with 100 ms Gaussian smoothing (n = 26). Arrowheads indicate mean correlation over neurons for each method and squares indicate interneurons. **(E)** False positive and detection rates for each neuron and algorithm (n = 26). Arrowheads indicate mean over neurons, squares indicate interneurons and the dashed line shows the true mean spontaneous firing rate. **(F)** SBM detection rates for single APs (black) and bursts of 2 or more APs (red, burst duration <100 ms), compared to overall detection rates for each neuron (squares indicate interneurons). **(G)** Rate of false positives inferred by the SBM without any true APs within 200 ms (y-axis) compared to overall false positive rates including overestimation of AP counts (x-axis). Squares indicate interneurons.

Figure 6–Figure supplement 1. Precision, recall and F1 scores.
Figure 6–Figure supplement 2. AP inference accuracy as a function of peak 1-AP fluorescence amplitude.
Figure 6–Figure supplement 3. Effect of imaging frame rate and SNR on the accuracy of AP sequences inferred by the SBM.
Figure 6–Figure supplement 4. Comparison of SBM AP inference accuracy using rate constants fit to *in vivo* data vs. rate constants fit from *in vitro* binding assay data.
Figure 6–Figure supplement 5. Accuracy of AP inference with one neuron of training data.
Figure 6–Figure supplement 6. Isolated single APs without visually apparent fluorescence increases.

supplement 1).

To characterize the agreement of true and inferred AP sequences beyond the single values provided by the correlation and F1 measures, we next computed the rates of missed APs and false positive detections when no APs were present (Figure 6e). For these comparisons we used a timing tolerance of 100 ms when matching the true and inferred AP sequences. The SBM detected 67.5 ± 29.2% of APs (n = 26 neurons), with a false positive rate of 0.07 ± 0.09 Hz. MLspike detected APs at a lower rate of 46.3 ± 24.3% ($p <$ 1e-6, t-test) but its false positive rate was not significantly different from the SBM (0.06 ± 0.06 Hz, p = 0.54). Similarly, c2s-s detected APs at the same rate as the SBM (68.3 ± 26.4%, p = 0.88) but with a false positive rate ten times the true median firing rate (1.57 ± 0.45 Hz), possibly because it was trained on data with firing rates of 2-5 Hz [126]. Compared to the SBM, lower detection and higher false positive rates were observed for c2s-t (50.5 ± 23.8%; 0.31 ± 0.10 Hz) and thr-σ (5.3±4.4%; 0.28±0.34 Hz). In 19/26 neurons tested, the SBM detected over half of all APs and less than half its inferred APs were false positives. Of the 7 neurons for which SBM did not satisfy this criterion, 4 were interneurons. In contrast, MLspike satisfied this criterion for 11/26 neurons, and c2s-t, c2s-s and thr-σ for no neurons.

We next examined which features of the AP sequence were associated with various types of inference errors for the SBM. About half the APs missed by the SBM (52 ± 32%, n = 26 neurons) were found to be isolated single APs (as opposed to bursts or longer discharge sequences). Overall, 58 ± 31% of isolated single APs were detected by the SBM, though this ranged over neurons from 3% to 100% (Figure 6f, green). For isolated bursts of 2 or more APs within 100 ms, the chance of detecting at least 1 AP was 99 ± 0% for pyramidal neurons (Figure 6f, red circles, n = 22), and 14 ± 7% for interneurons (Figure 6f, red squares, n = 4). When the SBM inferred false positive APs not present in the electrical recording, in 93 ± 12% of cases no true APs were present within 100 ms, while in the remaining cases APs were present but their number was overestimated (Figure 6g). In summary, false negative errors (missed APs) arose from a combination of failure to detect isolated single APs, underestimation of activity in interneurons and underestimation of AP counts in bursts. False positive errors arose predominantly from spurious detections when no APs were present, with overestimation of AP counts accounting for only a small fraction of these errors.

The SBM's accuracy depended on the characteristics of signal and noise in each neuron's fluorescence recordings. Neurons whose single APs evoked larger fluorescence increases exhibited higher correlations between true and SBM-inferred APs (Figure 6 - figure supplement 2, r = 0.42, p = 0.05, t-test, n = 22) and higher detection rates (r = 0.58, p = 0.005) but not lower false positive rates (r = 0.07, p = 0.74). Similarly, signal-to-noise ratio (SNR) was positively associated with correlation (r = 0.36, p = 0.05, Figure 6 - figure supplement 3) and detection rate (r = 0.51, p = 0.003) but unrelated to false positive rate (r = -0.04, p = 0.8). Imaging frame rate showed no significant association with correlation, detection rate or false positive rate over the range of 10-60 Hz ($p >$ 0.2).

Finally, we examined how AP inference accuracy depends on the type and quantity of training data used to fit SBM parameters. We first examined the effect of using GCaMP6s rate constants obtained from *in vitro* binding assays, while fitting physiological parameters from *in vivo* data with cross validation. This resulted in only a slight reduction in accuracy (correlation 0.81 ± 0.13, detection rate 65.2 ± 29.4%, false positive rate 0.07 ± 0.06 Hz, n = 26, Figure 6 - figure supplement 4) which did not significantly change correlation (p = 0.39). We also examined how our approach of model fitting followed by SMC-based AP inference would fare in an extremely data-poor scenario, when model parameters are fit from only a single neuron's data. To compute accuracy across our entire dataset, we carried out AP inference for each neuron based on parameters fit from the previous neuron (Figure 6 - figure supplement 5, methods). Correlation between true and inferred AP sequences was nearly unchanged when fitting SBM parameters to a single neuron's data, whether using GCaMP6s rate constants fit to *in vivo* data (r = 0.78 ± 0.13 vs. 0.83 ± 0.13, p = 0.20) or those fit to *in vitro* binding assays (0.80 ± 0.13 vs. 0.81 ± 0.13, p = 0.67). A much larger decrease in correlation values was observed when training c2s on a single neuron (r = 0.51 ± 0.18 vs. 0.67 ± 0.10, p = 0.0003); this difference might arise from c2s' greater number of parameters ($\tilde{1}000$) which could make it more susceptible to overfitting.

Overall, these results show that the SBM provides a clear advance in accuracy and robustness compared to currently available inference methods, for multiple error metrics and in both data-rich and data-poor conditions. Nonetheless, it is worth emphasizing that for some neurons many APs will be missed by all algorithms, as some single-AP responses are not distinguishable from noise (Figure 6 - figure supplement 6).

Linear readout of neural activity

We next examined whether APs were accurately inferred at all levels of neural activity, or only for certain firing rates. We first divided our entire *in vivo* dataset, ranging from isolated single APs to high frequency bursts (Figure 7a), into 500 ms windows and calculated the true firing rate in each one (0-24 Hz for pyramidal and 0-48 Hz for interneurons) along with each algorithm's inferred rate. This analysis revealed a tight linear dependence of the SBM's average inferred rate on the true rate (Figure 7b) that spanned the full range of firing rates present in our dataset and did not depend on the window size (Figure 7 - figure supplement 1). For the SBM, this linearity was observed in all pyramidal neurons (Figure 7b, black, $r^2 = 0.98 \pm 0.02$, n = 22) and interneurons (Figure 7b, red, $r^2 = 0.96 \pm 0.03$, n = 4), while other methods provided a less linear readout of neural activity ($p < 0.01$, rank sum tests, Figure 7 - figure supplement 2a-b). The average SBM-inferred rate grew with the true rate at a slope near 1 for pyramidal neurons (0.93 ± 0.20, n = 22), while lesser slopes were observed for other methods (Figure 7 - figure supplement 2c). For all true firing rates, the ratio of the SBM-inferred and true rates was close to 1, but this was not the case for other methods (Figure 7c). All methods tested underestimated firing rates in interneurons (Figure 7d, Figure 7 - figure supplement 2d). The SBM also provided the best agreement between the true and inferred values of neurons' overall mean spontaneous firing rates, and for pyramidal neurons this relationship was linear with a slope near 1 and y-intercept near 0 (Figure 7 - figure supplement 3). These results show that the SBM linearly reads out pyramidal neurons' firing rates from <0.1 Hz (mean spontaneous rates) to the maximum rate recorded in our dataset (> 20 Hz over 500 ms windows) and can do so with correct units, while for interneurons firing rate readout is linear but underestimates the true AP count.

We also examined whether the inter-spike intervals (ISIs) between nearby APs affect the number of APs inferred by the SBM. In recordings from pyramidal neurons, we first identified 2-AP bursts with ISIs up to 40 ms ($n = 528$, shortest 3.3 ms), and compared the ISI to the number of inferred APs (Figure 7e, blue). Calculating the average inferred AP count as a function of ISI, we found these quantities to be uncorrelated ($p = 0.92$, 10 ms time bins). Similar results were observed for 3-AP bursts (Figure 7e, green, $p = 0.88$, $n = 211$) and for interneurons (Figure 7f, $p = 0.96$, $n = 75$ for 2-AP and $p = 0.07$, $n = 59$ for 3-AP bursts). As expected from the above analyses of firing rate, the SBM underestimated AP counts in bursts for interneurons, but did so in a uniform way that did not depend on ISIs. These results show that SBM-inferred AP counts do not depend on ISIs, and that our approach can successfully enumerate APs separated by < 10 ms.

Precision of inferred AP times

When inferring events such as APs from measurements with limited SNR and temporal resolution, event detection times will generally not be precisely correct. Instead, they will exhibit some time difference from the nearest true event time, resulting in a probability distribution of timing errors that depends on the data, model and inference method. Investigating temporal precision is particularly important for AP times inferred from GECI fluorescence signals, as SNR can be <1 and the frame rate as low as 10 Hz (Figure 6 - figure supplement 3). To measure the accuracy of inferred AP times, we first examined how the correlation of true and inferred AP-sequences (Figure 6d) depends on the Gaussian filter size σ used to smooth both AP sequences before comparing them (Figure 8a). When the timing error between true and inferred APs is less than σ, the Gaussian filtering causes the smoothed AP sequences to overlap and increases correlation, so smaller σ values impose a more stringent requirement for accurate AP times. The SBM produced the highest correlations for all values of σ. Similarly, we calculated AP detection and false positive rates (Figure 8b-c) as functions of the timing tolerance within which true and inferred APs were considered as matching. For all timing tolerances, detection rates were similar for the SBM and c2s-s and lower for other methods, while false positive rates were similar for the SBM and MLspike and higher for other methods. We also examined cross-correlation as a function of the time lag from true to inferred APs (Figure 8d, figure 8 - figure supplement 1d, $\sigma =$ 25 ms); the SBM exhibited the highest cross-correlation peak (0.64 ± 0.18, Figure 8e) at the shortest time lag (-2.0 ± 29.7 ms, Figure 8f).

To further quantify timing precision, we analyzed the time differences between true and inferred APs for each algorithm. Using only isolated single APs (without other APs 1 s before and 0.5 s after), we calculated for each method the rate of inferred APs (or for FOOPSI and CFOOPSI, the unitless output) as a function of time difference from the true AP (Figure 8g, Figure 8 - figure supplement 1a-c). For each neuron with isolated single APs ($n = 22$), we then computed the average absolute time difference from the true AP for each algorithm's output (Figure 8h), which was lowest for the SBM (64 ± 28 ms). The distribution of SBM-inferred AP times peaked 17.4 ± 51.4 ms before the true AP time, while other methods exhibited larger positive or negative delays (Figure 8i).

To examine whether the SBM's timing errors can be attributed to some aspect of the image acquisition process itself, we also compared the absolute error and mean time delay for single APs to SNR and imaging frame rate (Figure 8 - figure supplement 2). Absolute timing error decreased with SNR ($r = -0.46$, $p = 0.015$) but not frame rate ($r = 0.18$, $p = 0.38$) while the mean time delay was closer to zero at higher frame rates ($r = 0.43$, $p = 0.024$) but not at higher SNR ($r = -0.20$, $p = 0.31$). The absolute time differences between true and SBM-inferred APs could also be predicted by the timing uncertainty values output by the SBM for each AP (Figure 8 - figure supplement 3), showing that the SBM can accurately quantify its own uncertainty regarding AP times.

Because inferred APs can precede true APs in time (Figure 8g), sensory-evoked APs detected optically may violate the assumption that causes precede effects [4, 50]. To illustrate this, we carried out simulations of L2/3 neurons' sensory responses (Figure 8 - figure supplement 4) which showed that optical AP detection can cause stimulus-evoked APs to be assigned to times before the stimulus. Therefore, in order to facilitate the use of optically detected APs for causal time series analysis we calculated the fraction of inferred APs before the true AP, and how this fraction can be reduced by shifting inferred APs forward in time. For isolated single APs, 90% of SBM-inferred APs can be assigned after the true AP time when using a forward shift of 114 ms, but when the latency of stimulus-evoked APs is known a smaller shift can be used (Figure 8 - figure supplement 5). Together, these results show that the SBM results in smaller timing errors and average time differences from true to inferred APs than other methods, but care must be taken when identifying causal effects from optically measured AP activity.

Quantifying sensory-evoked AP discharge

Having developed and validated the SBM using spontaneous activity in single neurons, we also applied it to infer sensory-evoked activity across neuronal populations. We recorded GCaMP6s fluorescence in visual cortex while presenting drifting grating stimuli (5 repetitions in 8 directions) and inferred AP times using the SBM, MLspike and thr-σ (Figure 9a-b).

Sensory stimulation at each neuron's preferred orientation resulted in SBM-inferred firing rates from 0-7 Hz (mean 0.9 ± 1.3 Hz, $n = 66$), while lower firing rates were inferred by MLspike (0.3 ± 0.4 Hz, $p = 2e-5$, t-test) and thr-σ (0.5 ± 0.2 Hz, $p = 0.04$), consistent with these algorithms' lower

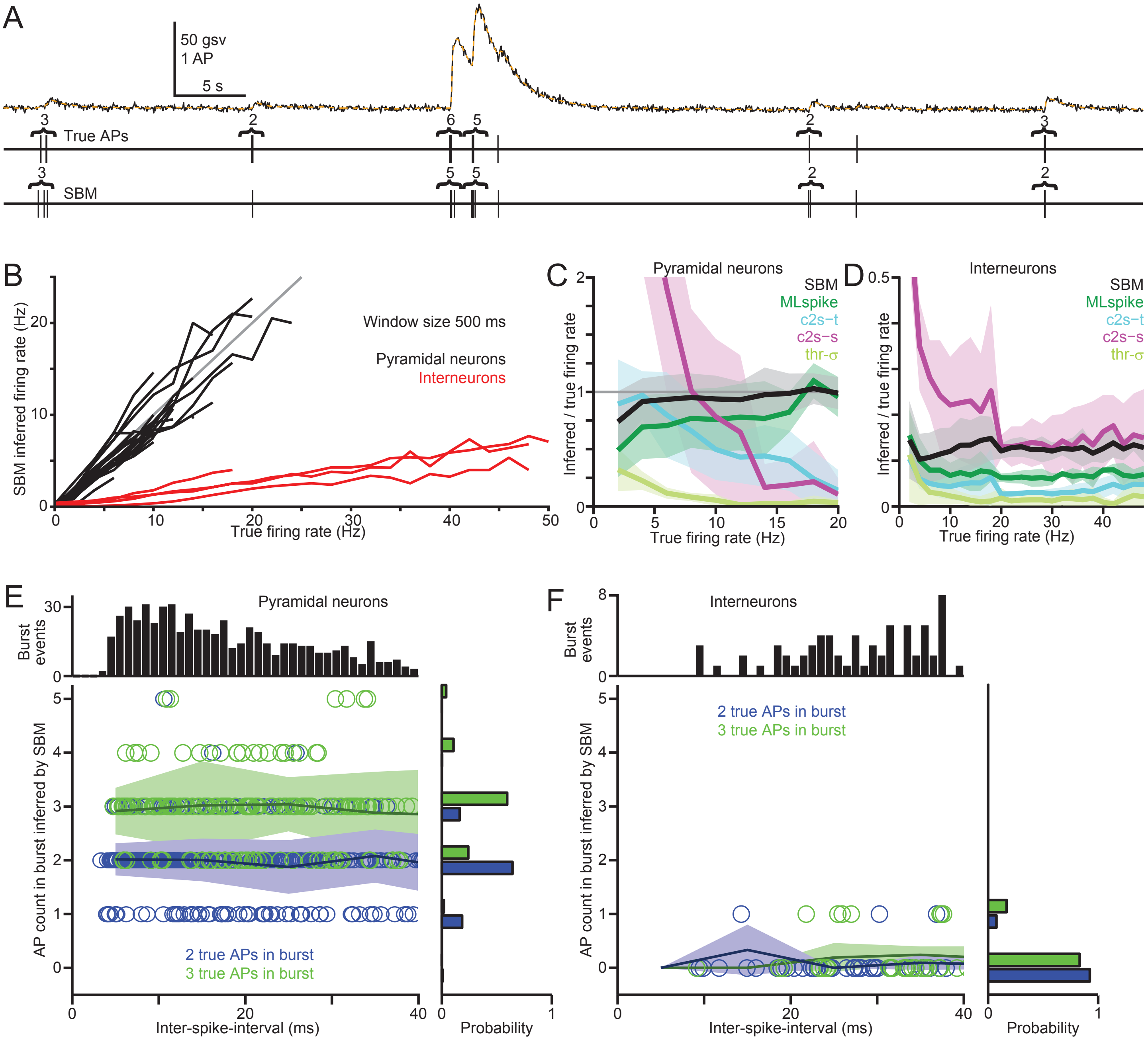

Figure 7: **(A)** Electrically detected APs and GCaMP6s fluorescence (black, upper) from a L2/3 mouse visual cortical pyramidal neuron, and SBM-inferred APs. Orange curve shows the SBM's posterior mean for denoised fluorescence. Numbers indicate the true and inferred AP counts for each burst; single APs are unlabeled. **(B)** Firing rate inferred by the SBM as a function of the true firing rate, calculated across the complete *in vivo* dataset using 500 ms windows with 90% overlap. Each curve corresponds to a single pyramidal neuron (black) or interneuron (red). **(C)** Ratio of inferred to true firing rates in pyramidal neurons, as a function of the true firing rate, for each inference method. Each curve shows an average over neurons (n = 22) and shaded regions show standard deviations. **(D)** As in (C), but for interneurons. **(E)** *Center*: AP counts inferred by the SBM in pyramidal neurons as a function of inter-spike interval (ISI) for bursts of 2 (blue) and 3 APs (green). Solid lines show average inferred AP counts in 10 ms ISI bins. For 3-AP bursts, the mean of the two ISIs was used. Inferred APs were counted from 200 ms before the first AP in the burst to 200 ms after the final AP, and only bursts flanked by 400 ms without other true APs before and after were included. *Top*: Histogram of ISIs for 2- and 3-AP bursts. *Right*: Probability distributions for the number of SBM-inferred APs for bursts of 2 (blue) and 3 true APs (green). **(F)** As in (E), but for interneurons.

Figure 7–Figure supplement 1. Linearity of SBM-based firing rate inference is robust to the choice of time window size.

Figure 7–Figure supplement 2. Linearity and slope of inferred firing rate as a function of true firing rate, for all inference methods.

Figure 7–Figure supplement 3. Accuracy of inferred mean spontaneous firing rates.

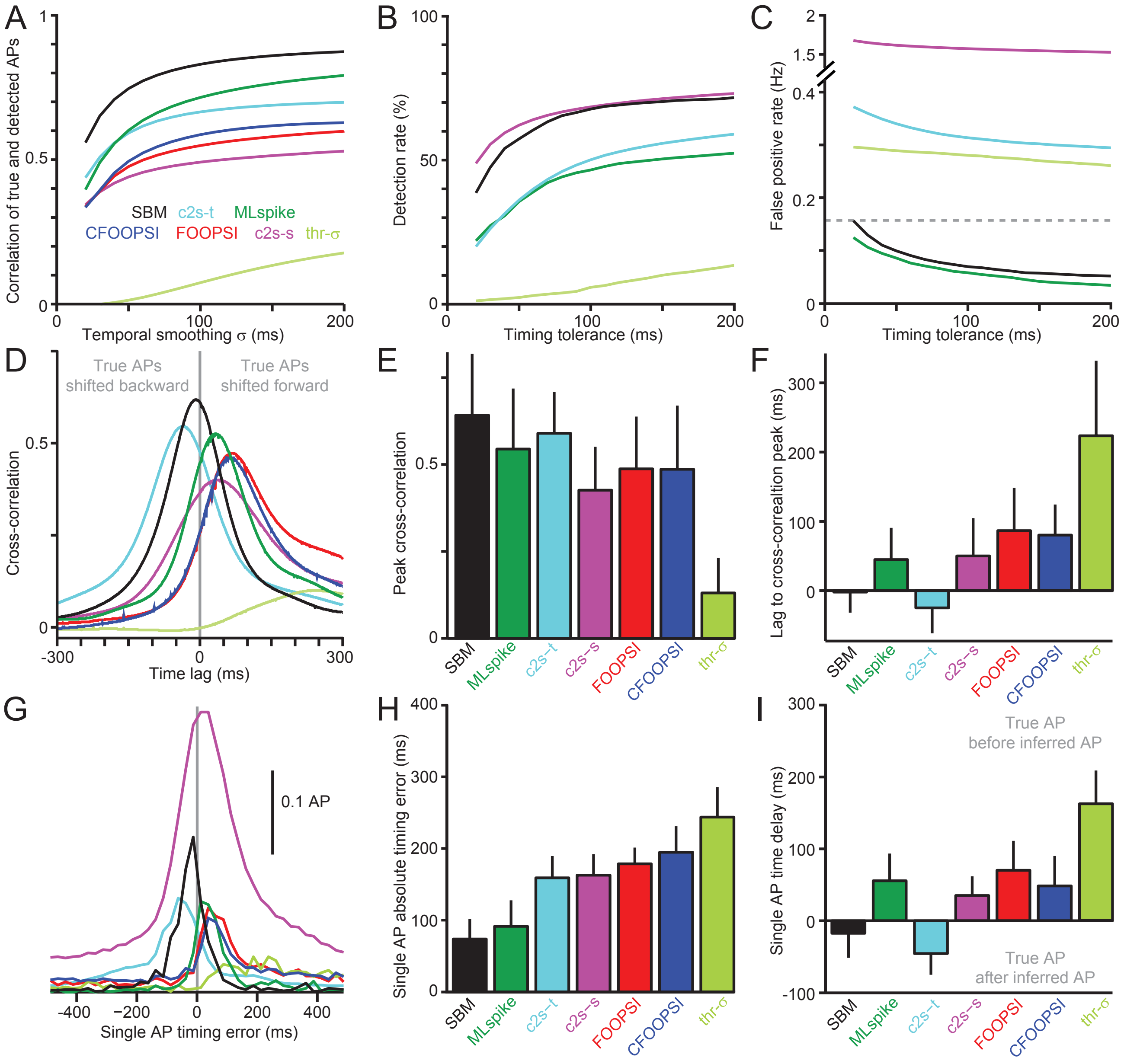

Figure 8: **(A)** Correlation between true and inferred AP sequences as a function of Gaussian smoothing σ and averaged over neurons ($n = 26$). **(B)** Detection rate as a function of timing tolerance. **(C)** False positive rate as a function of timing tolerance; dashed line indicates median spontaneous firing rate. **(D)** Cross-correlation between true and inferred APs for each inference method, as a function of time lag. Positive lags correspond to shifting true APs forward in time before comparing them to inferred APs. Cross-correlations were computed with a smoothing σ of 25 ms. **(E)** Mean and standard deviation over neurons of the peak cross-correlation. **(F)** Mean and standard deviation over neurons of the time lag for which cross-correlation was maximized. **(G)** Mean rates of AP inference relative to true AP times (gray vertical line) for single APs without other APs for 1 s before and 0.5 s after. The unitless outputs of CFOOPSI and FOOPSI have been rescaled for each neuron to give an average total output of 1 for each true single AP. **(H)** Mean and standard deviation over neurons of each algorithm's absolute timing error for isolated single APs. **(I)** Mean and standard deviation over neurons of each algorithm's average delay from true to inferred AP times for isolated single APs.

Figure 8–Figure supplement 1. Cross-correlation and single-AP timing accuracy in individual neurons

Figure 8–Figure supplement 2. Effect of imaging frame rate and SNR on timing accuracy of APs inferred by the SBM.

Figure 8–Figure supplement 3. Data-based evaluation of timing uncertainty values inferred by the SBM for isolated single APs.

Figure 8–Figure supplement 4. Simulations showing the effect of timing errors in optically detected APs on peri-stimulus time histograms (PSTHs).

Figure 8–Figure supplement 5. Forward time shifts limit too-early assignment of inferred APs.

AP detection rates (Figure 6e, Figure 7c-d). Orientation tuning curves calculated using SBM-inferred APs (Figure 9c, black) were similar to previous observations [93], including a diverse range of preferred orientations as well as untuned and nonresponsive neurons. Some pairs of adjacent neurons had opposite orientation preferences (Figure 9c, neurons **i** and **iv**), as previously observed using synthetic indicators in rats [97] and mice [88]. These results show that the SBM can be used to measure neurons' sensory tuning, and that quantitative estimation of this tuning can depend on the choice of AP inference method.

Discussion

We developed the SBM to address the complex, nonlinear and variable relationship between AP discharge and GCaMP6s fluorescence. By using quantitative biophysical modeling to link fluorescence and APs, it outperformed existing inference methods (figures 6-8) with more accurate firing rates, higher correlation and F1 values, higher detection rates, fewer false positives, greater linearity and more precise AP times. This improved accuracy can be explained by several differences from previous approaches. For methods based on linear deconvolution (FOOPSI [132] and CFOOPSI [105]) and thresholding (thr-σ [32]), accuracy is limited by the fact that GCaMP6s fluorescence neither increases linearly with AP counts nor crosses a fixed threshold upon AP discharge (Figure 1). For MLspike [29], the difference may arise because individual binding steps and endogenous buffers are not included in the phenomenological model, or because MLspike updates its internal state only once per image frame. c2s uses the largest number of parameters (1004) and is the most mathematically flexible method tested, but infers APs based only on fluorescence values within a one second window around each time point while other techniques look further into the past and future. In addition to the specific limitations of each of these methods, they all fail to address the variability of the AP-fluorescence relationship over neurons, which is a central challenge for AP inference from GECI fluorescence. While algorithms such as c2s which fix all parameters after fitting cannot account for this variability, the opposite extreme of estimating all parameters separately for each neuron would lead to overfitting and is likely to be computationally infeasible for models rich enough to describe nonlinear GECIs. A small number of parameters can be chosen to vary across neurons (for example, a single scaling factor as proposed in [29]), but for non-physical parameters such as polynomial coefficients or deep learning connection weights there is no reason for neuron-to-neuron variability to manifest in some parameters and not others. In contrast, the SBM captures this variability using the parameters $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG} (which are expected from their physical interpretations to vary over neurons), requires fewer total parameters than c2s and is more robust when trained on smaller datasets (Figure 6 - figure supplement 5).

The SBM's false positive rate was 0.07 Hz, bringing GCaMP6s into the range achieved by synthetic sensors such as OGB1 [51, 53, 71, 70]. However, even with the SBM GCaMP6s simply cannot match OGB1's sensitivity to single APs: excluding the 4 interneurons to match previous OGB1 measurements in pyramidal neurons brings the detection rate to only 78%, compared to 90-95% for OGB1[51, 53, 71, 70]. This may reflect an upper performance limit intrinsic to GCaMP6s that cannot be overcome by further improvements of the SBM or other methods, as visual inspection shows a total lack of fluorescence increase for some single APs (Figure 6 - figure supplement 6). Consequently, while bursts of 2 or more APs were nearly always detected, for many neurons isolated single APs were frequently missed (Figure 6f). These errors could complicate efforts to understand how neuronal populations carry out their essential functions, as even small numbers of APs in a single neuron can have strong effects on neuronal tuning through synaptic plasticity [102]. On the other hand, the fact that most of the SBM's false positive errors occurred in the absence of true APs (Figure 6g) suggests that most subsequent analyses would be unaffected by these errors.

The SBM provides the most linear readout of neural activity (Figure 7a-d , Figure 7 - figure supplements 1-3), with the ratio of true and inferred firing rates independent of the true rate (Figure 7c-d) and of ISIs within bursts (Figure 7e-f). This is an important characteristic for an AP inference algorithm, since it facilitates subsequent stages of quantitative analysis such as computation of sensory tuning curves [92]. Furthermore, for pyramidal neurons the SBM allows comparison of firing rates across neurons with correct units (Figure 7b-c, Figure 7 - figure supplement 2c), despite the variation in AP-evoked fluorescence across neurons (Figure 1c, Figure 3e-f).

Compared to alternative methods, SBM-based inference introduced both lesser absolute timing errors and smaller mean delays between true and inferred APs. AP timing plays a central role in dendritic integration [13], sensory coding [26, 67, 85, 135] and both *in vitro* [11, 81] and *in vivo* synaptic plasticity [86, 102]. Nonetheless, for GCaMP6s SBM-based inference produced timing errors far larger than the timescale of synaptic input integration [134] and often larger than the image acquisition time. Since timing errors were reduced in recordings with higher SNR and mean delays were reduced at higher frame rates (Figure 8 - figure supplement 2), timing precision can be modestly improved by imaging faster or reducing noise, but major gains may require switching to a faster sensor. Because the timing uncertainty arising from optical measurement of APs (Figure 8d-i) can reverse the apparent order of an inferred AP and the sensory stimulus that evoked it (Figure 8 - figure supplement 4), causal analysis of optically measured neural activity should either take the uncertainty of inferred AP times into account or apply a corrective time shift (Figure 8 - figure supplement 5) before further analysis.

A current limitation of our SBM/SMC approach is that it underestimates activity for interneurons (figures 6-7). This can be attributed to interneurons' negligible fluorescence increases for single APs (Figure 1 - figure supplement 3), as previously reported using OGB1 with post-hoc immunohistochemistry [69]. We emphasize that this shortcoming was specific to interneurons rather than high firing rates, as the algorithm accurately inferred APs in pyramidal neurons at firing rates >20 Hz and ISIs < 10 ms (Figure 7). Furthermore,

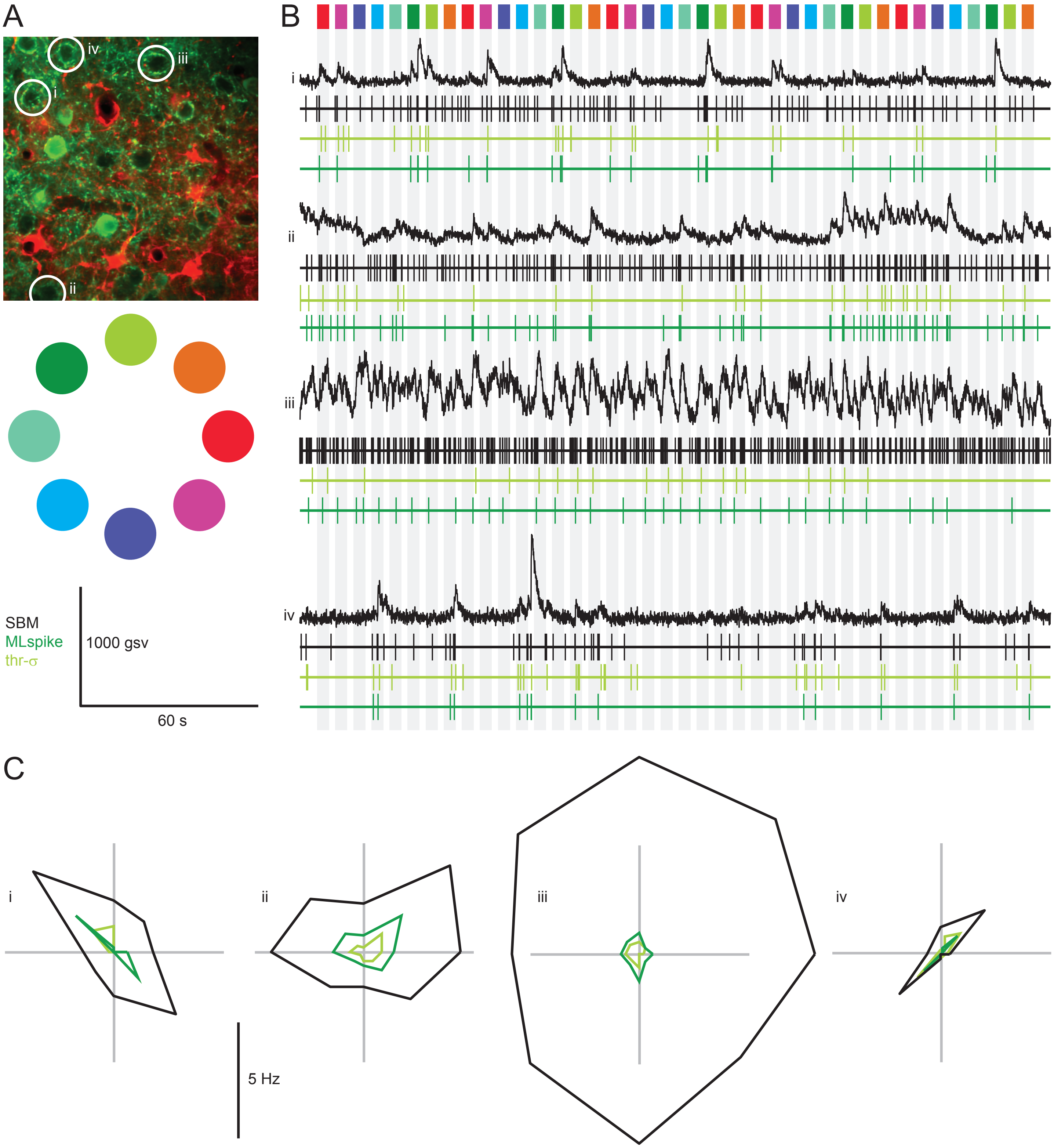

Figure 9: **(A)** Two-photon image of neuronal population (upper) expressing GCaMP6s (green) in L2/3 rat visual cortex, with astrocytes stained using sulforhodamine 101 (red) and imaged during visual stimulation with drifting gratings in 8 directions (colored circles, lower). **B** Fluorescence (black) recorded from 4 neurons from the population in (A) during presentation of drifting grating stimuli (gray bars; colors above indicate the direction of motion for each stimulus presentation). AP inference results are shown for the SBM (black), MLspike (dark green) and thr-σ (light green). **C** Orientation tuning curves showing the mean firing rate evoked by presentation of the drifting grating stimulus in 8 directions, for the 4 neurons in (A-B). Firing rates were calculated over the 2-second stimulus presentation.

the SBM's inferred firing rate grew linearly with the true rate in interneurons from 0 to >40 Hz, albeit with a slope <1 (Figure 7b,d). These results could be explained by stronger calcium buffering in interneurons, in which case SBM performance could be improved by relaxing the assumption that the same endogenous buffers are present in all neurons. Including per-neuron buffering parameters would require solving a more difficult parameter estimation problem for each neuron, possibly using recent machine learning techniques for parametric model fitting [78, 101, 128, 14]. More generally, the SBM could benefit from more extensive quantitative knowledge of endogenous buffers, extrusion mechanisms and calcium flow between the cytosol, nucleus and endoplasmic reticulum [110, 112, 115]. The SBM remains extremely simplistic in comparison to the diverse set of calcium-binding molecules present in neurons; for example, the InterPro database lists 826 different mouse proteins containing the calcium-binding EF hand domain pair alone [42].

SBM-based AP inference could also benefit from new methods for acquiring and pre-processing imaging data. New scanning techniques can increase imaging speed and the number of recorded neurons [20, 53, 103] towards the levels achieved by electrical approaches. Reduction in neuron-to-neuron variability might be achieved by using transgenic animals instead of viral delivery [24, 79] or by localizing the indicator to subcellular compartments [36, 72]. Accuracy could also benefit from improved methods for motion correction [32, 52, 105] and removal of contaminating signals from nearby structures [31, 89, 105].

Current knowledge and experimental tools provide only a partial picture of the calcium dynamics evoked by AP discharge in neuronal somata, but the SBM is broadly corroborated by past experimental measurements. The value of $\Delta[\mathrm{Ca}^{2+}]_{\mathrm{AP}} = 20.2$ µM calcium influx per AP is consistent with measurements in midbrain neurons [107] using the action-potential clamp method [77] that reported 4.32 pC of charge carried by calcium ions for each AP. For a spherical soma 10-15 µm in diameter, this would imply a concentration increase of 13-43 µM. While this far exceeds the sub-micromolar changes in free calcium typically observed upon AP discharge [116, 142], the SBM models most calcium ions inside the neuron as bound to endogenous buffers. For every free calcium ion, the SBM modelled 203 ions bound to endogenous buffers in a neuron at rest, consistent with previously reported values for pyramidal neurons *in vitro* [60, 80, 99].

Because the SBM describes the effect of AP discharge on fluorescence through the biophysical framework of mass action kinetics, we were able to apply it to *in vitro* binding assays as well as *in vivo* experiments. SBM parameters determined from global fits of multiple *in vitro* experiments can be used *in vivo*, both to predict fluorescence from APs (Figure 5) and to infer APs from fluorescence (Figure 6 - figure supplement 4). The fact the SBM can describe the same protein-ligand interaction under such divergent circumstances lends support to the model class as a quantitative account of calcium-GCaMP6s interactions. However, for both *in vitro* and *in vivo* data, individual parameters could not be precisely determined, though predictions of the system's output were reliable and robust (Figure 3 - figure supplement 9). This parameter sloppiness could arise from spectroscopically silent binding steps (including the first 3 steps when using the best-fit SBM parameters for *in vivo* data, see Table 1), insufficient temporal resolution to resolve the most rapid binding reactions, insufficient number or diversity of experimental conditions *in vitro*, the limited number of recorded neurons or diversity of AP sequences *in vivo* or lack of structural identifiability in the model itself [8]. This sort of parameter sloppiness has been observed in many other complex models in biochemistry and systems biology [45, 54, 108].

The SBM and associated techniques for model fitting and AP inference have a number of applications beyond detecting APs with GCaMP6s. These procedures could be applied without modification to GCaMP6f/m [18], other GCaMPs, recently developed red fluorescent versions [23] and future calmodulin-based sensors. Because the SBM describes both the calcium sensor protein and some aspects of neuronal physiology, it could also prove useful in investigating the physiological side-effects of GCaMP expression in virally transfected neurons [2, 139] and genetically modified animal lines [119]. Given that SBM parameters obtained *in vitro* can be used to predict fluorescence and infer APs *in vivo*, SBM parameters fit from *in vitro* binding assays might be useful in large-scale screening of candidate GECIs [18]. The model could also be extended to include photoconversion of calmodulin-based activity integrators [44], multiple protein products of the same GECI gene [23] and calcium dynamics in dendrites, axons, spines and boutons [21, 39, 75, 115, 121]. As a framework for linking APs to experimental observations through reaction kinetics, more general versions of the SBM could describe biochemical reaction networks coupled to neural activity [10]. The models, fitting procedures and software tools we developed for *in vitro* binding assays could also be applied to other calcium-binding proteins or different protein-ligand interactions, as existing analysis software cannot easily fit model parameters for multiple binding sites or combine equilibrium and kinetic data from multiple binding assays in a common global fit [141].

The approach we have taken here, of fitting parametric generative models to experimental data while incorporating domain-specific knowledge into design of the model class, has already proved useful in several areas of neuroscience [37, 58, 66, 43, 109, 95, 100] and other fields of biology [45, 54, 94, 96, 108]. By improving the accuracy and interpretability of data analysis, these results help justify the use of quantitative causal models and domain-specific knowledge over model-free black box methods for the analysis of complex biological data.

Methods and Materials

Mathematical notation

$\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2, while $\mathcal{N}\left(x; \mu, \sigma^2\right)$ denotes the pdf of this distribution at the value x. We define a sum indexed by the empty set as zero, and a product indexed by the empty set as one. The operator $\circ$ denotes convolution. For a vector $v \epsilon \mathbb{R}^n$, $\text{diag}(v)$ is an n x n matrix with v on the diagonal and zeros elsewhere. For a vector $x \epsilon \mathbb{R}^n$, we denote by $x_{j:k}$ the sub-vector starting at x_j and finishing with x_k.

Virus injection

Experiments were conducted according to German animal welfare regulations. Experimental subjects for our combined optical-electrical recordings were 6 male C57BL6 mice, 20-25g body weight at the time of the virus injection. Fluorescence imaging with visual stimulation was carried out 70g male Listar hooded rat. During all surgical and recording procedures, anesthetic depth was regularly monitored and body temperature was maintained at 37°C using a heating pad and thermal probe. Mice were anesthetized with an intraperitoneal bolus injection of ketamine and xylazine (120 mg/kg and 16 mg/kg respectively) while rats were anesthetized with fentanyl citrate, midazolam hydrochloride, and medetomidine hydrochloride (5 µg/kg, 2 mg/kg and 150 µg/kg respectively). Supplemental doses of anesthetic solution given as required. The target area for the virus injection (Lambda -1 mm, lateral 2.5 mm in mouse and Lambda +1.5 mm, lateral 4.5 mm in rat) was exposed and marked. A small craniotomy (~0.5 x 0.5 mm) was opened approx. 1 mm anterior to the marked target area at the same lateral coordinate and a small opening made in the underlying dura. To induce expression of GCaMP6s in cortical neurons, a glass pipette (tip opening ~20 µm) filled with virus solution (AAV1.Syn.GCaMP6s.WPRE.SV40, Penn State Vector Core, PA, USA) was advanced into the cortex to the target area at a 15-20° angle to the brain surface, and an injection was made over approximately 5 minutes (250-400 nL in mouse and 1.5 µL in rat). The surgical site was then protected with silicone (KwikSil, World Precision Instruments, FL, USA) and skin incision closed. Animals were given a bolus dose (5 mg/kg flunixin meglumin for mice; 120 µg)/kg naloxone hydrochloride, 200 µg)/kg flumazenil, 750 µg)/kg atipamezole) for post-operative analgesia, and allowed to recover. Additional, buprenorphine hydrochloride and carprofen (30 µg)/kg and 5 mg/kg respectively) were administered for the rat.

Imaging with simultaneous electrophysiological recording

After an expression time of 10-15 days, animals were anesthetized with urethane (1.6 mg/kg), a custom-made headplate fixed to the skull using dental cement (Paladur, Kulzer GmbH, Hanau, Germany), an approx. 3 x 3 mm craniotomy centered over the injection target site opened and the dura removed. Astrocytes were counterstained using sulforhodamine 101 (0.5 mM, Sigma-Aldrich, MO, USA), which was applied topically to the cortical surface for 60-120 s, and the exposed cortex then covered with agar (1.2%, Sigma-Aldrich, MO, USA, dissolved in artificial cerebrospinal fluid (ACSF) of the following composition in mM: NaCl, 135; KCl, 5.4; CaCl2, 1.8; MgCl2, 1; Hepes, 5) and a coverslip to minimize brain movement during subsequent multiphoton and electrophysiological recordings.

Labeled neurons and astrocytes were visualized using custom-built multiphoton microscopes. Excitation light was provided by a Ti:Sapphire pulsed laser system (Mai Tai, Spectra Physics, CA, USA) tuned to 920 nm. Datasets were acquired using either an Olympus 20x (XLUMPlanFl, Olympus, Tokyo, Japan) or a Zeiss 20x (WPlan-APOCHROMAT, Zeiss, Oberkochen, Germany) objective lens. The scanning system consisted either of a conventional or resonance galvanometric system. With the conventional galvanometric system, imaging was carries out at 64 x 128 pixel resolution, acquired at frame rates of 10.4, 15.6, 18.6 or 37.2 Hz. With the resonance galvanometric system datasets comprised either 512x512, or 1024 x 256 pixel images, acquired frame rates of 30 and 60 Hz respectively. Motion within each frame of the imaging datasets was corrected using the method of [52] using red fluorescence from sulforhodamine 101.

Cell-attached electrophysiological recordings were made using glass pipettes (6-12 MΩ resistance) filled with ACSF containing either 2.5 µM Alexa 594 (ThermoFisher, MA, USA) or 2.5 µM Alexa 488 (ThermoFisher, MA, USA). Electrophysiological signals were amplified using a Multiclamp 700B (Molecular Devices, CA, USA), equipped with a CV 7B headstage. Signals were lowpass filtered at 10 kHz (Bessel filter) and digitized at 20 kHz (Power 1401, Cambridge Electronic Design, Cambridge, UK). Cells were visually targeted under multiphoton guidance for acquiring simultaneous imaging and cell-attached datasets.

Imaging with simultaneous visual stimulation

After an expression time of 14 days, the rat was anesthetized with Ketamine-Medetomidine (100 mg/kg and 2 mg/kg respectively) with supplemental doses of anesthetic solution given as required. A custom-made headplate fixed to the skull, a craniotomy was opened and the dura was removed with the same procedure as for simultaenous optical/electrical recordings. Visual stimuli were generated using a custom-written Matlab script and the Psychophysics Toolbox, and presented with a monitor (Faytech, 16 x 12 cm, 60 Hz refresh rate) placed 15 cm in front of the rat, covering approximately 56° x 42° of the visual field. Full-length drifting bar gratings (0.05 cycle/°, 2 cycle/s, 8 orientations, 2 s duration, 3 s inter-stimulus-interval, 100% contrast) were binocularly presented with 5 repetitions. Two-photon imaging with another custom-built microscopy visualized labelled neurons and astrocytes similarly as described above, but apart from using an objective lens (Throl Optics).

Expression and purification of GCaMP6s

GCaMP6s was expressed in E. coli BL21 (DE3) pLysS (Novagen) in LB medium supplemented with 180 µg/mL Ampicillin and 20 µg/mL Chloramphenicol for 24h at 25°C in the absence of IPTG. The pellet was resuspended in 30 mM Hepes, pH 7.2, 500 mM NaCl, 10 mM Imidazole, 1 mM PMSF, and a Roche cOmplete tablet (EDTA-free), lysed by sonication and cell debris was removed by ultracentrifugation. The cleared lysate was incubated with Ni-NTA Superflow beads (Qiagen). After washing with high-salt buffer (1000 mM NaCl), bound protein was eluted in a step gradient with elution buffer (30 mM Hepes, pH 7.8, 500 mM NaCl, and 500 mM imidazole). The pool was subjected to gel filtration (Superdex 200, GE Healthcare) in 30 mM MOPS, pH 7.2, 150 mM KCl.

Preparation of low calcium buffer (LCB)

All plastic ware and stirring bars were incubated for 30 min in 20 mM Tris (pH 8) with 50 µM EGTA, then washed 15 times with clean water to remove all traces of EGTA. Standard buffer for the measurements was 30 mM MOPS and 100 mM KCl, prepared with trace select water (Fluka), the pH was adjusted to 7.2 (final) using KOH. The buffer solutions were incubated with BAPTA-based polystyrene beads (BAPTA-sponge S, ThermoFisher) on a rotating plate for 24-36 hours at RT to reduce background calcium levels. Calcium contamination was determined using Fura-2 Thermo Fisher, F-1200) by fluorescence spectroscopy.

Calcium removal by refolding or dialysis

To obtain nearly calcium-free protein levels (calcium-free GCaMP6s) we unfolded GCaMP6s in denaturation buffer (6 M Guanidinium-HCl in 30 mM MOPS, pH 7.2, 100 mM KCl, 2 mM EGTA) followed by repeated concentration and dilution in denaturation buffer using AMICON Ultra centrifugal filters (Millipore). The protein was refolded by fast dilution into 30 mM MOPS, pH 7.2, 100 mM KCl, 2 mM EGTA, 0.5 M Arginine, and 1 M Guanidinium-HCl. The buffer was then exchanged to LCB using AMICON devices. Alternatively, calcium was removed by dialysis instead of refolding. The protein pool was dialysed in slide-a-lyzer buttons (2 K MWCO) for 48 hours at RT and protected from light, against 30 mM MOPS, 100 mM KCl, 10 mM EGTA, 10 mM EDTA, pH 7.5. The buffer was exchanged to LCB by multiple rounds of dilution and concentration at 4°C using 15 mL Amicon concentration devices (10 K MWCO). During calcium removal and in all subsequent steps, we exclusively employed chemicals and water (Fluka 95305) of the highest available grade and pre-cleaned all plastic ware.

Fluorescence spectra

GCaMP6s was diluted to a nominal concentration of 2 µM in LCB with 2 mM BAPTA and 0 to 2 mM total calcium (final concentrations taking into account the reagent purity values determined during data fitting are shown in the main and supplementary figures). Spectra were recorded at pH 7.2 at 37 C on a PTI spectrometer in photon counting mode (at photon counts below 700000, to avoid nonlinearities in the photon counting module). Excitation wavelength was varied from 350 to 505 nm. The colored spectrum shown in Figure 4d was taken from Wikimedia commons[1] and is available under the creative commons license.

Isothermal titrational calorimetry

ITC measurements were carried out in an ITC200 micro-calorimeter (MicroCal, Malvern Instruments) at 37C. The cell and Hamilton syringe were incubated for 1 hour with 20 mM Tris/HCl (pH 8) and 50 µM EDTA, washed twice with the same solution, then washed twenty times with trace select water (Sigma Aldrich). Finally, the cell and syringe were washed four times with LCB.

All solutions were prepared in LCB (pH 7.2) using non-autoclaved Eppendorf tubes and pipette tips. Protein concentrations were calculated from absorption at 280 nm.

Nominal concentrations of 400 to 1200 µM CaCl_2 was titrated into 20 to 60 µM calcium-free GCaMP6s. The experiments consisted of 18 to 30 injections with volumes between 2 and 1.2 µL, 90 s spacing, and constant stirring at 1000 rpm. The initial active cell volume was 203.4 µL. The heat of a 0.2 µL dummy injection at the beginning of each titration was not analysed but its volume and calcium content was taken into account. Baseline correction and peak integration were carried out using [68]. The integrated heat signal was corrected for the heat of dilution (obtained from reference titrations of syringe component into buffer).

[1] commons.wikimedia.org/w/index.php?title=File:Rendered_Spectrum.png&oldid=110367311

Stopped-flow fluorimetry

Measurements were carried out in an SX20 stopped-flow device (Applied Photophysics) at 37°C in LCB (pH 7.2). The injection volume was set to 145 µL for each syringe. Calcium release from GCaMP6s was measured by mixing 1 µM GCaMP6s and 20 µM total $\mathrm{CaCl_2}$ in LCB (syringe 1) with 10 mM BAPTA in LCB (syringe 2).

In order to measure calcium binding to GCaMP6s, 1 µM protein in the presence of 50 µM BAPTA was mixed with a defined free calcium concentration (2 mM BAPTA and 0-1.8 mM total calcium) in LCB. These concentrations are nominal values, whereas corrected concentrations taking into account reagent purity values were calculated during data fitting and are reported in the main text and figures.

The instrument was incubated in 50 µM EGTA in LCB over-night before experiments. Syringes were washed using trace select water four times prior to the experiment, then an additional two times using low calcium buffer. Fluorescence was excited at 488 nm (bandwidth 0.5 nm) and a 515 nm cut-off filter was used for the emission. The digitization rate was adjusted to match the speed of the observed kinetics (80 kHz for the fastest trace).

Extraction of fluorescence signals from *in vivo* imaging data

We used an feature extraction method based on nonnegative matrix factorization [76] to separate fluorescence signals from overlapping structures, similar to previously published techniques [105, 31, 89]. The aim of this method was not to detect fluorescent structures such as neurons and dendrites, but rather to isolate the signals from a single structure that was already identified and marked with a region of interest (ROI). In-depth exploration of the capabilities and limitations of this approach as compared to alternative methods are beyond the scope of the present work and will be described elsewhere (Voit et al., in final preparation).

Biophysical model

The sequential binding model is a generative, parametric biophysical model designed to describe GCaMP6s and other GECIs both *in vitro* and *in vivo*. Here we first describe the model of reaction kinetics that make up the core of the SBM, while the specific ways in which various data types depend on these kinetics are described in later sections. For a GECI G with m calcium binding sites and total concentration $[\mathrm{G}]^{\mathrm{total}}$, we model the sequential binding steps:

$$m\mathrm{Ca}^{2+} + \mathrm{G} \xrightleftharpoons[k_1^-]{k_1^+} (m-1)\mathrm{Ca}^{2+} + \mathrm{CaG} \xrightleftharpoons[k_2^-]{k_2^+} \cdots \xrightleftharpoons[k_{m-1}^-]{k_{m-1}^+} \mathrm{Ca}^{2+} + \mathrm{Ca}_{m-1}\mathrm{G} \xrightleftharpoons[k_m^-]{k_m^+} \mathrm{Ca}_m\mathrm{G} \tag{8}$$

where k_j^+ and k_j^- are the kinetic rate constants for binding and unbinding of the m-th calcium ion. The net rate of each reaction in the forward direction is then

$$r_j = r_j^+ - r_j^- = k_j^+[\mathrm{Ca}^{2+}][\mathrm{Ca}_{j-1}\mathrm{G}] - k_j^-[\mathrm{Ca}_j\mathrm{G}] \qquad 1 \le j \le m \tag{9}$$

The SBM also includes n_{buffers} additional molecules B_1, B_2, ..., $\mathrm{B}_{n_{\mathrm{buffers}}}$ binding a single calcium ion, with rates constants b_ℓ^+ and b_ℓ^- and total concentrations $[\mathrm{B}]_\ell^{\mathrm{total}}$. Finally, free calcium is extruded from the neuron at the rate r_{ex}, which for the standard SBM is

$$r_{\mathrm{ex}} = \frac{[\mathrm{Ca}^{2+}] - [\mathrm{Ca}^{2+}]_{\mathrm{rest}}}{\tau_{\mathrm{ex}}} \tag{10}$$

We also implemented a more complicated extrusion mechanism (Figure 3 - figure supplements 3-4), in which n_{ex} different extrusion reactions proceed with Michaelis-Menten kinetics. In this case, extrusion reaction p saturates at a maximum rate of v_p^{ex} and the dependence on free calcium concentration is described by a Michaelis constant K_p^{ex}, yielding the total extrusion rate

$$r_{\mathrm{ex}} = \sum_{p=1}^{n_{\mathrm{ex}}} v_p^{\mathrm{ex}} \left(\frac{[\mathrm{Ca}^{2+}]}{[\mathrm{Ca}^{2+}] + K_p^{\mathrm{ex}}} - \frac{[\mathrm{Ca}^{2+}]_{\mathrm{rest}}}{[\mathrm{Ca}^{2+}]_{\mathrm{rest}} + K_p^{\mathrm{ex}}} \right) \tag{11}$$

For both extrusion mechanisms, $r_{\mathrm{ex}} = 0$ when $[\mathrm{Ca}^{2+}] = [\mathrm{Ca}^{2+}]_{\mathrm{rest}}$. When analyzing data from *in vitro* binding assays, we set $r_{\mathrm{ex}} = 0$.

Together, these reactions define the rate equation for our model [87, 137], a system of coupled nonlinear first-order ordinary differential equations.

$$\frac{d[\mathrm{Ca}_j\mathrm{G}]}{dt} = \begin{cases} -r_1 & j=0 \\ r_{j-1} - r_j & 0<j<m \\ r_m & j=m \end{cases} \tag{12}$$

$$\frac{d[\mathrm{CaB}_\ell]}{dt} = b_\ell^+[\mathrm{Ca}^{2+}]([\mathrm{B}]_\ell^{\mathrm{total}} - [\mathrm{CaB}_\ell]) - b_\ell^-[\mathrm{CaB}_\ell] \tag{13}$$

$$\frac{d[\mathrm{Ca}^{2+}]}{dt} = -\sum_{j=1}^{m} r_j - \sum_{\ell=1}^{n_{\mathrm{buffers}}} \frac{d[\mathrm{CaB}_\ell]}{dt} - r_{\mathrm{ex}} \tag{14}$$

This system of $m + n_{\mathrm{buffers}} + 2$ equations is slightly redundant, since $\sum_{j=0}^{m}[\mathrm{Ca}_j\mathrm{G}] = [\mathrm{G}]^{\mathrm{total}}$ is fixed. However, we explicitly calculated all $m+1$ indicator binding state concentrations due to the nature of our numerical ODE solver (see below).

Equilibrium state for known free calcium concentration

For a fixed calcium concentration $[\mathrm{Ca}^{2+}]$, we calculate the equilibrium concentrations for all molecular species by setting derivatives to zero and solving the resulting system of linear equations. This gives for the GECI G the relations:

$$\frac{[\mathrm{Ca}_j\mathrm{G}]_{\mathrm{eq}}}{[\mathrm{Ca}_{j-1}\mathrm{G}]_{\mathrm{eq}}} = [\mathrm{Ca}^{2+}]k_j^+/k_j^- \qquad 1 \le j \le m \tag{15}$$

$$\frac{[\mathrm{Ca}_j\mathrm{G}]_{\mathrm{eq}}}{[\mathrm{G}]_{\mathrm{eq}}} = [\mathrm{Ca}^{2+}]^j\beta_j \qquad 1 \le j \le m \tag{16}$$

$$\beta_j = \prod_{u=1}^{j} k_u^+/k_u^- \tag{17}$$

$$[\mathrm{Ca}_j\mathrm{G}]_{\mathrm{eq}} = [\mathrm{G}]^{\mathrm{total}} \frac{[\mathrm{Ca}^{2+}]^j\beta_j}{1+\sum_{h=1}^{m}[\mathrm{Ca}^{2+}]^h\beta_h} \qquad 0 \le j \le m \tag{18}$$

Eq. 18 is known as the Adair-Klotz equation [87, 137] and the β's are macroscopic association constants. Similarly, for the buffers

$$[\mathrm{CaB}_\ell] = [\mathrm{B}]_\ell^{\mathrm{total}} \frac{[\mathrm{Ca}^{2+}]b_\ell^+/b_\ell^-}{1+[\mathrm{Ca}^{2+}]b_\ell^+/b_\ell^-} \tag{19}$$

Equilibrium state for known total calcium concentration

For certain *in vitro* experiments, the total concentration of calcium $[\mathrm{Ca}^{2+}]_{\mathrm{total}}$ is known but the free concentration $[\mathrm{Ca}^{2+}]$ at equilibrium is not. In this case we have the relation:

$$[\mathrm{Ca}^{2+}]_{\mathrm{total}} = [\mathrm{Ca}^{2+}] + \sum_{j=1}^{m}[\mathrm{Ca}_j\mathrm{G}] + \sum_{\ell=1}^{n_{\mathrm{buffers}}}[\mathrm{CaB}_\ell] \tag{20}$$

$$= [\mathrm{Ca}^{2+}] + [\mathrm{G}]^{\mathrm{total}}\frac{\sum_{h=1}^{m} h[\mathrm{Ca}^{2+}]^h\beta_h}{1+\sum_{h=1}^{m}[\mathrm{Ca}^{2+}]^h\beta_h} + \sum_{\ell=1}^{n_{\mathrm{buffers}}}[\mathrm{B}]_\ell^{\mathrm{total}}\frac{[\mathrm{Ca}^{2+}]b_\ell^+/b_\ell^-}{1+[\mathrm{Ca}^{2+}]b_\ell^+/b_\ell^-} \tag{21}$$

Multiplying by the denominators gives a polynomial equation for $[\mathrm{Ca}^{2+}]$ of degree $m + n_{\mathrm{buffers}} + 1$. Because free and total calcium concentrations are strictly increasing functions of each other, this polynomial must posses exactly one real root between zero and $[\mathrm{Ca}^{2+}]_{\mathrm{total}}$, which we obtain numerically using the NumPy function **roots**. Substituting the value of $[\mathrm{Ca}^{2+}]_{\mathrm{rest}}$ obtained from eq. 21 into eq. 18-19 gives the concentrations of all other molecular species.

Differentiability of the equilibrium state Our procedures for fitting the SBM to *in vitro* data require calculating the derivatives of equilibrium binding state concentrations with respect to model parameters. For this purpose, we can obtain the derivative of $[\mathrm{Ca}^{2+}]$ with respect to any total reagent concentrations or rate constants by implicit differentiation with respect to polynomial coefficients. Suppose that x_R is a root of some polynomial $P(x)$ with degree d. Then considering x_R as a function of the polynomial coefficients, we have for the j^{th} order coefficient p_j:

$$0 = P(x_R) \tag{22}$$

$$\frac{d}{dp_j}0 = x_R^j + \left.\frac{dP(x)}{dx}\right|_{x=x_R} \frac{dx_R}{dp_j} \tag{23}$$

$$\frac{dx_R}{dp_j} = -x_R^j \left.\left(\frac{dP(x)}{dx}\right)^{-1}\right|_{x=x_R} \tag{24}$$

The second line follows from the product and chain rules. Note when the x_R represents the free calcium concentration being calculated from total reagent concentrations, the polynomial derivative $\frac{dP(x)}{dx}$ is zero only when $x_R = 0$, i.e. when both free and total calcium concentrations are zero.

In vitro data model

We analyzed data from three *in vitro* binding assays: fluorescence spectra, isothermal titration calorimetry (ITC) and stopped-flow fluorimetry.

For spectral and stopped-flow data, the calcium buffer BAPTA [129] was used to control calcium concentration. Using the Maxchelator program [9], we calculated an initial value for the dissociation constant $K_d = b^-/b^+ = 222.8$ nM at 37 °C, ionic strength 0.15 and pH 7.2. We constrained the K_d in subsequent fitting procedures to lie within 5% of this value. While BAPTA's kinetics have not been as thoroughly investigated as its affinity, [91] reported an off-rate of $b^- = 79.0$ s^{-1} at 22 °C. We therefore initialized $b^- = 90$ s^{-1} for data at 37 °C and constrained $1/b^-$ to lie within 50% to 150% of its initial value of 11.1 ms.

Effective concentration and calcium contamination of GCaMP6s and BAPTA All *in vitro* experiments involved solutions containing the calcium sensor GCaMP6s, calcium ions and in some cases the calcium buffer BAPTA. For each of these, nominal concentration values were calculated based on dilution factors, molecular weight (for BAPTA) and UV absorption at 280 nm (for GCaMP6s). However, true concentrations can differ substantially from nominal concentrations, due imperfect theoretical predictions of protein extinction coefficients [47], increase in the mass of salts through absorption of water and variations in purity across protein purifications or batches of chemical reagents [74, 82, 9, 98]. The effects of calcium contamination can also be significant, especially when considering calcium concentrations approaching the extremely low levels ($<$ 100 nM) present in neurons at rest [116, 60, 80, 142], which we aimed to reproduce in the *in vitro* binding assays.

Therefore, for each pool of purified GCaMP6s or batch of BAPTA, we modeled the ratio of true to nominal concentration as a free parameter, which we refer to as the "purity" of the reagent. We also included for each GCaMP6s pool or BAPTA batch a "calcium contamination" parameter, defined as the ratio of excess total calcium concentration to the reagent's nominal concentration. We also included a different GCaMP6s contamination value for each ITC experiment, since the absence of BAPTA makes these experiments extremely sensitive to the handling of the protein. We did not include a purity value for the calcium added intentionally from diluted 0.1 M stock, since its concentration errors are likely to be minuscule and since rescaling all concentrations and making a compensatory change to other parameters would not change the data values predicted by the model. Thus our standard calcium stock acts as a unit of concentration or "Urmeter" for our model fits.

Data model for fluorescence spectra Spectral data were represented as a matrix F of fluorescence values, where $F_{\lambda,v}$ indicates the fluorescence detected with excitation wavelength λ in experimental condition v. For each spectroscopy experiment, the experimental conditions differed only in the amount of total calcium present. Based on the nominal reagent concentrations, contaminations and purities, we calculated the true concentrations of GCaMP6s, BAPTA and total calcium for each condition. We next calculated free calcium concentration from eq. 21 and all binding state concentrations from eq. 18. Denoting the calculated concentration of binding state i in condition v by $[\mathrm{Ca}_i\mathrm{GCaMP6s}]_v$, we modeled fluorescence according to

$$\hat{F}_{\lambda,v} = \sum_{i=0}^{m} F_{\lambda,i}[\mathrm{Ca}_i\mathrm{GCaMP6s}]_v \tag{25}$$

$F_{\lambda,i} > 0$ denotes the fluorescence of binding state i when excited by wavelength λ.

Data model for isothermal titration calorimetry Data for a single ITC experiment were represented as a vector $Q\epsilon\mathbb{R}^{n_Q}$ of integrated peak heats. We first calculated the true concentrations of calcium and GCaMP6s for the syringe and the intial solution in the reaction chamber given the nominal concentrations, purities and contaminations. For each injection, we then updated the concentrations using the standard perfusion model [38]. Specifically, for a reagent X with total concentration $[X]_0$ in the cell (after the dummy injection) and total concentration $[X]_S$ in the syringe, the concentration in the cell after the i^{th} injecion will be:

$$[X]_i = [X]_{i-1}\frac{V_{\text{cell}} - V_{\text{inj}}}{V_{\text{cell}}} + [X]_S\frac{V_{\text{inj}}}{V_{\text{cell}}} \tag{26}$$

$$= [X]_0\left(\frac{V_{\text{cell}} - V_{\text{inj}}}{V_{\text{cell}}}\right)^i + [X]_S\left[1 - \left(\frac{V_{\text{cell}} - V_{\text{inj}}}{V_{\text{cell}}}\right)^i\right] \tag{27}$$

After calculating total concentrations of GCaMP6s and calcium at each stage of the ITC experiment, we calculated free calcium concentration from eq. 21 and all binding state concentrations from eq. 18. We next calculated for each mixture the total enthalpy difference from the calcium-free state arising from binding of calcium to the indicator:

$$\Delta H = \sum_{j=1}^{m} h_j[\text{Ca}_j\text{GCaMP6s}] \tag{28}$$

where h_j is the enthalpy difference between one mole of indicator with j calcium ions bound and one mole of calcium-free indicator. Finally, denoting the enthalpy difference after i injections by ΔH_i, we model the integrated peak heat by

$$\hat{Q}_i = \Delta H_i - \Delta H_{i-1}\frac{V_{\text{cell}} - V_{\text{inj}}}{V_{\text{cell}}} \tag{29}$$

The modeled peak heat $\hat{Q}$ does not contain a term for the pre-reaction enthalpy of each injection, since the syringe did not contain GCaMP6s. Note that we do not explicitly consider heats of dilution, as the heat of dilution for the injection of concentrated calcium was already subtracted, and the dilution heat of GCaMP6s in the reaction cell is negligible compared to binding enthalpy.

Data model for stopped-flow fluorimetry Data for a single stopped-flow experiment were represented as a vector F of fluorescence observations F_i at a sequence of times after mixing. For each syringe, we calculated free calcium concentration from eq. 21 and all binding state concentrations from eq. 18, giving the initial equilibrium states before mixing. We then calculated the non-equilibrium binding state concentrations immediately after mixing as the average of the two equilibrium states (since the syringe volumes are equal). We next integrated the rate equation (12-14) from time zero to t_{dead}, representing the ”dead time” delay between initial combination of the two solutions in the mixing chamber and observations made on solution in the fluorescence cell. We then calculated the model's prediction of time-dependent fluorescence by

$$\hat{F}_i = \alpha_{\text{stopped-flow}}\sum_{j=0}^{m} F_{\lambda,j}[\text{Ca}_j\text{GCaMP6s}]_i \tag{30}$$

This formula differs from eq. 25 only in the presence of an additional scaling factor, which captures the differences between stopped-flow and spectral data in collection efficiency, excitation intensity, data units, etc. We then integrated the rate equation until the next fluorescence observation and repeated the procedure until the last data point.

In vivo data model

Time step For imaging data with a time step of Δ between image frames, the SBM used a time step of $\delta < \Delta$ to describe the AP sequence and all concentrations over time. Except where otherwise noted, we chose δ to be the largest possible value less than 10 ms for which Δ/δ is an integer.

AP discharge Each neuron's AP sequence was modeled as an independent Bernoulli variable for each time step. The probability of AP discharge was

$$p_{\text{AP}} = \delta\chi \tag{31}$$

where χ is the neuron's firing rate. The prior distribution of firing rate over neurons was modeled as a Gamma distribution $\chi \sim \Gamma(k_\chi, \theta_\chi)$, where k_χ is a shape parameter and θ_χ is a scale parameter. We modeled each AP as an instantaneous increase in $[\text{Ca}^{2+}]$ by $\Delta[\text{Ca}^{2+}]_{\text{AP}}$.

Fluorescence model Fluorescent measurements are acquired as the neuron's soma is scanned by the laser focus of the two-photon microscope. The infrared laser light excites the fluorophore leading to emission of visible photons, some of which are captured by the objective lens and detected by the photomultiplier tube (PMT). We model the expected value of the i^{th} fluorescence measurement by:

$$\mathbb{E}[F_i] = \hat{F} = (F_{\text{cyt}} + F_{\text{BG}})\frac{B_i}{\phi_0[\text{GCaMP6s}]^{\text{total}}} \tag{32}$$

$$F_{\text{cyt}} = \sum_{j=0}^{m} \phi_j[\text{Ca}_j\text{GCaMP6s}]_i \tag{33}$$

Neither the values of ϕ_j nor their ratios need be consistent with the values of $F_{\lambda,j}$ for *in vitro* data, due to the difference between one- and two-photon excitation. F_{BG} represents fluorescence from out-of-focus structures that does not depend on binding state concentrations in the neuron. B_i represents many factors influencing the scale of the fluorescence measurement, some of which may vary over time. These include laser power, light scattering and absorption by tissue, changes in blood flow through vessels above the image plane, photobleaching of the fluorophore, collection efficiency, artifacts arising from uncorrected lateral or axial motion, PMT quantum efficiency, amplifier gain and the analog-to-digital (A2D) converter's input range and bit depth. This formulation does not include an additive offset in the fluorescence values, and thus assumes that the recorded fluorescence values will on average be zero in the absence of incoming photons. This can be achieved by subtracting from the fluorescence data an offset recorded with the laser shutter closed before or after each imaging file, a standard feature in many imaging systems such as Scanimage [106].

To ensure that B_i is positive while allowing it to drift over time, we define $B_i = e^{\rho}$ and modeled ρ as a Brownian motion with variance σ_ρ^2 sec^{-1}. That is, given the value of ρ at time t_1 seconds, its value at a later time t_2 has conditional distribution

$$\rho(t_2)|\rho(t_1) \sim \mathcal{N}(\rho(t_1), (t_2 - t_1)\sigma_\rho^2) \tag{34}$$

We set $\sigma_\rho = 0.01$ for the analysis of all *in vivo* data.

Finally, we define $\psi_j = \frac{\phi_j - \phi_0}{\phi_0}$ to give a parameterization that does not depend on the units of fluorescence or concentration:

$$\hat{F} = \left(1 + \sum_{j=1}^{m} \psi_j \frac{[\text{Ca}_j\text{GCaMP6s}]}{[\text{GCaMP6s}]^{\text{total}}} + \frac{F_{\text{BG}}}{\phi_0[\text{GCaMP6s}]^{\text{total}}}\right) B_i \tag{35}$$

$$\psi_j = \frac{\phi_j - \phi_0}{\phi_0} \tag{36}$$

We assume that calcium binding does not decrease the indicator's brightness, so each $\psi_j > 0$.

Joint prior distribution for $[\textbf{GCaMP6s}]^{\textbf{total}}$ and $F_{\textbf{BG}}$ Since resting calcium concentration is consistently 50-80 nM [116, 60, 80, 142], neurons with higher GECI concentrations would be expected to have higher brightness relative to the background fluorescence at rest. We therefore modeled the joint distribution over neurons on $[\text{GCaMP6s}]^{\text{total}}$ and $F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}}$ using a 2D log-normal prior:

$$\begin{bmatrix} \log([\text{GCaMP6s}]^{\text{total}}) \\ \log(F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}}) \end{bmatrix} \sim \mathcal{N}(\mu_G, \Sigma_G) \tag{37}$$

where $F_{\text{cyt}}^{\text{eq}} = \sum_{j=0}^{m} \phi_j[\text{Ca}_j\text{GCaMP6s}]_{\text{eq}}$, with equilibrium concentrations calculated for $[\text{Ca}^{2+}] = [\text{Ca}^{2+}]_{\text{rest}}$. Note that $[\text{Ca}^{2+}]_{\text{rest}}$ is not an estimated parameter, but has been fixed to 50 nM (see results).

Fluorescence observation noise For an expected fluorescence value $\hat{F}$ predicted by the SBM, we modeled the probability distribution on fluorescence observations as

$$F \sim \mathcal{N}\left(\hat{F}, \frac{g\hat{F} + \sigma_F^2}{n_{A2D}}\right) \tag{38}$$

Where g is an unknown gain factor used to incorporate photon shot noise and other signal-dependent noise, n_{A2D} is the number of analog-to-digital values averaged when calculating the neuron's fluorescence for each image frame and σ_F represents signal-independent noise. We implemented our model fitting and AP inference algorithms with the option to incorporate calibrated values of g and σ_F, but when estimating noise properties from fluorescence data alone (see below), we set $g = 0$ and estimated only σ_F. Thus for SBM-based AP inference in this study, n_{A2D} is known, $g = 0$ and a different value of σ_F is estimated for each neuron based on its fluorescence signals.

Numerical solution of ODEs

Model fitting and AP inference with the SBM require integrating the rate equation (12-14) forward in time to calculate predicted fluorescence given an AP sequence. When fitting models to *in vitro* data, we used a standard ODE solver available as SciPy library function. We chose the LSODA solver [63] based on the backward differentiation formulas [6] with adaptive step size determination, since we did not know the time constants of the various binding steps in advance and since the stopped-flow data achieved measurement rates as high as 80 kHz. Another reason to use adaptive step sizes is that *in vitro* experiments can produce much larger and more rapid changes in free calcium than those observed in neurons.

However, fitting *in vivo* data required us to design and use a custom solver. This is because when we use our model to detect APs in a sequential Monte Carlo framework (see below), we will be carrying out many simulations in parallel using a GPU implementation. In order to run efficiently on a GPU, these simulations must maintain a low memory footprint per thread, must not allocate and deallocate memory and must avoid branch points such as ”IF statements” (to avoid desynchronizing parallel simulations and causing the threads in non-active branches to wait). We therefore designed a custom ODE solver for use with AP inference on the GPU. We also used our custom ODE solver when fitting SBM parameters to *in vivo* fluorescence signals together with known AP sequences, both for speed and to ensure that any approximations or errors would be apparent in the model fits.

Backward Euler method

Let $y \epsilon \mathbb{R}^8$ be the vector of binding state concentrations:

$$y = \begin{bmatrix} [\mathrm{Ca}^{2+}] \\ [\mathrm{G}] \\ [\mathrm{Ca_1GCaMP6s}] \\ [\mathrm{Ca_2GCaMP6s}] \\ [\mathrm{Ca_3GCaMP6s}] \\ [\mathrm{Ca_4GCaMP6s}] \\ [\mathrm{CaB_1}] \\ [\mathrm{CaB_2}] \end{bmatrix} \tag{39}$$

We designed our solver to use the backward Euler method, defined by:

$$y(t+\delta) = y(t) + \delta f(y(t+\delta)) \tag{40}$$

where $f(y)$ is the vector of derivatives $\frac{dy}{dt}$, in our case defined by the rate equation (12-14). The backward Euler method possesses several important stability properties [16], allowing us to use a relatively large step size of 10 ms without strongly affecting the fluorescence predicted by the model (Figure 5 - figure supplement 1).

Applying eq. 40 requires solving a system of nonlinear equations, for which we use Newton's method, an iterative technique based on linearizing f using its Jacobian J. Given a current estimate $y^i(t+\delta)$ of $y(t+\delta)$, we approximate

$$f(y(t+\delta)) \approx f(y^i(t+\delta)) + J(y(t+\delta) - y^i(t+\delta)) \tag{41}$$

$$J = \partial f / \partial y \Big|_{y=y^i(t+\delta)} \tag{42}$$

We then use this approximation to repeatedly refine the estimate of $y(t+\delta)$ by solving a system of linear equations. Initializing with $y^0(t+\delta) = y(t)$, we have:

$$y(t+\delta) \approx y(t) + \delta \left[f(y^i(t+\delta)) + J(y(t+\delta) - y^i(t+\delta)) \right] \tag{43}$$

$$(I - \delta J)\left(y(t+\delta) - y^i(t+\delta)\right) \approx y(t) - y^i(t+\delta) + \delta f(y^i(t+\delta)) \tag{44}$$

Setting $\Delta y = y(t+\delta) - y^i(t+\delta)$ and $z = y(t) - y^i(t+\delta) + \delta f(y^i(t+\delta))$ gives:

$$(I - \delta J)\Delta y = z \tag{45}$$

$$\tag{46}$$

solving for Δy then allows us to compute the next iterative estimate $y^{i+1}(t+\delta) = y^i(t+\delta) + \Delta y$.

In order to solve this linear system with a low memory footprint and no branch points, we introduce a direct solver that generalizes the tridiagonal matrix algorithm (see below). To avoid additional branch points we also fixed the step size to 10 ms and the number of Newton iterations to 3. We show that this does not lead to inaccurate predictions of *in vivo* GCaMP6s fluorescence even at high firing rates (Figure 5 - figure supplement 1), though a shorter step size or a greater number of Newton iterations might be needed for larger calcium fluxes or faster sensors.

Jacobian

To apply the backward Euler method, we use the following Jacobian matrix:

$$J = \frac{df(y)}{dy} = \begin{bmatrix} J_{11} & J_{12} & J_{13} & J_{14} & J_{15} & J_{16} & J_{17} & J_{18} \\ J_{21} & J_{22} & J_{23} & 0 & 0 & 0 & 0 & 0 \\ J_{31} & J_{32} & J_{33} & J_{34} & 0 & 0 & 0 & 0 \\ J_{41} & 0 & J_{43} & J_{44} & J_{45} & 0 & 0 & 0 \\ J_{51} & 0 & 0 & J_{54} & J_{55} & J_{56} & 0 & 0 \\ J_{61} & 0 & 0 & 0 & J_{65} & J_{66} & 0 & 0 \\ J_{71} & 0 & 0 & 0 & 0 & 0 & J_{77} & 0 \\ J_{81} & 0 & 0 & 0 & 0 & 0 & 0 & J_{88} \end{bmatrix} \tag{47}$$

where the elements of J are defined as follows for $0 \leq j \leq 4$ and $1 \leq \ell \leq 2$:

$$J_{11} = \frac{\partial^2[\mathrm{Ca}^{2+}]}{\partial t \partial[\mathrm{Ca}^{2+}]} = -\sum_{j=1}^{4} k^+[\mathrm{Ca}_{j-1}\mathrm{GCaMP6s}] - \sum_{\ell=1}^{2} b_\ell^+([\mathrm{B}]_\ell^{\mathrm{total}} - [\mathrm{CaB}_\ell]) - \frac{\partial r_{\mathrm{ex}}}{\partial[\mathrm{Ca}^{2+}]} \tag{48}$$

$$J_{1,j+1} = \frac{\partial^2[\mathrm{Ca}^{2+}]}{\partial t \partial[\mathrm{Ca}_j\mathrm{GCaMP6s}]} = k_j^- - k_{j+1}^+[\mathrm{Ca}^{2+}] \tag{49}$$

$$J_{1,\ell+6} = \frac{\partial^2[\mathrm{Ca}^{2+}]}{\partial t \partial[\mathrm{CaB}_\ell]} = b_\ell^- + b_\ell^+[\mathrm{Ca}^{2+}] \tag{50}$$

$$J_{j+1,1} = \frac{\partial^2[\mathrm{Ca}_j\mathrm{GCaMP6s}]}{\partial t \partial[\mathrm{Ca}^{2+}]} = k_j^+[\mathrm{Ca}_{j-1}\mathrm{GCaMP6s}] - k_{j+1}^+[\mathrm{Ca}_j\mathrm{GCaMP6s}] \tag{51}$$

$$J_{j+1,j} = \frac{\partial^2[\mathrm{Ca}_j\mathrm{GCaMP6s}]}{\partial t \partial[\mathrm{Ca}_{j-1}\mathrm{GCaMP6s}]} = k_j^+[\mathrm{Ca}^{2+}] \tag{52}$$

$$J_{j+1,j+1} = \frac{\partial^2[\mathrm{Ca}_j\mathrm{GCaMP6s}]}{\partial t \partial[\mathrm{Ca}_j\mathrm{GCaMP6s}]} = -k_j^- - k_{j+1}^+[\mathrm{Ca}^{2+}] \tag{53}$$

$$J_{j+1,j+2} = \frac{\partial^2[\mathrm{Ca}_j\mathrm{GCaMP6s}]}{\partial t \partial[\mathrm{Ca}_{j+1}\mathrm{GCaMP6s}]} = k_{j+1}^- \tag{54}$$

$$J_{\ell+6,1} = \frac{\partial^2[\mathrm{CaB}_\ell]}{\partial t \partial[\mathrm{Ca}^{2+}]} = b_\ell^+([\mathrm{B}]_\ell^{\mathrm{total}} - [\mathrm{CaB}_\ell]) \tag{55}$$

$$J_{\ell+6,\ell+6} = \frac{\partial^2[\mathrm{CaB}_\ell]}{\partial t \partial[\mathrm{CaB}_\ell]} = -b_\ell^- - b_\ell^+[\mathrm{Ca}^{2+}] \tag{56}$$

with the convention that $k_0^- = k_5^+ = [\mathrm{Ca}_{-1}\mathrm{GCaMP6s}] = 0$. For the standard SBM $\frac{\partial r_{\mathrm{ex}}}{\partial[\mathrm{Ca}^{2+}]} = 1/\tau_{\mathrm{ex}}$, while with Michaelis-Menten extrusion $\frac{\partial r_{\mathrm{ex}}}{\partial[\mathrm{Ca}^{2+}]} = \sum_{p=1}^{n_{\mathrm{ex}}} \frac{v_p^{\mathrm{ex}} K_p^{\mathrm{ex}}}{([\mathrm{Ca}^{2+}]+K_p^{\mathrm{ex}})^2}$.

Generalized tridiagonal solver

Except for its first row and column, J is a tridiagonal matrix. Here we describe an $O(n)$ direct method (algorithm 1) for solving such a system of linear equations, generalizing the tridiagonal algorithm [25] to the present case. We consider the $(n+1)$-dimensional linear system

$$Mx = z \tag{57}$$

where

$$M = \begin{bmatrix} m & v_1 & v_2 & v_3 & \cdots & v_{n-2} & v_{n-1} & v_n \\ u_1 & b_1 & c_1 & 0 & \cdots & 0 & 0 & 0 \\ u_2 & a_1 & b_2 & c_2 & \cdots & 0 & 0 & 0 \\ u_3 & 0 & a_2 & b_3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ u_{n-2} & 0 & 0 & 0 & \cdots & b_{n-2} & c_{n-2} & 0 \\ u_{n-1} & 0 & 0 & 0 & \cdots & a_{n-2} & b_{n-1} & c_{n-1} \\ u_n & 0 & 0 & 0 & \cdots & 0 & a_{n-1} & b_n \end{bmatrix} \tag{58}$$

$$z = \begin{bmatrix} w \\ d_1 \\ \vdots \\ d_n \end{bmatrix} \tag{59}$$

Following the original tridiagonal algorithm, we make a forward pass that sets a to zero and b to one, then a backward pass that sets c to zero. During the backward pass, we also set v to zero, and during both passes we keep track of the resulting changes to u, v, w, m and d. Finally, we make another forward pass that sets u to zero while changing only d, at which point we have reduced M to the identity matrix so that we can return the modified z as the desired solution x. Elements set to zero or one are not actually modified in memory since they will not be accessed again.

Algorithm 1 Generalized tridiagonal solver

```
$c_1 \leftarrow c_1/b_1$
$d_1 \leftarrow d_1/b_1$
$u_1 \leftarrow u_1/b_1$
for $i = 2$ to $n$ do
    $p \leftarrow b_i - a_i c_{i-1}$
    $d_i \leftarrow (d_i - a_i d_{i-1})/p$
    $u_i \leftarrow (u_i - a_i u_{i-1})/p$
    if $i < n$ then
        $c_i \leftarrow c_i/p$
    end if
end for
for $i = n - 1$ to $0$ do
    if $i > 0$ then
        $d_i \leftarrow d_i - c_i d_{i+1}$
        $u_i \leftarrow u_i - c_i u_{i+1}$
    end if
    $m \leftarrow m - v_{i+1} u_{i+1}$
    $w \leftarrow w - v_{i+1} d_{i+1}$
end for
$w \leftarrow w/m$
for $i = 1$ to $n$ do
    $d_i \leftarrow d_i - w u_i$
end for
return $\begin{bmatrix} w \\ d \end{bmatrix}$
```

In our implementations we unroll the loops and use different variable names for each vector element, so that no array indexing or branch statements remain. We also skip the absent off-diagonal elements for endogenous buffers.

Fitting SBM parameters to *in vitro* binding assay data

We developed procedures to perform global fitting of all *in vitro* parameters and implemented them in the Python programming language. The rate equation and its Jacobian with respect to all molecular species' concentrations were implemented in Python and NumPy, then compiled using Numba for speed.

Objective function

We minimized the weighted sum of squares

$$e_{\text{vitro}}(\Theta) = \omega_{\text{spectra}} \sum_{u \epsilon U_{\text{spectra}}} (\hat{X}_u(\Theta) - X_u)^2 + \omega_{\text{ITC}} \sum_{u \epsilon U_{\text{ITC}}} (\hat{X}_u(\Theta) - X_u)^2 + \omega_{\text{stopped-flow}} \sum_{u \epsilon U_{\text{stopped-flow}}} (\hat{X}_u(\Theta) - X_u)^2 \quad (60)$$

Θ is a vector of SBM parameters, u is an index over all observed data values X_u, U_d is the set of data values belonging to each data modality and ω_d is a weighting factor for data modality. The ω's were chosen so that each data type would make approximately the same contribution to the objective function despite different data units, noise levels and numbers of observations. We first fit data of each type (spectra, ITC and stopped-flow) independently to determine the r.m.s. residual σ_d for each. We then normalized each data type by the corresponding squared residual, and further normalized by the number of observation for each data type:

$$\omega_d^{-1} = N_d \sigma_d^2 \quad (61)$$

where N_d is the number of data values for data type d. Thus our objective function is not affected by rescaling the data units or by adding multiple copies of the data for one data type. It also weights noisier measurements less heavily when fitting the parameters.

Gradients

Gradients were calculated in closed form for ITC and spectral data, and using finite differences for stopped-flow data. Finite differences were calculated using **SciPy**'s **approx_derivative** function, using the default step size and central (3-point) differences. For parameter values at the upper or lower bounds, **approx_derivative** calculates derivatives using the tangent to a 3-point quadratic fit.

Profiling over coefficients

For all three types of data, the value predicted by the model can be written as a matrix product of regressors and coefficients:

$$\hat{X} = Ac \quad (62)$$

where rows of X are data points. For spectral and stopped-flow data, each row of A contains the sensor's binding state concentrations for one data point, whereas for ITC data each row contains concentration changes after each injection for binding states with at least one calcium ion bound. For stopped-flow and ITC data, c is a vector of fluorescence values or enthalpy differences for each binding state, whereas for spectral data c is a matrix whose rows contain the excitation spectra of the binding states. Note that for clarity in eq. 62 we have ignored the weights ω but the results generalize in a trivial way to the case where weights are present.

Instead of directly optimizing over all model parameters, we can first calculate A as a nonlinear function of the rate constants, reagent purities, etc. and then optimize $\hat{c} = \left(A^T A\right)^{-1} A^T X$ by linear regression. In the case where bounds on c prevent us from using a simple linear regression, we use **SciPy**'s **optimize.nnls** function (where only a positivity constraint exists) or **optimize.lsq_linear** (for arbitrary upper and lower bounds). In either case, we obtain the global optimum on c since we are minimizing a convex function on a convex set. This allows us to greatly reduce the number of free parameters, especially for spectral data with many excitation wavelengths. It also speeds convergence by avoiding small alternating updates of c and A.

Let $\Theta = [\Theta_A \;\; \Theta_c]^T$, where A depends only on Θ_A and Θ_c is simply c in vector form. As long as the columns of A are linearly independent, $\hat{c}$ is well defined and we have

$$e_c(\Theta_A) = \underset{c}{\operatorname{argmin}} \|X - Ac\|_2^2 = \|X - A\hat{c}\|_2^2 \quad (63)$$

and

$$\frac{de_c(\Theta_A)}{d\Theta_A} = \frac{\partial \|X - Ac\|_2^2}{\partial \Theta_A} + \left.\frac{\partial \|X - Ac\|_2^2}{\partial c}\right|_{c=\hat{c}} \frac{d\hat{c}}{d\Theta_A} \quad (64)$$

By definition, each element of $\hat{c}$ is either at a local optimum or at a bound. For an element $\hat{c}_i$ at an optimum, $\frac{\partial \|X-Ac\|_2^2}{\partial c_i} = 0$. Elements at a bound and not at an optimum will not change upon an infinitesimal change in Θ_A, so for these elements $\frac{d\hat{c}_i}{d\Theta_A} = 0$. Consequently, the second term on the right hand side of eq. 64 is a dot product whose additive sub-terms are all zero. Thus, after optimizing with respect to c we can calculate the gradients of the objective function with respect to Θ_A while treating $c = \hat{c}$ as a constant.

Using eq. 64 does not merely speed up the calculation of gradients. When using an iterative solver (see below), the current gradient and approximation to the Hessian matrix are used to compute a search direction, which is used as the basis for a line search. However, by optimizing out the coefficients we can make larger steps in Θ_A since c will be automatically adjusted during the line search to minimize the objective function.

Reparameterization

We used an alternative parameterization for the kinetic properties of GCaMP6s that possessed two advantages over on- and off-rates. First, it allowed us to separate equilibrium and kinetic responses into separate groups of parameters, so that fits to spectral and ITC data would depend on a smaller number of parameters. Second, since the transitions through unstable intermediate binding states might be extremely rapid, the values of some rate constants might be quite large. This is problematic since we do not want to set a limit that may reduce the expressiveness of our model, but very large or small rate constants may lead to numerical instability when integrating the rate equation or solving for the equilibrium state.

For the case of a single binding site, a natural alternative is to use the dissociation constant $K_d = k^-/k^+$ to describe affinity and the time constant $\tau_{\text{obs}} = k^- + K_d k^+ = 2k^-$ to describe kinetics. Both parameters have simple, intuitive interpretations: K_d is the ligand concentration at which the binding site is half-saturated, and the τ_{obs} is the time constant with which the difference in saturation from 0.5 decays to 0 when free ligand concentration is fixed to the K_d. K_d's can also be calculated for molecules such as GCaMP that bind multiple ligands, in which case the j^{th} $K_d = k_j^-/k_j^+$ is the ligand concentration at which the $(j-1)^{\text{th}}$ and j^{th} binding states have equal concentrations. However, this equality may occur at a ligand concentration where both the $(j-1)^{\text{th}}$ and j^{th} binding sites are almost totally filled or totally empty, so the K_d's do not tell us the concentrations at which the binding actually happens.

We therefore introduce a new quantity K_j^{50}, defined as the free ligand concentration at which precisely half the molecules have j or more ligands bound. Thus by eq. 18 we have the relations:

$$\sum_{h=0}^{j-1} (K_j^{50})^h \beta_h = \sum_{h=j}^{m} (K_j^{50})^h \beta_h \tag{65}$$

Therefore, given the β's we can determine each K^{50} by solving a polynomial equation, and given the K^{50}'s we can determine the β's by solving a system of linear equations. Unlike for the K_d's, if we know that the molecule transitions from the apo to the saturated state occur over some ligand concentration range, then all the K^{50}'s must lie in this range.

To describe the kinetics, we define $\tau_j = (k_j^- + k_j^+ K_j^{50})^{-1}$, which is the time constant with which the $(j-1)^{\text{th}}$ and j^{th} binding states would reach equilibrium if the ligand concentration where set to K_j^{50} and all other binding state transitions were ignored.

Based on exploratory fits with multiple initializations, we set following loose constraints on the parameters: 10 nM $\leq K_1^{50} \leq$ 500 nM, 1 nM $\leq K_j^{50} - K_{j-1}^{50} \leq$ 300 nM, 0.1 ms $\leq \tau_j \leq$ 300 ms. Note that at any particular moment the actual kinetics of a binding step can be faster or slower than the corresponding τ, as $[\text{Ca}^{2+}]$ may be greater or lesser than the corresponding K^{50}.

Initialization

Purities were initialized to 1, BAPTA contamination was initialized to 0 and GCaMP6s contamination was initialized to 1 (i.e. 25% saturation). Dead time was initialized to 1.5 ms. Values for the K^{50}'s and τ's were initialized randomly 50 times within their limits using latin hypercube sampling [83] as implemented in the **lhs** function of the **pyDOE** package. For each initialization, the K^{50}'s and τ's were fixed for the first 10 l-bfgs-b iterations.

Optimization

We minimized the objective function with the **minimize** function in the **optimize** module of SciPy, using the l-bfgs-b algorithm with options maxls = 100, mintol = 0.0 and maxcor = 20. To ensure that poor approximations of the Hessian matrix did not lead to false convergence, we repeatedly restarted the minimization function until the library reported convergence in a single iteration. We did not limit the number of iterations. Optimization with multiple initializations was parallelized using the **multiprocessing** python module.

Optimization proceeded with the following parameter bounds: 10 nM $\leq K_1^{50} \leq$ 500 nM, 2.5 nM $\leq K_j^{50} \leq$ 1.5 µM for $j > 1$, 1 ms $\leq \tau_j \leq$ 5 s, $F_{\lambda,i} > 0$, $1\,\text{ms} \leq t_{\text{dead}} \leq 2\,\text{ms}$. Based on previous studies, purity values were limited to the range 0.9 to 1.1 for GCaMP6s [47] and 0.75 to 1 for BAPTA [98]. Calcium contamination was limited to the range 0 to 0.01 for BAPTA and 0 to 10 for GCaMP6s.

Fitting SBM parameters to *in vivo* fluorescence signals using known AP times

We developed procedures to perform global fitting of all SBM parameters to *in vivo* data and implemented them in the MATLAB programming language (Mathworks). We integrated the rate equation (12-14) using our custom ODE solver (see above), implemented through C MEX functions with openMP for parallelization over multiple fluorescence time series. We fit separate values of $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG} for each neuron. We fit our model only to fluorescence measurements with at least 1 AP in the preceding 20 seconds; other data were not included in the objective function (but were still used when calculating baseline fluorescence). Based on preliminary fits, we set $n_{\text{buffers}} = 2$. We explore other numbers of buffers in Figure 3 - figure supplements 3-4.

Normalization

Before further analysis, each sequence of fluorescence measurements was normalized by its median value. This allowed us to combine recordings from multiple microscope configurations, zoom factors and imaging rates for the same neuron even when absolute fluorescence values differed across these recordings. In the following material on model fitting procedures, F_i refers to these normalized measurements.

Simulation time steps

For neurons with data at multiple imaging rates, different values of δ were calculated for each fluorescence time series. The time of each fluorescence observation t_j was defined as the time when the laser focus passed through the center of the neuron.

All electrically detected AP times were rounded to the nearest simulation time step. For each simulation time step, we first incremented $[\text{Ca}^{2+}]$ by $\Delta[\text{Ca}^{2+}]_{\text{AP}}$ for each AP present, then integrated the rate equation (12-14) forward in time by δ. The first simulation time step begins Δ before the first fluorescence measurement (but much earlier for AP inference; see below). After every Δ/δ simulation time steps, we encounter a fluorescence observation for which we generate a predicted value based on the current binding state concentrations.

For example, if $\Delta = 99$ ms, then $\delta = 9.9$ ms, the first simulation time step begins 99 ms before the first fluorescence observation and the first prediction of a fluorescence value occurs after 10 ODE integrations of 9.9 ms each.

An exception to the above occurred when electrically recorded APs occurred before the first fluorescence measurement. In this case we included additional simulation time steps with the same value of δ, so that all APs up to 20 seconds before the first fluorescence measurement were included.

Objective function

We fit SBM parameters to *in vivo* data by minimizing the mismatch between predicted and observed values for normalized fluorescence. In order to weight all neurons in our datasets equally, we averaged square residuals over time for each neuron and then summed the result over neurons:

$$e_{\text{vivo}}(\Theta) = \sum_{\text{neurons}} \mathbb{E}_i(\hat{F}_i - F_i)^2 \tag{66}$$

where the expectation with respect to i is an average over images frames for each neuron. We chose this objective function to ensure that all neurons contributed equally to the objective function, regardless of recording duration or imaging frame rate.

Model prediction of fluorescence using known baseline fluorescence values

The formula for model-predicted fluorescence (eq. 35) contains the unknown, time-varying scaling factor B_i. To avoid having to explicitly determine the B_i for each time point, we use the concept of baseline fluorescence F_{BL}, defined as the fluorescence that would be observed for a neuron at rest, with $[\text{Ca}^{2+}] = [\text{Ca}^{2+}]_{\text{rest}}$ and all binding state concentrations at their equilibrium values. Thus we have by eq. 35

$$F_{\text{BL}} = \left(1 + \sum_{j=1}^{m} \psi_j \frac{[\text{Ca}_j\text{GCaMP6s}]_{\text{EQ}}}{[\text{GCaMP6s}]^{\text{total}}} + \frac{F_{\text{BG}}}{\phi_0 [\text{GCaMP6s}]^{\text{total}}}\right) B_i \tag{67}$$

With $F_{\text{cyt}}^{\text{eq}} = \sum_{j=0}^{m} \phi_j [\text{Ca}_j\text{GCaMP6s}]_{\text{EQ}}$, we can derive the relation

$$\frac{F_{\text{cyt}} + F_{\text{BG}}}{F_{\text{cyt}}^{\text{eq}} + F_{\text{BG}}} = \frac{\hat{F}\phi_0[\text{GCaMP6s}]^{\text{total}}/B_i}{F_{\text{BL}}\phi_0[\text{GCaMP6s}]^{\text{total}}/B_i} = \frac{\hat{F}}{F_{\text{BL}}} \tag{68}$$

This quantity is simply the ratio of predicted fluorescence to baseline fluorescence, leading to eq. 7. Thus, if we known F_{BL} we can calculate $\hat{F}$ without explicitly calculating B.

Estimation of baseline fluorescence

For each contiguous sequence of T fluorescence measurements, we estimated baseline drift using fluorescence values and electrically detected APs. We first identified all fluorescence measurements $> 1.5\Delta$ before subsequent APs and > 6 seconds after previous APs, for which a binary mask M^0 was set to 1. To initialize baseline fluorescence, we interpolated and smoothed fluorescence measurements, using a kernel K_b consisting of a sum of two Gaussian functions with standard deviations 5 seconds and 500 ms. We calculated

$$b_0 = \frac{(M^0 F) \circ K_b}{M^0 \circ K_b} \tag{69}$$

where the multiplication and division are carried out elementwise.

This initial estimate showed less perturbation due to AP discharges than a rolling average or median, but still increased during long AP bursts. To refine it further, we minimized an objective function e_{BL} that penalized both mismatch between data and predictions as well as excessive baseline fluctuations, while assuming exponential decay of fluorescence to baseline starting 5.5 seconds after AP activity.

We first computed a second binary mask M^1, defined in the same way as M^0 but starting 5.5 seconds after each AP instead of 6. Let the j-th contiguous sequence of time points for which $M^1 = 1$ be denoted G_j, and let $i_j^0 = \min(G_j)$ be the first time point in G_j. For each such sequence we introduce a free variable A_j representing the initial ratio of fluorescence to baseline for the first time point in G_j. We then sought to minimize the objective function

$$e_{\mathrm{BL}} = \sum_j \sum_{i \epsilon G_j} \left(F_i - \beta - b_i(1 + A_j e^{-(i-i_j^0)/\tau_{\mathrm{BL}}})\right)^2 + \frac{1}{\sigma_{\mathrm{BL}}^2} \sum_{i=2}^{T} (b_i - b_{i-1})^2 \tag{70}$$

$$= \sum_i M_i^1 \left(F_i - \beta - b_i \nu_i\right)^2 + \frac{1}{\sigma_{\mathrm{BL}}^2} b^T Q b \tag{71}$$

Where $Q_{00} = Q_{TT} = 1$, $Q_{ii} = 2$ for $1 < i < T$, $Q_{i,i+1} = Q_{i+1,i} = -1$ and $Q_{ik} = 0$ otherwise. We have also defined $\nu_t = A_j e^{-(t-t_j^0)/\tau_{\mathrm{BL}}}$ if t is in some G_j and $\nu_t = 1$ otherwise. We used $\sigma_{\mathrm{BL}}^2 = \delta/(4 \text{ minutes})$, where δ is the time between fluorescence measurements. That is, we expect the baseline to drift with a standard deviation equal to its starting value every four minutes.

The free parameters of this objective function are the baseline values $b_i > 0$, the offset β, the initial amplitudes for each segment $A_j > 0$ and the time constant $\tau_{\mathrm{BL}} > 0$. The objective function consists of squared residual terms over time points for which $M^1 = 1$, as well as a regularization terms over all time points.

We alternated between minimizing e_{BL} over $(\{A_j\}, \tau_{\mathrm{BL}})$, β and b. We first optimized jointly over τ_{BL} and all the A_j. We used golden section search with parabolic interpolation over τ_{BL} (Matlab's **fminbnd** function), while calculating the optimal value for each A_j in closed form given τ_{BL} using a rectified linear regression. τ_{BL} was initialized at 500 ms.

We then updated β in closed form to

$$\beta = \mathbb{E}_{M_i^1 = 1} \left(F_i - b_i s_i\right) \tag{72}$$

We then minimized e_{BL} over $b > 0$. If we ignore the positivity constraint, setting the derivative to zero yields the system of linear equations.

$$\left(\mathrm{diag}\left(\begin{bmatrix} M_1^1 \nu_1^2 \\ M_2^1 \nu_2^2 \\ \vdots \\ M_T^1 \nu_T^2 \end{bmatrix} \right) + \frac{1}{\sigma_{\mathrm{BL}}^2} Q \right) b = \begin{bmatrix} M_1^1 \nu_1 F_1 \\ M_2^1 \nu_2 F_2 \\ \vdots \\ M_T^1 \nu_T F_T \end{bmatrix} \tag{73}$$

We accept this solution for b if it is nonnegative for all t. Otherwise, we must solve a quadratic programming problem (we implemented this contingency using MATLAB's **quadprog** function, but this code was never called for baseline fitting of any data).

We iterated these alternating updates until the improvement in e_{BL} after all three updates together was less than machine precision (MATLAB's **eps** function, about 2.2e-16), or when the improvement was less than 0.001% of the previous objective function value. Computation time for baseline fitting was negligible compared to the overall model-fitting procedure.

Finally, we calculate $F_{\mathrm{BL}} = b + \beta$. The results of this model fitting procedure are shown in Figure 3 - figure supplement 1.

Table 3: Multiple initializations for K_j^{50}

Initialization	equilibrium binding state sequence	K_1^{50} (nM)	K_2^{50} (nM)	K_3^{50} (nM)	K_4^{50} (nM)
1	$0 \to 4$	350	352.5	355	357.5
2	$0 \to 3 \to 4$	100	102.5	105	600
3	$0 \to 2 \to 4$	100	102.5	597.5	600
4	$0 \to 2 \to 3 \to 4$	100	102.5	351.3	600
5	$0 \to 1 \to 4$	100	595	597.5	600
6	$0 \to 1 \to 3 \to 4$	100	348.8	351.3	600
7	$0 \to 1 \to 2 \to 4$	100	348.8	597.5	600
8	$0 \to 1 \to 2 \to 3 \to 4$	100	266.7	433.3	600

Closed-form solution for F_{BG}

When all other parameters of the model are fixed, we can minimize e_{vivo} (eq. 66) with respect to F_{BG} in closed form. For each neuron, we have by eq. 68

$$\begin{aligned} F - \hat{F} &= (F - F_{\mathrm{BL}}) - (\hat{F} - F_{\mathrm{BL}}) && (74) \\ &= (F - F_{\mathrm{BL}}) - F_{\mathrm{BL}} \left(\frac{F_{\mathrm{cyt}} + F_{\mathrm{BG}}}{F_{\mathrm{cyt}}^{\mathrm{eq}} + F_{\mathrm{BG}}} - 1 \right) && (75) \\ &= (F - F_{\mathrm{BL}}) - \frac{1}{1 + F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}}} F_{\mathrm{BL}} (F_{\mathrm{cyt}}/F_{\mathrm{cyt}}^{\mathrm{eq}} - 1) && (76) \end{aligned}$$

so minimizing e_{vivo} amounts to a simple linear regression with unknown slope $1/(1 + F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}})$. For model fitting, we imposed the restriction $0.001 < F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}} < 20$.

Initial parameter values

We initialized $[\mathrm{GCaMP6s}]^{\mathrm{total}} = 20$ µM and $\psi = [0.45, 0.45, 22.5, 45]^T$. The $[\mathrm{Ca}^{2+}]_{\mathrm{rest}}$ was initialized to 50 nM [60, 116, 80, 142]. As previous studies suggest the extrusion time constant for calcium τ_{ex} is around 5 ms in dendrites [60, 115], we initialized this parameter to 20 ms to describe slower extrusion at the soma. The total calcium influx per AP $\Delta[\mathrm{Ca}^{2+}]_{\mathrm{AP}}$ was initialized to 5 µM to be consistent with past measurements of total calcium influx [107] as well as smaller measured increases in free calcium [60, 142]. We initialized the endogenous buffers K_d's to 500 nM and their time constants to 15 ms and 1 s. We initialized all total buffer concentrations at the same value, chosen so that the calcium binding ratio of a neuron without GCaMP6s at rest was 125 [60, 99, 80]. The calcium binding ratio of buffer ℓ was defined as the derivative of the concentration of calcium bound to the buffer with respect to free calcium concentration:

$$\kappa_\ell = \left. \frac{\partial [\mathrm{CaB}_\ell]_{\mathrm{EQ}}}{\partial [\mathrm{Ca}^{2+}]} \right|_{[\mathrm{Ca}^{2+}] = [\mathrm{Ca}^{2+}]_{\mathrm{rest}}} = \frac{b_\ell^- / b_\ell^+}{([\mathrm{Ca}^{2+}]_{\mathrm{rest}} + b_\ell^- / b_\ell^+)^2} \qquad (77)$$

To describe the affinity and kinetics of calcium binding to GCaMP6s, we used the same reparameterization as for *in vitro* data (see above), consisting of time constants τ_j and affinity parameters K_j^{50}. We initialized each τ_j to 100 ms. We used 8 different initializations for the K_j^{50} (Table 3), corresponding to the 8 possible ascending sequences from zero to four ions bound. These 8 initializations corresponded to the presence or absence of the 3 intermediate binding states at equilibrium as free calcium concentration was increased (absent binding states could still occur transiently in binding state kinetics evoked by AP discharge). When a a binding state $[\mathrm{Ca}_j\mathrm{GCaMP6s}]$ was absent from a the equilibrium binding state sequence of an initialization (Table 3, second column), we also imposed the restriction $K_{j+1}^{50} - K_j^{50} < 5\,\mathrm{nM}$ during optimization. After the optimization converged, we then released these restrictions and optimized further.

Optimization

We computed the gradients of the error function using finite differences. For each parameter, we set the step size for the finite gradient as 5e-7 or 5e-7 times the absolute value of the parameter value, whichever was larger. We used a positive step only (asymmetric finite differences). We optimized over F_{BG} for each neuron in closed form during finite difference calculation and line search.

Using this procedure for gradient calculation, we minimized e_{vivo} using the sequential quadratic programming technique available in MATLAB R2014a's **fmincon** function. We used the following settings: algorithm = 'sqp', tolfun = 1e-7, tolx = 1e-8, tolcon = 1e-10, gradobj = 'on'. We did not limit the number of iterations.

Optimization proceeded with the same parameter bounds on K_j^{50} and τ_j as for *in vitro* data, and the additional bounds 2.5 ms $\leq \tau_{\text{ex}} \leq$ 300 ms, 10 nM $\leq \Delta[\text{Ca}^{2+}]_{\text{AP}} \leq$ 100 µM, $10 \leq \phi_4 \leq 80$, $\phi_j \geq \phi_{j-1}$, 250 µs $\leq 1/b_\ell^- \leq$ 30 second, 100 nM $\leq b_\ell^-/b_\ell^+ \leq$ 100 µM and $[\text{B}]_\ell^{\text{total}} \geq 0$. We also enforced the bound 0.5 µM $\leq [\text{GCaMP6s}]^{\text{total}} \leq$ 300 µM for each neuron.

Hyperparameter fitting

After model-fitting, we obtained maximum likelihood estimators of the firing rate hyperparameters $k_\lambda, \theta_\lambda$ by applying MATLAB's **gamfit** function to the true firing rates. We also fit μ_G, Σ_G by maximum likelihood, yielding:

$$\mu_G = \mathbb{E}\left[\begin{array}{c}\log([\text{GCaMP6s}]^{\text{total}})\\ \log(F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}})\end{array}\right] \tag{78}$$

$$\Sigma_G = \text{Cov}\left(\left[\begin{array}{c}\log([\text{GCaMP6s}]^{\text{total}})\\ \log(F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}})\end{array}\right]\right) \tag{79}$$

where the expectation and covariance are over neurons. We refer to μ_G, Σ_G as hyperparameters since they define a distribution on the parameters F_{BG} and $[\text{GCaMP6s}]^{\text{total}}$, but for fixed values of F_{BG} and $[\text{GCaMP6s}]^{\text{total}}$ in a specific neuron they do not affect the likelihood of that neuron's fluorescence data.

Leave-one-out cross validation

In order to validate our procedures for AP inference and for estimation of F_{B} and $[\text{GCaMP6s}]^{\text{total}}$ from fluorescence measurements alone (see below), we first performed model fitting on all neurons but one in our dataset. We then discarded the neuron-specific parameters $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG} and transferred the remaining parameters and hyperparameters to the last neuron. This allowed us to infer the remaining neuron's AP sequence along with χ, $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG} from fluorescence data alone (see below), without using any parameters trained on its data.

When fitting the SBM using only a single neuron as training data, the algorithm was trained on neuron 1 and tested on neuron 2, trained on neuron 2 and tested on neuron 3, etc. If the neuron previous to the testing neuron was an interneuron (and hence could not be fit using our optimization procedures due to the lack of silent periods for fluorescence baseline estimation), the previous pyramidal neuron was used instead.

Sequential Monte Carlo-based AP inference

We aim to infer the presence or absence of AP discharge at each SBM simulation time step j, denoted $s_j \epsilon \{0, 1\}$. We first describe procedures for inferring the posterior probability that each $s_j = 1$ given fluorescence data when all parameters are known, followed by procedures for inferring neuron-specific data from fluorescence data alone.

We used Sequential Monte Carlo to approximate posterior distributions over a vector y of hidden states using finite weighted sums of delta functions. Each combination of state vector and weight is termed a particle. For the SBM y consists of s, ρ and all molecular species concentrations. We update the particles once for every fluorescence measurement, yielding one new value for ρ and Δ/δ new values for s and the binding state concentrations. Thus at the i^{th} fluorescence measurement the state vector includes spiking and binding state concentrations for the simulation time step range $U_i = (i-1)\Delta/\delta + 1 : (i-1)\Delta/\delta$. For $m = 4$ binding steps and $n_{\text{buffers}} = 2$, we have the state vector

$$y_i = \left[\begin{array}{c} s_{U_i}\\ [\text{Ca}^{2+}]U_i\\ [\text{G}]_{U_i}\\ [\text{Ca}_1\text{GCaMP6s}]_{U_i}\\ [\text{Ca}_2\text{GCaMP6s}]_{U_i}\\ [\text{Ca}_3\text{GCaMP6s}]_{U_i}\\ [\text{Ca}_4\text{GCaMP6s}]_{U_i}\\ [\text{CaB}_1]_{U_i}\\ [\text{CaB}_2]_{U_i}\\ \rho_i \end{array}\right] \tag{80}$$

Filtering distribution

Given the SBM model parameters, the filtering distribution $P(y_i|F_{1:i})$ describes the posterior probability distribution on the hidden states y at the time of the i^{th} fluorescent measurement F_i, given all fluorescence data up to and including F_i. We approximate the filtering with a finite number of samples, using the standard SMC estimator:

$$P(y_i|F_{1:i}) \approx \sum_{k=1}^{N_{\text{particles}}} W_{ki} \mathbb{1}_{y_i=\tilde{y}_{ki}} \tag{81}$$

where $\tilde{y}_{ki}$ is the state vector value for the k^{th} particle at the ith fluorescence measurement, $\mathbb{1}_{y_i=\tilde{y}_{ki}}$ is an indicator function and $\sum_k W_{ki} = 1$ for all i. In eq. 81, y_i are unknown hidden variables whose probability distribution we approximating, and the particle states $\tilde{y}_{ki}$ are known values we are using for this approximation. The particles are propagated from t_{i-1} to t_i by sampling a new y for each k from the proposal $q(y_i|y_{i-1})$. The weights are then updated according to the standard "sequential importance sampling" SMC update:

$$w_{ki} = W_{k,i-1} \frac{P(\tilde{y}_{ki}|\tilde{y}_{k,i-1})P(F_i|\tilde{y}_{ki})}{q(\tilde{y}_{ki}|\tilde{y}_{k,i-1})} \tag{82}$$

$$W_{ki} = \frac{w_{ki}}{\sum_k w_{ki}} \tag{83}$$

The data likelihood $P(F_i|\tilde{y}_{ki})$ is defined by eq. 38, while the transition probability $P(\tilde{y}_{ki}|\tilde{y}_{k,i-1})$ is defined by eq. 31 and 34:

$$P(F_i|\tilde{y}_{ki}) = \mathcal{N}\left(F_i; \hat{F}(\tilde{y}_{ki}), \frac{g\hat{F}(\tilde{y}_{ki}) + \sigma_F^2}{n_{A2D}}\right) \tag{84}$$

$$P(\tilde{y}_{ki}|\tilde{y}_{k,i-1}) = P(\rho_i|\rho_{i-1})P(S_{U_i}) = \mathcal{N}\left(\rho_{ki}; \rho_{k,i-1}, \delta\sigma_\rho^2\right) \prod_{j \epsilon U_i} (\delta\chi)^{s_{kj}} (1-\delta\chi)^{1-s_{kj}} \tag{85}$$

Our proposal q first samples spiking from a Bernoulli distribution $q_s(j)$ that is different for each j but does not depend on the state vector (see below). Once s_j has been sampled for each $j \epsilon U_i$, binding state concentrations are calculated deterministically by integrating the rate equation with our custom ODE solver. We sample ρ from the transition prior, a normal distribution with variance $\Delta\sigma_\rho^2$, so that the terms involving ρ cancel from the numerator and denominator of eq. 82. Together these updates produce an SMC estimate of $P(y_i|F_{1:i})$ given F_i and the SMC estimate of $P(y_{i-1}|F_{1:i-1})$.

Resampling

After the SMC update has been carried out many times, if the proposal distribution does not precisely match the posterior most of the particle weights will approach zero and a small number of particles will dominate the filtering distribution. This situation, known as particle degeneracy [35] can lead to increasing variance in the estimates of the filtering distribution for increasing t. To prevent this, a resampling operation is carried out in which particles are sampled with replacement from a discrete distribution $q_{\text{RS}}(k)$ on $\{1, \cdots, N_{\text{particles}}\}$ using the systematic resampling technique [73]. For each particle index $1 \leq k \leq N_{\text{particles}}$, a new index k' is sampled from q_{RS}. The weights and states are then updated according to:

$$y_{ki} \leftarrow y_{k'i} \tag{86}$$

$$w_{ki} \leftarrow w_{k'i}/q_{\text{RS}}(k') \tag{87}$$

$$W_{ki} \leftarrow \frac{w_{ki}}{\sum_k w_{ki}} \tag{88}$$

Most SMC algorithms choose q_{RS} to be equal to the current filtering distribution so that W_{ki} are all $1/N_{\text{particles}}$ after resampling, but other distributions have been used as well [41]. We used specialized distributions tailored to our model to reduce the variance of SMC algorithm results, as described below.

After each update of the particle weights, we calculated the effective sample size

$$N_{\text{eff}} = \frac{1}{\sum_k W_{ki}^2} \tag{89}$$

and resampled if $N_{\text{eff}} < N_{\text{particles}}/2$. When illustrating the particle filter technique in Figure 5a-c, we resampled at every fluorescence measurement.

Initial state distribution

To assign weights and states at the first time step we sample random AP sequences before the first fluorescence measurement and calculate their probabilities based on χ. We also sample initial values of ρ from a broad prior $\pi_\rho(\rho_1)$. We first defined F_{init} as the 10th percentile of the first 120 seconds of fluorescence values. If this gave a negative value, then σ_F was assigned to F_{init} instead. We next sampled ρ_1 from the prior

$$\pi_\rho(\rho_1) = \mathcal{N}\left(\rho_1; \log\left(\frac{F_{\text{init}}}{F_{\text{EQ}}}\right), V_\rho^0\right) \tag{90}$$

With $V_\rho^0 = \log(1.05)^2$.

For each particle k we set $[\text{Ca}^{2+}] = [\text{Ca}^{2+}]_{\text{rest}}$ and set the concentrations of all other molecular species to their equilibrium values using eq. 18. We next a random AP sequence s_{k,U_1}, with U_1 in this case indexing 2.5 seconds before the first fluorescence measurement, and update the concentrations of indicator and buffer binding states accordingly by integrating the rate equation. As for the standard SMC state update, the probabilities of randomly sampling an AP for each simulation time step are taken from q_s, the form of which is specified below. We then calculated for each particle k the maximum likelihood estimate of ρ given the indicator binding state concentrations and a fluorescence observation of F_{init}:

$$\hat{\rho}_{k,1} = \underset{\rho_1}{\operatorname{argmax}}\, P\left(F_{\text{init}} \middle| \rho_1, \begin{array}{c}[\text{Ca}_j\text{GCaMP6s}]_k\\ 0 \le j \le m\end{array}\right) \tag{91}$$

$$= \log(F_{\text{init}}) - \log\left(1 + \sum_{j=1}^{m} \psi_j \frac{[\text{Ca}_j\text{GCaMP6s}]_k}{[\text{GCaMP6s}]^{\text{total}}} + \frac{F_{\text{BG}}}{\phi_0[\text{GCaMP6s}]^{\text{total}}}\right) \tag{92}$$

and calculated the Laplace approximation [19] to the likelihood of F_{init} given ρ, centered at $\hat{\rho}_{k,1}$:

$$P^k_{\text{Laplace}}(F_{\text{init}}|\rho_1) = \mathcal{N}\left(\rho_1; \hat{\rho}_{k,1}, V^k_{\text{Laplace}}\right) \tag{93}$$

$$V^k_{\text{Laplace}} = \frac{gF_{\text{init}} + \sigma_F^2}{F_{\text{init}}^2 n_{\text{A2D}}} \tag{94}$$

Using this approximation and the prior π_ρ, we calculated the sampling distribution for each k by multiplying the prior and the Laplace approximation:

$$q_\rho\left(\rho_1 \middle| F_{\text{init}}, \begin{array}{c}[\text{Ca}_j\text{GCaMP6s}]_k\\ 0 \le j \le m\end{array}\right) = \mathcal{N}\left(\rho_1; V_\rho^*\left(\frac{\hat{\rho}_{k,1}}{V^k_{\text{Laplace}}} + \frac{\log(F_{\text{init}}/F_{\text{EQ}})}{V_\rho^0}\right), V_\rho^*\right) \tag{95}$$

$$V_\rho^* = \left((V^k_{\text{Laplace}})^{-1} + (V_\rho^0)^{-1}\right)^{-1} \tag{96}$$

We then sampled each $\rho_{k,1}$ and calculated weights as the ratio of prior and sampling probabilities:

$$W_{k,1} = \frac{\pi_\rho(\rho_{k,1}) \prod_{j \epsilon U} (\delta\chi)^{s_{kj}} (1-\delta\chi)^{1-s_{kj}}}{q_\rho\left(\rho_1 \middle| F_{\text{init}}, \begin{array}{c}[\text{Ca}_j\text{GCaMP6s}]_k\\ 0 \le j \le m\end{array}\right) \prod_{j \epsilon U_1} q_s(j)^{s_{kj}} (1-q_s(j))^{1-s_{kj}}} \tag{97}$$

Smoothing distribution

To estimate the smoothing distribution $P(y_i|F_{1:T})$, we used a standard filter-smoother [73] with lag of 500 ms. We kept track of each particle's history of previous state values as we advanced the filtering distribution forward in time. Then, at each time point the current filtering weights together with the state values of each particle's ancestor 500 ms back in time define the smoothing distribution. For particle k at fluorescence measurement i, the smoothing weight W^*_{ki} is defined as the summed weights of its descendants 500 ms after t_i (or for $t_i > t_T - 0.5$, descendant weights at t_T). $W^*_{ki} = 0$ for particles with no descendants 500 ms in the future. Smoothing weights are then used to calculate posterior means and variances of y:

$$\mathbb{E}_{\text{smc}}(y_i|F_{1:T})) = \sum_k W^*_{ki} \tilde{y}_{ki} \tag{98}$$

$$\text{Var}_{\text{smc}}(y_i|F_{1:T})) = \sum_k W^*_{ki} \left(\tilde{y}_{ki} - \mathbb{E}_{\text{smc}}(y_i|F_{1:T}))\right)^2 \tag{99}$$

where the squaring in the definition of Var_{smc} is carried out elementwise.

Sampling and resampling techniques for SMC-based AP inference with the SBM

The quality of an SMC algorithm, measured by the number of particles required to obtain accurate filtering and smoothing distributions, depends strongly on the choice of proposal distribution q [34, 19, 35]. When the generated samples match the posterior poorly, many will be rejected during resampling events and most computation will be wasted. For the SBM, the key question is whether the algorithm extends AP sequences forward in time in a way consistent with fluorescence data. Successful inference requires at least some particles to generate an AP at the time of each true AP, and for these particles to survive all resampling steps until the filter-smoother lag has been reached. However, if the needed APs are not sampled or are lost to resampling, the algorithm will fail to detect the true AP. In this section, we describe three techniques for designing proposal and resampling distributions to generate the needed spikes and ensure their survival through resampling steps.

Increased AP discharge probability in the sampling distribution For $\delta = 10$ ms and our median firing rate of $\chi = 0.16$ Hz, the prior probability $P(s_j = 1)$ is only $p_{\mathrm{AP}} = \chi\delta = 0.0016$. Therefore, even with many particles the chance of simply failing to generate spike trains consistent with the data is not insignificant when sampling s_j from its prior distribution. We therefore sample APs with a higher probability than the prior to generate a greater diversity of AP sequences. While this mismatch between the transition prior and proposal will cause an overall increase in the frequency of resampling when APs do not occur, the fit to the data will be better when APs are present.

We sampled APs with probability $\max(p_{\mathrm{AP}}, 0.01)$. We demonstrate how sampling APs at a higher rate than the prior can reduce variability in the SMC algorithm's output in Figure 5 - figure supplement 2.

Multiround filtering with time-varying proposals To further improve the quality of sampled AP sequences, we first run the SMC algorithm with increased probability of AP sampling as above to obtain the smoothing distribution and $\bar{s}_j = \mathbb{E}_{\mathrm{smc}}[s_j|F_{1:T}]$. We then designed a time-varying proposal on AP discharge

$$q(s) = \max(\bar{s} \circ K_q, 0.01) \tag{100}$$

where $\circ$ denotes convolution and K_q is a Gaussian kernel with $\sigma = 100$ ms. We extend the calculation of both $\bar{s}$ and the time-varying proposal $q(s)$ to the full 2.5 seconds before the first fluorescence measurement for which SBM simulations are carried out (see section "Initial state distribution" above). We show how this technique further reduces Monte Carlo variance in Figure 5 - figure supplement 2.

Resampling to the proposal For GECIs in general and for GCaMP6s in particular, an additional difficulty for SMC algorithms arises from the delay between AP discharge and peak fluorescence. The first fluorescence measurement after AP discharge may not show a strong increase in fluorescence, and for higher imaging rates this can increase to 2 or 3 measurements. Consequently, for particles generating an AP at the correct time, the resulting increase in their weights relative to other particles without the correct AP may not appear until the algorithm has passed over at least one second of fluorescence measurements. So even even if APs are sampled by the SMC algorithm at the correct time, they may be lost during resampling steps before the predicted fluorescence increase has been fully observed. Since the p_{spike} is low, sampling an AP will decrease the probability weight of a particle at the next fluorescence measurement compared to other particles (eq. 82). This will both reduce N_{eff} to make resampling more likely and increase the chance that the particle with the AP will be lost.

This problem is not alleviated by using an increased or time-varying $q(s)$, since while this causes more APs to be sampled when needed it also decreases the weights of particles with APs even further due to the presence of q in the denominator of eq. 82. We might sometimes be fortunate enough not to encounter a resampling event between the sampling of an AP and the subsequent correct prediction of a fluorescence increase, but this will not always be the case. The possibility of these two different outcomes, where the spikes needed to match the data are retained or are lost to resampling, sharply increases the variability of the SMC algorithm over multiple runs.

We therefore introduce a novel modification of the SMC algorithm which we term "resampling to the proposal," designed specifically to deal with state transitions that have strong but delayed effects on the predict measurements. In addition to the standard raw weights w_{ki} and normalized weights W_{ki}, we also maintain a second set of raw weights w^q_{ki} and normalized weights W^q_{ki}. The second set of weights, which we refer to as "proposal weights" are obtained by calculating the particle weight updates as if the time-varying proposal on APs were also the prior on APs. Both sets of weights share the same state variables $\tilde{y}$. This leads to the update rules:

$$w^q_{ki} = W^q_{k,i-1} P(F_i|\tilde{y}_{ki}) \tag{101}$$

$$w_{ki} = w^q_{ki} \prod_{j \epsilon U_i} \frac{p_{\mathrm{spike}}}{q(s_j)} \tag{102}$$

with normalized weights calculated as above. Note that s does not appear in the update of w^q_{ik} since the proposal and prior on s are the same, so the numerator and denominator in eq. 82 cancel. Thus particles with more APs do not have lower resampling weights w^q_{ik}, although they do have lower standard weights w_{ik}.

To reduce the chance of losing needed APs to resampling events, we simply set the resampling distribution to be equal to the normalized proposal weights: $q_{RS}(k) = W^q_{ik}$. Resampling is used to select a single set of indices, but the standard and resampling weights are updated differently. The updates now take the form

$$w_{ki} \leftarrow w_{k'i}/w^q_{k'i} \tag{103}$$

$$w^q_{ki} \leftarrow 1 \tag{104}$$

with normalized weights calculated as above. We calculate the effective sample size using the proposal weights W^q_{ik}.

Finally, when calculating the smoothing distribution we use the original weights W_{ki} to calculate W^*_{ki} as before. Unlike the proposal weights, the standard weights used for smoothing take into account χ through a weight decrease for every discharged AP, but they also reflect fluorescence measurements acquired after those APs. This technique allowed us to sample APs with high probability from our time-varying prior, keep those spikes through resampling events and still obtain the smoothing distribution for the original problem. We show how this further reduces Monte Carlo variance of the SMC algorithm in figure 5 - figure supplement 2. In order to illustrate the SMC technique in Figure 5c, the area of the black dots was chosen to be proportional to proposal weights, using a q based on increased AP discharge probability but not multiround filtering.

To our knowledge, resampling to the proposal is the first approach to use two particle filters with the same states but different weights to deal with lags between state changes and changes in the observed data. Previous approaches to dealing with such lags have generally focused on modifying SMC states in the past using MCMC steps [46, 33], but these approaches are often unreliable for time series more than a few hundred measurements [3] and could impose impractical computational and memory requirements for GPU-based SMC with many particles. However, the idea of coupling two particle filters to use the same indices during resampling is not unprecedented, and has been applied to compute finite difference gradients of the data likelihood with respect to static model parameters [22]. However, the use of proposal weights for resampling distinguishes resampling to the proposal from previous approaches, and when combined with proposal distributions based on multi-round filtering may prove useful for a wide range of nonlinear filtering problems.

Fitting a spike train to posterior moments

The SMC algorithm described above calculates each $\bar{s}_j$. These values are discretized in time with a step size of δ, and represent the marginals of the joint posterior distribution provided by the SMC smoother. While for certain applications, such as calculation of receptive fields, $\bar{s}$ itself can be used directly, at other times it is preferable to infer a single AP sequence from a fluorescence time series.

We therefore developed a procedure to infer a single, non-time-discretized AP sequence from $\bar{s}$. Inspection of the time-varying posterior spiking probability from the SMC smoother revealed that true, electrically detected APs gave rise to Gaussian-like peaks in $\bar{s}$, which can be interpreted as uncertainty in the times of APs. We therefore fit to $\bar{s}$ an approximation consisting of a sum of n_{APs} Gaussian functions with means μ_k and standard deviations σ_k. We modeled the individual values of $\bar{s}_j$ as integrals of the Gaussian functions over each time step of length δ:

$$\bar{s}_j \approx \sum_{k=1}^{n_{\text{APs}}} \Phi\left(\frac{t_j + \delta/2 - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{t_j - \delta/2 - \mu_k}{\sigma_k}\right) \tag{105}$$

which we optimize in least squares with the restriction that $\delta/2 \leq \sigma_k \leq 50$ ms.

We determine n_{APs} and the values of μ_k and σ_k using a greedy algorithm. First, we attempt to add a new spike while varying the new σ value in 100 steps, ranging from 1% to 99% of the interval between the minimum and maximum allowed values. For each value of σ for the new AP, we shift the mean to each t_j, and choose the location which decreases the sum of square errors most. We then adjust all the μ's and σ's to minimize the sum of square errors using Gauss-Newton optimization, while allowing both to vary continuously without discretization. The algorithm terminates when no new spikes can be added to reduce the sum of square errors, yielding the μ's as spike times and the σ's as temporal uncertainties. For the step of adding new spikes, we precompute a lookup table of all necessary Φ values for speed. This method is illustrated in Figure 5 - figure supplement 6.

Parameter estimation from fluorescence data alone

Most of the SBM's parameters are determined by fitting *in vitro* binding assay data or fitting *in vivo* fluorescence data with true AP times known. However, when detecting APs from fluorescence recordings without simultaneous cell-attached recordings or where the true APs have been held out for testing, the remaining parameters $(\chi, \sigma_F, [\text{GCaMP6s}]^{\text{total}}, F_{\text{BG}})$ must be determined from fluorescence alone. For this purpose, we use a heuristic method to determine σ_F and an expectation-maximization method for χ, while $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG} are determined using an SMC-based estimate of the marginal data likelihood.

Estimation of σ_F

We estimated σ_F as

$$\hat{\sigma}_F = n_{\text{A2D}} \frac{\text{median}_i(|F_i - F_{i-1}|)}{2\text{erf}^{-1}(0.5)} \tag{106}$$

where erf^{-1} is the inverse error function. The numerator is the median absolute fluorescence difference across neighboring image frames, while the denominator is median absolute difference for two values drawn from a standard normal distribution.

EM estimation of χ

With σ_F known and for fixed values of $[\text{GCaMP6s}]^{\text{total}}$ and F_{BG}, we perform maximimum a posteriori (MAP) estimation of χ using the SMC version of the Expectation-Maximization (EM) algorithm [28]. In the E-step, use the SMC algorithm to estimate the smoothed posterior on s given F. In the M-step, we use this posterior to maximize the expected complete log-likelihood times the parameter prior:

$$Q(\Theta|\Theta_0) = P(\Theta) \int P(y_{1:T}|F_{1:T}, \Theta_0) \left[\log P(F_{1:T}|y_{1:T}, \Theta) + \log P(y_{1:T}|\Theta)\right] dy_{1:T} \tag{107}$$

where Θ_0 is the current estimate of the parameters and Θ is the new estimate. The SMC version of the EM algorithm uses the SMC smoothing distribution, which is a sum of delta functions, to approximate $P(y_{1:T}|F_{1:T}, \Theta_0)$. Thus maximization of eq. 107 over χ reduces to maximization of a gamma prior times a product of Bernoulli likelihoods, which we consider in the log domain:

$$\begin{aligned} \log(P(\chi)) + \sum_{i,k} W^*_{ki} \sum_{j \epsilon U_i} s_{jk} \log(\chi\delta) + (1 - s_{jk}) \log(1 - \chi\delta) = \\ (k_\chi - 1)\log(\chi) - \chi/\theta_\chi + \left(\sum_j \bar{s}_j\right) \log(\chi\delta) + \left(T\Delta/\delta - \sum_j \bar{s}_j\right) \log(1 - \chi\delta) \end{aligned} \tag{108}$$

Having computed $\sum_j \bar{s}_j$ using the SMC smoother, we maximized the expression in eq. 108 numerically with the Matlab function **fminbnd**.

We found that a single EM-iteration was sufficient to estimate χ, in the sense that multiple iterations did not improve inference of the AP sequence or the marginal data likelihood (see below). Since we already use two SMC passes over the data to reduce Monte Carlo variance (see section "Multiround filtering with time-varying proposals" above), we simply estimated χ after the first round of SMC and used this value for the second round.

Marginal likelihood

In addition to providing filtering and smoothing distributions, SMC algorithms can also be used to estimate the marginal data likelihood $P(F_{1:T}|\Theta)$, where Θ is a vector of model parameters that determine the transition prior, the initial state distribution and/or the distribution on observed data values given the model state. We seek to compute this function in order to perform model fitting: that is, to find values of Θ consistent with our data. The marginal likelihood can be a difficult quantity to calculate since it involves integration over the joint distribution on hidden states over all time points:

$$P(F_{1:T}|\Theta) = \int P(F_{1:T}|y_{1:T}, \Theta) P(y_{1:T}|\Theta) dy_{1:T} \tag{109}$$

We follow the standard approach of decomposing

$$P(F_{1:T}|\Theta) = P(F_1) \prod_{i=2}^{T} P(F_i|F_{1:i-1}, \Theta) \tag{110}$$

The terms of this product, which are the likelihoods of each fluorescence measurement condition on previous measurements, can be calculated using the SMC weights [104]:

$$P_{\mathrm{SMC}}(F_1|\Theta) = \frac{\sum_k w_{k,1}}{N_{\mathrm{particles}}} \tag{111}$$

$$P_{\mathrm{SMC}}(F_i|F_{1:i-1},\Theta) = \sum_k w_{ki} \tag{112}$$

Maximization of tempered posterior

For a given choice of $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG} we can obtain an estimate $\hat{\chi}$ of χ via EM, and then calculate the SMC estimate of the marginal likelihood. Multiplying the result by the prior on χ, $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG} yields

$$\mathcal{H}([\mathrm{GCaMP6s}]^{\mathrm{total}}, F_{\mathrm{BG}}) \tag{113}$$

$$= P(F_{1:T}|[\mathrm{GCaMP6s}]^{\mathrm{total}}, F_{\mathrm{BG}}, \hat{\chi}, \hat{\sigma}_F, \Theta_{\mathrm{fixed}})\mathcal{N}\left(\left[\log([\mathrm{GCaMP6s}]^{\mathrm{total}}), \log(F_{\mathrm{BG}}/F^{\mathrm{eq}}_{\mathrm{cyt}})\right]^T; \mu_G, \Sigma_G\right) P(\chi) \tag{114}$$

$$\propto P([\mathrm{GCaMP6s}]^{\mathrm{total}}, F_{\mathrm{BG}}, \hat{\chi}|F_{1:T}, \hat{\sigma}_F, \Theta_{\mathrm{fixed}}) \tag{115}$$

Where Θ_{fixed} contains SBM parameters that do not vary over neurons. Since $\hat{\chi}$ is a partial optimizer of the posterior given the other parameters, maximizing $\mathcal{H}$ over $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG} yields a maximum a posteriori estimate over all three parameters.

Direct maximization of $\mathcal{H}$ is not possible, since even for 10^5 particles the variance of $\log \mathcal{H}$ for fixed input can reach 10 for as little as 30 seconds of imaging data (Figure 5 - figure supplement 2). Instead, we seek to characterize the shape of the function $\mathcal{H}$ and find its maximum in a way that is robust to the variability in the calculated value of $\mathcal{H}([\mathrm{GCaMP6s}]^{\mathrm{total}}, F_{\mathrm{BG}})$ over multiple runs arising from random number generation within the SMC algorithm.

To do this, we use the adaptive independent Metropolis Hastings (AIMH) technique of [48]. This method draws samples from a density proportional to a function while simultaneously fitting a Gaussian mixture to the sample set by harmonic K-means clustering. The algorithm was implemented in Matlab as described in [48], and run for 400 iterations. For harmonic K-means clustering, we used the **kmeanhar** implementation of [140]. Our implementation accepts multiple time series for the same neuron, but by default we limit the length of the input data to 12500 fluorescence measurements for speed. Because we are interested only in the maximum of $\mathcal{H}$ and not in characterizing the entire distribution, we increased the acceptance rate by "tempering" $\mathcal{H}$ by a temperature $\mathcal{T}$. That is, instead of using $\mathcal{H}$ as the target distribution for the AIMH technique, we used $\mathcal{H}^{1/\mathcal{T}}$. To choose the temperature, we first calculated $\mathcal{H}$ 15 times at the initial values of $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ and F_{BG}, which we set to equal μ_G. We then choose $\mathcal{T}$ so that $\log(\mathcal{H}^{1/\mathcal{T}})$ would have a variance of 0.1 or less:

$$\mathcal{T} = \max\left(\sqrt{10 \cdot \mathrm{Var}(\mathcal{H}(\mu_G))}, 10\right) \tag{116}$$

After 400 AIMH iterations, we returned the mode of the Gaussian mixture as a point estimate for $([\mathrm{GCaMP6s}]^{\mathrm{total}}, F_{\mathrm{BG}})$, along with the corresponding values of $\hat{\chi}, \hat{\sigma}_F$. These procedures are illustrated in Figure 5 - figure supplement 5.

Other inference methods

We implemented Matlab wrappers for all other AP inference methods, a Matlab GUI capable of running all algorithms on individual data segments, neurons and datasets and calculating accuracy measures based on the results. All individual and group data comparisons of various algorithms were carried out on the same data with the same pre-processing, including feature extraction and motion correction. For inference methods

FOOPSI

Fluorescence data were first detrended using Matlab's built-in **detrend** function. Next, the minimum value was subtracted from the detrended data, the result was divided by its maximum and a value of machine epsilon (2.22e-16) was added. Finally, the result was passed into the **fast_oopsi** function, with the parameter tau_c = 0.5.

CFOOPSI

Unmodified fluorescence values were passed to the **constrained_foopsi** function, with no additional inputs. Commit 08468bc1b25c9b8617b861a6404bbf4576cf156c was retrieved from
github.com/epnev/constrained-foopsi, and was used with CVX version 2.1, Build 1116 (d4cc5c5).

c2s

For c2s-s, we used the **c2s preprocess** command to upsample fluorescence data to 100 Hz as descriped in [Theis et al.], while providing metadata indicating which fluorescence time series were from the same neuron. We then used the **c2s predict** command to infer spike probabilities. For c2s-t, we used the **c2s leave-one-out** command to carry out training and to infer spike probabilities with cross validation. To calculate the number of parameters used by c2s, we trained it on our complete dataset (n = 26 neurons) with standard settings, for which c2s selected 10 PCA components consisting of 1000 elements but only 945 degrees of freedom after accounting for orthonormality constraints. The spike triggered mixture model used 20 feature parameters, 3 bias parameters, 6 weights and 30 predictor coefficients, for a total of 1004 parameters. Commit 9f12398a33a17e557b4a689f1dce902123c6f2eb was retrieved from github.com/lucastheis/c2s.

MLspike

We used MLspike with the published parameters for GCaMP6s [29], as the autocalibration feature failed whenever tested on our GCaMP6s data. These parameters were: a = 0.113, tau = 1.87, pnonlin = [0.81, -0.056], drift.parameter = 0.01. One call was made to the **tps_mlspikes** function for each different imaging rate in a neuron's data, with the parameters and the frame interval passed along with fluorescence data. For an image frame i for which a single AP was inferred, we assigned it to the midway point between the times when the neuron was scanned: $(t_{i-1} + t_i)/2$. For an image frame for which k APs were inferred, we distributed the inferred APs evenly across the interval between the two scan times, assigning the j^{th} AP time as $t_{i-1} + \Delta(j - 0.5)/k$. Commit 048122135c7d77457ee8c8c026a572ac40739c3f was retrieved from github.com/MLspike/spikes, along with commit 048122135c7d77457ee8c8c026a572ac40739c3f from github.com/MLspike/brick.

thr-σ

We first estimated fluorescence baseline F_{BL} as the 8th percentile of fluorescence values over a 15 s window around each data point. For data points within 7.5 s of the start or end of a fluorescence recording, the first or last fluorescence value was repeated to substitute for the missing values. We then calculated a normalized $\Delta F/F_0$ measure as $y = (F - F_{\text{BL}})/\text{median}(F_{\text{BL}})$. We then divided the fluorescence recording into 1 s windows and calculated the standard deviation of y in each window, and estimated the standard deviation of fluorescence noise σ_{noise} as the median over 1 s windows of the standard deviation of y. We then identified time points where y cross a threshold of $4\sigma_{\text{noise}}$. After each threshold crossing, additional threshold crossing events were ignored until after y decreased below $2\sigma_{\text{noise}}$. For each threshold crossing, an AP was assigned at the maximum value of y among the values after y increased over $4\sigma_{\text{noise}}$ and before it decreased below $2\sigma_{\text{noise}}$.

Accuracy measures

To calculate the correlation of two AP sequences with AP times $\{a_i\}$ and $\{b_j\}$, we convolved each AP sequence with a Gaussian kernel of width σ leading to a sum of Gaussian functions:

$$Z_a(t) = \sum_i \mathcal{N}\left(t; a_i, \sigma^2\right) \tag{117}$$

$$Z_b(t) = \sum_j \mathcal{N}\left(t; b_j, \sigma^2\right) \tag{118}$$

We then calculated the means of Z_a, Z_b, their product and their squares over the time window $[t_0, t_1]$ over which the AP sequences are defined:

$$\overline{Z_a} = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} Z_a dt = \frac{1}{t_1 - t_0} \sum_i \left[\Phi\left(\frac{t_1 - a_i}{\sigma}\right) - \Phi\left(\frac{t_0 - a_i}{\sigma}\right)\right] \tag{119}$$

$$\overline{Z_a^2} = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} Z_a^2 dt = \frac{1}{t_1 - t_0} \sum_{i_1, i_2} G(a_{i_1}, a_{i_2}) \tag{120}$$

$$\overline{Z_a Z_b} = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} Z_a Z_b dt = \frac{1}{t_1 - t_0} \sum_{i,j} G(a_i, b_j) \tag{121}$$

$$G(u, v) = \mathcal{N}\left(u; v, 2\sigma^2\right) \left[\Phi\left(\frac{t_1 - (u+v)/2}{\sigma/\sqrt{2}}\right) - \Phi\left(\frac{t_0 - (u+v)/2}{\sigma/\sqrt{2}}\right)\right] \tag{122}$$

where Φ is the Gaussian cdf and $\overline{Z_b}$, $\overline{Z_b^2}$ are defined correspondingly. Finally we calculated the correlation as

$$\rho = \frac{\overline{Z_a Z_b} - \overline{Z_a} \cdot \overline{Z_b}}{\sqrt{\overline{Z_a^2} \cdot \overline{Z_b^2}}} \tag{123}$$

To calculate the average absolute timing error, we first collected all the outputs of each algorithm up to 500 ms before and after each isolated single AP. For algorithm that produced AP times (the SBM, MLspike and thr-σ), we then simply calculated the mean absolute time difference from the true AP. For other algorithms which produce expected AP counts (c2s) or unitless outputs (FOOPSI and CFOOPSI), we calculated $\left(\sum_i |t_i - t_{\mathrm{AP}}| y_i\right) / \left(\sum_i y_i\right)$, where y_i is the algorithm's output at time t_i.

Statistical analysis

To calculated the average fluorescence evoked by single APs in pyramidal neurons, we included only those without other APs in the preceding 5.5 s. We first calculated fluorescence baseline F_{BL} and $\Delta F/F_0$ values using fluorescence and true AP times as described above in the section "Estimation of baseline fluorescence" above. The baseline-fitting procedure also determined the time constant τ_{BL} with which fluorescence decays back to baseline after starting 5.5 s after AP discharge, as well as an exponential fit to all $\Delta F/F_0$ values at least 5.5 s after any APs. Therefore, to remove any effects of earlier AP discharge >5.5 s before single APs, we extrapolated the exponential fit forward through the fluorescence values around the single AP, and subtracted this from the $\Delta F/F_0$ values. We then subtracted the mean of $\Delta F/F_0$ values for image frames 0 to 200 ms before the AP. Finally, we calculated the $\Delta F/F_0$ values around single APs, using 75 ms bins with a bin edge centered on the time of AP discharge. For interneurons, the requirement that no other APs were present in the preceding 300 ms was used instead. To calculate fluorescence responses to bursts of multiple APs, we included bursts with up to 200 ms between the first and final AP, and no other APs afterwards, with a bin edge centered on the first AP time.

The amplitude or "height" of the response was defined as the maximum of the average from 0 to 300 ms for single APs, and 0 to 500 ms for bursts. Signal-to-noise ratio (SNR) was defined as the ratio of the single AP response amplitude to the standard deviation of the fluorescence noise. To calculate the noise standard deviation, we used all $\Delta F/F_0$ values 300 ms to 50 ms before isolated single APs as defined above.

All values are reported as mean +/- standard deviation unless otherwise stated.

Data and code availability

All AP sequences, fluorescence signals and *in vitro* binding assay data are provided along with metadata (supplementary data). Raw fluorescence movies will be available in the future at caesar.de/sbm. Pre-alpha releases of code for analyzing *in vivo* and *in vitro* data are available at `http://github.com/dgreenberg/sbmvivo` and `http://github.com/dgreenberg/sbmvitro`.

Acknowledgments

We thank Stephan Irsen and Daniel Rudolph for assistance with cluster computing, Alexandr Klioutchnikov for assistance with *in vivo* optical imaging, Jeanine Klotz for technical assistance and Norbert Brenner for initial assistance with protein purification. We thank Jakob Macke for fruitful discussions and comments on the manuscript.

Author contributions

In vivo GCaMP6s imaging experiments with simultaneous electrical recordings were designed by JNDK, DJW and DSG, and performed by DJW and UC. Imaging experiments with visual stimulation were performed by TH. *In vitro* binding assay experiments were designed by SW, YG, RS and DG and performed by SW. The SBM, model-fitting procedures and AP inference method were developed by DSG. *In vivo* data were analyzed by DSG and KMV. *in vitro* data were analyzed by DSG, AM, SW and YG. KMV and DSG developed the feature extraction technique. JTV helped design an earlier SMC method for small-molecule indicators. DSG and JNDK ran the project and wrote the paper.

References

[1] J. Akerboom et al. “Optimization of a GCaMP calcium indicator for neural activity imaging”. In: *J Neurosci* 32.40 (2012), pp. 13819–40.

[2] Jasper Akerboom et al. “Crystal Structures of the GCaMP Calcium Sensor Reveal the Mechanism of Fluorescence Signal Change and Aid Rational Design”. In: *Journal of Biological Chemistry* 284.10 (2009), pp. 6455–6464. DOI: 10.1074/jbc.M807657200.

[3] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. “Particle markov chain monte carlo methods”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3 (2010), pp. 269–342.

[4] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. “Identification of causal effects using instrumental variables”. In: *Journal of the American statistical Association* 91.434 (1996), pp. 444–455. ISSN: 0162-1459.

[5] Toshihiko Aosaki and Haruo Kasai. “Characterization of two kinds of high-voltage-activated Ca-channel currents in chick sensory neurons”. In: *Pflügers Archiv* 414.2 (1989), pp. 150–156.

[6] Uri M Ascher and Linda R Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*. Vol. 61. Siam, 1998.

[7] Lauren M. Barnett, Thomas E. Hughes, and Mikhail Drobizhev. “Deciphering the molecular mechanism responsible for GCaMP6m’s Ca2+-dependent change in fluorescence”. In: *PLoS One* 12.2 (2017), e0170934. DOI: 10.1371/journal.pone.0170934.

[8] R. Bellman and K. J. Astrom. “On structural identifiability”. In: *Mathematical Biosciences* 7.3 (1970), pp. 329–339. ISSN: 0025-5564. URL: http://www.sciencedirect.com/science/article/pii/002555647090132X.

[9] Donald M Bers, Chris W Patton, and Richard Nuccitelli. “A practical guide to the preparation of Ca2+ buffers”. In: *Methods in cell biology*. Vol. 99. Elsevier, 2010, pp. 1–26.

[10] Upinder S. Bhalla. “Molecular computation in neurons: a modeling perspective”. In: *Current Opinion in Neurobiology* 25 (2014), pp. 31–37.

[11] Guo-qiang Bi and Mu-ming Poo. “Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type”. In: *The Journal of Neuroscience* 18.24 (1998), p. 10464.

[12] T. J. Blanche et al. “Polytrodes: high-density silicon electrode arrays for large-scale multiunit recording”. In: *J Neurophysiol* 93.5 (2005), pp. 2987–3000.

[13] Tiago Branco, Beverley A. Clark, and Michael Häusser. “Dendritic Discrimination of Temporal Input Sequences in Cortical Neurons”. In: *Science* 329.5999 (2010), p. 1671.

[14] Johann Brehmer et al. “A Guide to Constraining Effective Field Theories with Machine Learning”. In: *arXiv preprint arXiv:1805.00020* (2018).

[15] Marisa Brini et al. “Neuronal calcium signaling: function and dysfunction”. In: *Cellular and molecular life sciences* 71.15 (2014), pp. 2787–2814.

[16] John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley and Sons, 2016.

[17] ELHD Carbone and HD Lux. “A low voltage-activated, fully inactivating Ca channel in vertebrate sensory neurones”. In: *Nature* 310.5977 (1984), pp. 501–502.

[18] T. W. Chen et al. “Ultrasensitive fluorescent proteins for imaging neuronal activity”. In: *Nature* 499.7458 (2013), pp. 295–300.

[19] Zhe Chen. “Bayesian filtering: From Kalman filters to particle filters, and beyond”. In: *Statistics* 182.1 (2003), pp. 1–69.

[20] Adrian Cheng et al. “Simultaneous 2-photon calcium imaging at different cortical depths in vivo with spatiotemporal multiplexing”. In: *Nature methods* 8.2 (2011), pp. 139–142.

[21] Joseph Cichon and Wen-Biao Gan. “Branch-specific dendritic Ca2+ spikes cause persistent synaptic plasticity”. In: *Nature* 520.7546 (2015), pp. 180–185.

[22] A Coquelin, Romain Deguest, and Remi Munos. “Perturbation analysis for parameter estimation in continuous space HMMs”. In: *Submitted to IEEE Transactions on Signal Processing* (2008).

[23] Hod Dana et al. “Sensitive red protein calcium indicators for imaging neural activity”. In: *eLife* 5 (2016), e12727.

[24] Hod Dana et al. “Thy1-GCaMP6 Transgenic Mice for Neuronal Population Imaging In Vivo”. In: *PLoS One* 9.9 (2014), e108697.

[25] Biswa Nath Datta. *Numerical linear algebra and applications*. Vol. 116. Siam, 2010.

[26] R. Christopher deCharms and Michael M. Merzenich. “Primary cortical representation of sounds by the coordination of action-potential timing”. In: *Nature* 381.6583 (1996), pp. 610–613.

[27] Pierre Del Moral. “Feynman-Kac Formulae”. In: *Feynman-Kac Formulae*. Springer, 2004, pp. 47–93.

[28] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), pp. 1–38.

[29] Thomas Deneux et al. “Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo”. In: *Nature communications* 7 (2016), p. 12190.

[30] W. Denk, J. H. Strickler, and W. W. Webb. “Two-photon laser scanning fluorescence microscopy”. In: *Science* 248.4951 (1990), pp. 73–6.

[31] Ferran Diego and Fred A. Hamprecht. *Sparse space-time deconvolution for Calcium image analysis*. 2014.

[32] Daniel A. Dombeck et al. “Imaging large scale neural activity with cellular resolution in awake mobile mice”. In: *Neuron* 56.1 (2007), pp. 43–57.

[33] Arnaud Doucet, Mark Briers, and Stéphane Sénécal. “Efficient block sampling strategies for sequential Monte Carlo methods”. In: *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 693–711.

[34] Arnaud Doucet, Nando De Freitas, and Neil Gordon. *Sequential Monte Carlo methods in practice*. 2001.

[35] Arnaud Doucet and Adam M Johansen. “A tutorial on particle filtering and smoothing: Fifteen years later”. In: *Handbook of nonlinear filtering* 12.656-704 (2009), p. 3.

[36] Timothy W. Dunn et al. “Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion”. In: *eLife* 5 (2016), e12741.

[37] Gaute T Einevoll et al. “Modelling and analysis of local field potentials for studying the function of cortical circuits”. In: *Nature Reviews Neuroscience* 14.11 (2013), pp. 770–785.

[38] Mohamed El Harrous, Stanley J Gill, and A Parody-Morreale. “Description of a new Gill titration calorimeter for the study of biochemical reactions. I: assembly and basic response of the instrument”. In: *Measurement Science and Technology* 5.9 (1994), p. 1065.

[39] Guido C. Faas et al. “Calmodulin as a direct detector of Ca2+ signals”. In: *Nat Neurosci* 14.3 (2011), pp. 301–304.

[40] Pc Fatt and BL Ginsborg. “The ionic requirements for the production of action potentials in crustacean muscle fibres”. In: *J Physiol* 142.3 (1958), pp. 516–543.

[41] Paul Fearnhead and Peter Clifford. “On-line inference for hidden Markov models via particle filters”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.4 (2003), pp. 887–899.

[42] Robert D. Finn et al. “InterPro in 2017—beyond protein family and domain annotations”. In: *Nucleic Acids Research* 45.Database issue (2017), pp. D190–D199.

[43] Dimitry Fisher et al. “A Modeling Framework for Deriving the Structural and Functional Architecture of a Short-Term Memory Microcircuit”. In: *Neuron* 79.5 (2013), pp. 987–1000.

[44] Benjamin F. Fosque et al. “Labeling of active neural circuits in vivo with designed calcium integrators”. In: *Science* 347.6223 (2015), p. 755.

[45] Fabian Fröhlich et al. “Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks”. In: *PLoS computational biology* 13.1 (2017), e1005331.

[46] Walter R Gilks and Carlo Berzuini. “Following a moving target—Monte Carlo inference for dynamic Bayesian models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.1 (2001), pp. 127–146.

[47] Stanley C. Gill and Peter H. von Hippel. “Calculation of protein extinction coefficients from amino acid sequence data”. In: *Analytical Biochemistry* 182.2 (1989), pp. 319–326. ISSN: 0003-2697.

[48] Paolo Giordani and Robert Kohn. “Adaptive Independent Metropolis–Hastings by Fast Estimation of Mixtures of Normals”. In: *Journal of Computational and Graphical Statistics* 19.2 (2010), pp. 243–259.

[49] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. “Novel approach to nonlinear/non-Gaussian Bayesian state estimation”. In: *IEE Proceedings F - Radar and Signal Processing* 140.2 (1993), pp. 107–113.

[50] Clive WJ Granger. “Investigating causal relations by econometric models and cross-spectral methods”. In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438. ISSN: 0012-9682.

[51] D. S. Greenberg, A. R. Houweling, and J. N. Kerr. “Population imaging of ongoing neuronal activity in the visual cortex of awake rats”. In: *Nat Neurosci* 11.7 (2008), pp. 749–51.

[52] D. S. Greenberg and J. N. Kerr. “Automated correction of fast motion artifacts for two-photon imaging of awake animals”. In: *J Neurosci Methods* 176.1 (2009), pp. 1–15.

[53] Benjamin F. Grewe et al. “High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision”. In: *Nat Meth* 7.5 (2010), pp. 399–405.

[54] Ryan N Gutenkunst et al. “Universally sloppy parameter sensitivities in systems biology models”. In: *PLoS computational biology* 3.10 (2007), e189.

[55] Kenneth D. Harris et al. “Accuracy of Tetrode Spike Separation as Determined by Simultaneous Intracellular and Extracellular Measurements”. In: *J Neurophysiol* 84.1 (2000), pp. 401–414.

[56] Christopher D. Harvey, Philip Coen, and David W. Tank. “Choice-specific sequences in parietal cortex during a virtual-navigation decision task”. In: *Nature* 484.7392 (2012), pp. 62–68.

[57] P. Haugland. *Handbook of Biological Fluorescent Probes and Research Chemicals*. Eugene, Oregon: Molecular Probes, 1996.

[58] Martin Havlicek et al. “Physiologically informed dynamic causal modeling of fMRI data”. In: *NeuroImage* 122 (2015), pp. 355–372.

[59] Nordine Helassa et al. “Design and mechanistic insight into ultrafast calcium indicators for monitoring intracellular calcium dynamics”. In: *Scientific reports* 6 (2016), p. 38276. ISSN: 2045-2322.

[60] F. Helmchen, K. Imoto, and B. Sakmann. “Ca2+ buffering and action potential-evoked Ca2+ signaling in dendrites of pyramidal neurons”. In: *Biophys J* 70.2 (1996), pp. 1069–81.

[61] T. Hendel et al. “Fluorescence changes of genetic calcium indicators and OGB-1 correlated with neural activity and calcium in vivo and in vitro”. In: *J Neurosci* 28.29 (2008), pp. 7399–411.

[62] Bertil Hille. *Ion channels of excitable membranes*. Vol. 507. Sinauer Sunderland, MA, 2001.

[63] Alan C Hindmarsh. “ODEPACK, a systematized collection of ODE solvers”. In: *Scientific computing* (1983), pp. 55–64.

[64] S. A. Hires, L. Tian, and L. L. Looger. “Reporting neural activity with genetically encoded calcium indicators”. In: *Brain Cell Biol* 36.1-4 (2008), pp. 69–86.

[65] David H Hubel and Torsten N Wiesel. “Receptive fields of single neurones in the cat’s striate cortex”. In: *J Physiol* 148.3 (1959), pp. 574–591.

[66] Quentin JM Huys and Liam Paninski. “Smoothing of, and parameter estimation from, noisy biophysical recordings”. In: *PLoS computational biology* 5.5 (2009), e1000379.

[67] Roland S. Johansson and Ingvars Birznieks. “First spikes in ensembles of human tactile afferents code complex spatial fingertip events”. In: *Nat Neurosci* 7.2 (2004), pp. 170–177.

[68] Sandro Keller et al. “High-precision isothermal titration calorimetry with automated peak-shape analysis”. In: *Analytical chemistry* 84.11 (2012), pp. 5066–5073.

[69] A.M. Kerlin et al. “Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex”. In: *Neuron* 67.5 (2010), pp. 858–871.

[70] J. N. Kerr, D. Greenberg, and F. Helmchen. “Imaging input and output of neocortical networks in vivo”. In: *Proc Natl Acad Sci U S A* 102.39 (2005), pp. 14063–8.

[71] J. N. Kerr et al. “Spatial organization of neuronal population responses in layer 2/3 of rat barrel cortex”. In: *J Neurosci* 27.48 (2007), pp. 13316–28.

[72] Christina K. Kim et al. “Prolonged, brain-wide expression of nuclear-localized GCaMP3 for functional circuit mapping”. In: *Frontiers in Neural Circuits* 8 (2014), p. 138.

[73] Genshiro Kitagawa. “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models”. In: *Journal of Computational and Graphical Statistics* 5.1 (1996), pp. 1–25.

[74] Martin Klabusay and John R. Blinks. “Some commonly overlooked properties of calcium buffer systems: a simple method for detecting and correcting stoichiometric imbalance in CaEGTA stock solutions”. In: *Cell Calcium* 20.3 (1996), pp. 227–234.

[75] Helmut J. Koester and Bert Sakmann. “Calcium dynamics associated with action potentials in single nerve terminals of pyramidal cells in layer 2/3 of the young rat neocortex”. In: *J Physiol* 529.Pt 3 (2000), pp. 625–646.

[76] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), p. 788.

[77] R Llinas, M Sugimori, and SM Simon. “Transmission by presynaptic spike-like depolarization in the squid giant synapse”. In: *Proceedings of the National Academy of Sciences* 79.7 (1982), pp. 2415–2419.

[78] Jan-Matthis Lueckmann et al. “Flexible statistical inference for mechanistic models of neural dynamics”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 1289–1299.

[79] Linda Madisen et al. “Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance”. In: *Neuron* 85.5 (2015), pp. 942–958.

[80] M. Maravall et al. "Estimating intracellular calcium concentrations and buffering without wavelength ratioing". In: *Biophys J* 78.5 (2000), pp. 2655–2667.

[81] Henry Markram et al. "Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs". In: *Science* 275.5297 (1997), p. 213.

[82] John A. S. McGuigan, James W. Kay, and Hugh Y. Elder. "Critical review of the methods used to measure the apparent dissociation constant and ligand purity in Ca2+ and Mg2+ buffer solutions". In: *Progress in Biophysics and Molecular Biology* 92.3 (2006), pp. 333–370.

[83] Michael D McKay, Richard J Beckman, and William J Conover. "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code". In: *Technometrics* 21.2 (1979), pp. 239–245.

[84] Bruce L McNaughton, John O'Keefe, and Carol A Barnes. "The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records". In: *J Neurosci Methods* 8.4 (1983), pp. 391–397.

[85] Markus Meister, Leon Lagnado, and Denis A. Baylor. "Concerted Signaling by Retinal Ganglion Cells". In: *Science* 270.5239 (1995), p. 1207.

[86] C. Daniel Meliza and Yang Dan. "Receptive-Field Modification in Rat Visual Cortex Induced by Paired Visual Stimulation and Single-Cell Spiking". In: *Neuron* 49.2 (2006), pp. 183–189.

[87] Elliott W. Montroll. "On Coupled Rate Equations with Quadratic Nonlinearities". In: *Proceedings of the National Academy of Sciences of the United States of America* 69.9 (1972), pp. 2532–2536.

[88] T.D. Mrsic-Flogel et al. "Homeostatic Regulation of Eye-Specific Responses in Visual Cortex during Ocular Dominance Plasticity". In: *Neuron* 54.6 (2007), pp. 961–972.

[89] Eran A. Mukamel, Axel Nimmerjahn, and Mark J. Schnitzer. "Automated analysis of cellular signals from large-scale calcium imaging data". In: *Neuron* 63.6 (2009), pp. 747–760.

[90] Junichi Nakai, Masamichi Ohkura, and Keiji Imoto. "A high signal-to-noise Ca2+ probe composed of a single green fluorescent protein". In: *Nature biotechnology* 19.2 (2001), pp. 137–141.

[91] Mohammad Naraghi. "T-jump study of calcium binding kinetics of calcium chelators". In: *Cell Calcium* 22.4 (1997), pp. 255–268.

[92] Ian Nauhaus, Kristina J Nielsen, and Edward M Callaway. "Nonlinearity of two-photon Ca2+ imaging yields distorted measurements of tuning for V1 neuronal populations". In: *Journal of Neurophysiology* 107.3 (2011), pp. 923–936.

[93] Cristopher M. Niell and Michael P. Stryker. "Highly Selective Receptive Fields in Mouse Visual Cortex". In: *The Journal of Neuroscience* 28.30 (2008), p. 7520.

[94] Edward J O'Brien, Jonathan M Monk, and Bernhard O Palsson. "Using Genome-scale Models to Predict Biological Capabilities". In: *Cell* 161.5 (2015), pp. 971–987.

[95] Timothy O'Leary, Alexander C. Sutton, and Eve Marder. "Computational models in the age of large datasets". In: *Current Opinion in Neurobiology* 32 (2015), pp. 87–94.

[96] Andrew C. Oates et al. "Quantitative approaches in developmental biology". In: *Nature Reviews Genetics* 10 (2009), p. 517.

[97] Kenichi Ohki et al. "Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex". In: *Nature* 433.7026 (2005), p. 597.

[98] S. Oiki, T. Yamamoto, and Y. Okada. "Apparent stability constants and purity of Ca-chelating agents evaluated using Ca-selective electrodes by the double-log optimization method". In: *Cell Calcium* 15.3 (1994), pp. 209–216.

[99] Jiri Palecek, Mario B. Lips, and Bernhard U. Keller. "Calcium dynamics and buffering in motoneurones of the mouse spinal cord". In: *J Physiol* 520.Pt 2 (1999), pp. 485–502.

[100] Liam Paninski et al. "A new look at state-space models for neural data". In: *Journal of Computational Neuroscience* 29.1 (2010), pp. 107–126.

[101] George Papamakarios and Iain Murray. "Fast epsilon-free Inference of Simulation Models with Bayesian Conditional Density Estimation". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1028–1036.

[102] Verena Pawlak et al. "Changing the responses of cortical neurons from sub- to suprathreshold using single spikes in vivo". In: *eLife* 2 (2013), e00012.

[103] Nicolas C Pegard et al. "Three-dimensional scanless holographic optogenetics with temporal focusing (3D-SHOT)". In: *Nature communications* 8.1 (2017), p. 1228. ISSN: 2041-1723.

[104] Michael K Pitt. *Smooth particle filters for likelihood evaluation and maximisation*. University of Warwick, Department of Economics, 2002.

[105] Eftychios A Pnevmatikakis et al. “Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data”. In: *Neuron* 89.2 (2016), pp. 285–299.

[106] Thomas A Pologruto, Bernardo L Sabatini, and Karel Svoboda. “ScanImage: flexible software for operating laser scanning microscopes”. In: *Biomedical engineering online* 2.1 (2003), p. 13.

[107] Michelino Puopolo, Elio Raviola, and Bruce P. Bean. “Roles of Subthreshold Calcium Current and Sodium Current in Spontaneous Firing of Mouse Midbrain Dopamine Neurons”. In: *The Journal of Neuroscience* 27.3 (2007), p. 645.

[108] Andreas Raue et al. “Lessons Learned from Quantitative Dynamical Modeling in Systems Biology”. In: *PLoS One* 8.9 (2013), e74335.

[109] Esteban Real et al. “Neural Circuit Inference from Function to Structure”. In: *Current Biology* 27.2 (2017), pp. 189–198.

[110] W. G. Regehr. “Interplay between sodium and calcium dynamics in granule cell presynaptic terminals”. In: *Biophys J* 73.5 (1997), pp. 2476–2488.

[111] F. Sala and A. Hernández-Cruz. “Calcium diffusion modeling in a spherical neuron. Relevance of buffering properties”. In: *Biophys J* 57.2 (1990), pp. 313–324.

[112] Vladislav M. Sandler and Jean-Gaël Barbara. “Calcium-Induced Calcium Release Contributes to Action Potential-Evoked Calcium Transients in Hippocampal CA1 Pyramidal Neurons”. In: *The Journal of Neuroscience* 19.11 (1999), p. 4325.

[113] T. R. Sato et al. “The functional microarchitecture of the mouse barrel cortex”. In: *PLoS Biol* 5.7 (2007), e189.

[114] J. Sawinski et al. “Visually evoked activity in cortical cells imaged in freely moving animals”. In: *Proc Natl Acad Sci U S A* 106.46 (2009), pp. 19557–62.

[115] Volker Scheuss et al. “Nonlinear [Ca2+] Signaling in Dendrites and Spines Caused by Activity-Dependent Depression of Ca2+ Extrusion”. In: *The Journal of Neuroscience* 26.31 (2006), p. 8183.

[116] J. Schiller, F. Helmchen, and B. Sakmann. “Spatial profile of dendritic calcium transients evoked by action potentials in rat neocortical pyramidal neurones”. In: *J Physiol* 487.Pt 3 (1995), pp. 583–600.

[117] Daniel J Simons and Thomas A Woolsey. “Functional organization in mouse barrel cortex”. In: *Brain research* 165.2 (1979), pp. 327–332.

[118] Daniel J Simons et al. “Responses of barrel cortex neurons in awake rats and effects of urethane anesthesia”. In: *Experimental brain research* 91.2 (1992), pp. 259–272.

[119] Nicholas A. Steinmetz et al. “Aberrant Cortical Activity in Multiple GCaMP6-Expressing Transgenic Mouse Lines”. In: *eneuro* 4.5 (2017).

[120] Christoph Stosiek et al. “In vivo two-photon calcium imaging of neuronal networks”. In: *Proceedings of the National Academy of Sciences* 100.12 (2003), pp. 7319–7324.

[121] Wenzhi Sun et al. “Thalamus provides layer 4 of primary visual cortex with orientation- and direction-tuned inputs”. In: *Nat Neurosci* 19.2 (2016), pp. 308–315.

[122] Xiaonan R. Sun et al. “Fast GCaMPs for improved tracking of neuronal activity”. In: *Nature communications* 4 (2013), 10.1038/ncomms3170.

[123] Yvonne N. Tallini et al. “Imaging cellular signals in the heart in vivo: Cardiac expression of the high-signal Ca2+ indicator GCaMP2”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.12 (2006), pp. 4753–4758.

[124] I Tasaki, A Watanabe, and L Lerman. “Role of divalent cations in excitation of squid giant axons”. In: *American Journal of Physiology–Legacy Content* 213.6 (1967), pp. 1465–1474.

[125] Patrick Theer and Winfried Denk. “On the fundamental imaging-depth limit in two-photon microscopy”. In: *Journal of the Optical Society of America A* 23.12 (2006), pp. 3139–3149.

[126] Lucas Theis et al. “Supervised learning sets benchmark for robust spike detection from calcium imaging signals”. In: *bioRxiv* (2015).

[127] Lin Tian et al. “Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators”. In: *Nat Meth* 6.12 (2009), pp. 875–881.

[128] Dustin Tran, Rajesh Ranganath, and David M Blei. “Deep and Hierarchical Implicit Models”. In: *arXiv preprint arXiv:1702.08896* (2017).

[129] R. Y. Tsien. “New calcium indicators and buffers with high selectivity against magnesium and protons: design, synthesis, and properties of prototype structures”. In: *Biochemistry* 19.11 (1980), pp. 2396–404.

[130] Meel Velliste et al. “Cortical control of a prosthetic arm for self-feeding”. In: *Nature* 453.7198 (2008), p. 1098.

[131] Joshua T. Vogelstein et al. “Spike Inference from Calcium Imaging Using Sequential Monte Carlo Methods”. In: *Biophys J* 97.2 (2009), pp. 636–655.

[132] Joshua T Vogelstein et al. “Fast non-negative deconvolution for spike train inference from population calcium imaging”. In: *Journal of Neurophysiology* 104.6 (2010), pp. 3691–3704.

[133] Qi Wang et al. “Structural basis for Calcium Sensing by GCaMP2”. In: *Structure (London, England : 1993)* 16.12 (2008), pp. 1817–1827. ISSN: 0969-2126. DOI: 10.1016/j.str.2008.10.008.

[134] Jack Waters and Fritjof Helmchen. “Background synaptic activity is sparse in neocortex”. In: *Journal of Neuroscience* 26.32 (2006), pp. 8267–8277.

[135] Michael Wehr and Gilles Laurent. “Odour encoding by temporal sequences of firing in oscillating neural assemblies”. In: *Nature* 384.6605 (1996), pp. 162–166.

[136] Matthew A Wilson and Bruce L McNaughton. “Dynamics of the hippocampal ensemble code for space”. In: *Science* 261.5124 (1993), pp. 1055–1058.

[137] Jeffries Wyman and Stanley J Gill. *Binding and linkage: functional chemistry of biological macromolecules.* University Science Books, 1990.

[138] E. Yaksi and R. W. Friedrich. “Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging”. In: *Nat Methods* 3.5 (2006), pp. 377–83.

[139] Yaxiong Yang et al. “Improved calcium sensor GCaMP-X overcomes the calcium channel perturbations induced by the calmodulin in GCaMP”. In: *Nature communications* 9.1 (2018), p. 1504. ISSN: 2041-1723.

[140] Bin Zhang. “Generalized k-harmonic means”. In: *Hewlett-Packard Laboratoris Technical Report* (2000).

[141] Huaying Zhao, Grzegorz Piszczek, and Peter Schuck. “SEDPHAT – a platform for global ITC analysis and global multi-method analysis of molecular interactions”. In: *Methods (San Diego, Calif.)* 76 (2015), pp. 137–148.

[142] Kaiyu Zheng et al. “Time-Resolved Imaging Reveals Heterogeneous Landscapes of Nanomolar Ca2+ in Neurons and Astroglia”. In: *Neuron* 88.2 (2015), pp. 277–288.

[143] Weijian Zong et al. “Fast high-resolution miniature two-photon microscopy for brain imaging in freely behaving mice”. In: *Nature methods* 14.7 (2017), p. 713. ISSN: 1548-7105.

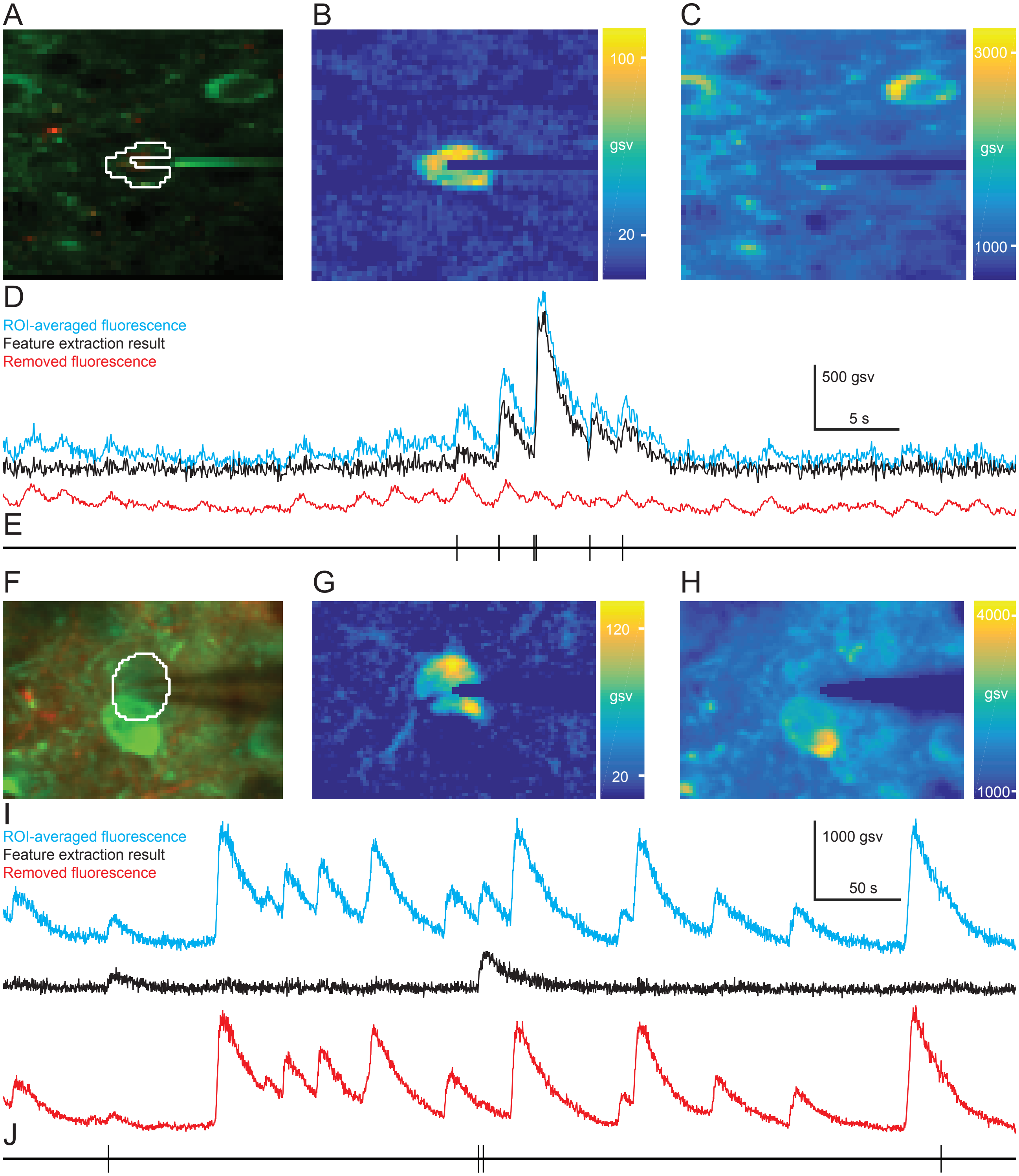

Figure 1–Figure supplement 1. **(A)** Two-photon image acquired using galvanometric scanning (18.6 Hz), showing a neuronal population expressing GCaMP6s (green) in L2/3 of mouse visual cortex, with astrocytes stained using SR101 (red) and electrical recording of APs in a single pyramidal neuron. **(B)** Image showing the mean GCaMP6s fluorescence for each pixel after removal of extraneous signals using the feature extraction algorithm (see methods). **(C)** Image showing the mean GCaMP6s fluorescence removed from each pixel by the feature extraction algorithm. **(D)** Fluorescence time series calculated by a simple average (cyan) over pixels within a region of interest (ROI, white countour in (A)), along with the fluorescence removed by the feature extraction algorithm (red) and the resulting estimate of cytosolic fluorescence (black). Note the ongoing fluctuations in ROI fluorescence due to contamination from the neuropil background. **(E)** Electrically recorded AP times for the data shown in (D). **(F-J)**. As in (A-D), but for a second neuron recorded using resonance scanning (60 Hz). Note that in this case, most of the removed fluorescence arises from a second neuron, which due to the microscope's point spread function overlaps the neuron recorded electrically.

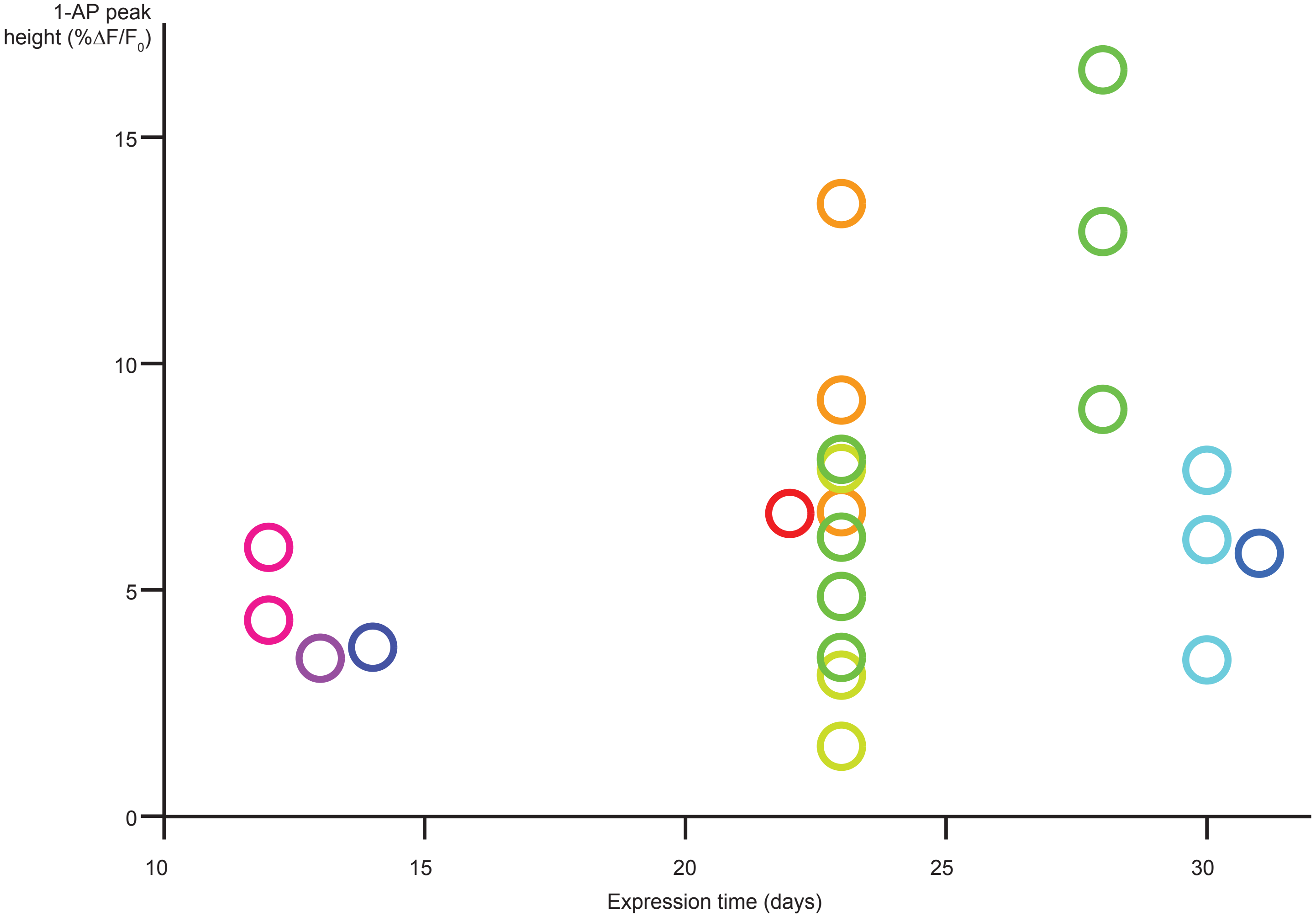

Figure 1–Figure supplement 2. Peak GCaMP6s fluorescence increase evoked by isolated single APs as a function of time since viral injection. Each circle represents one pyramidal neuron, while each color represents one animal.

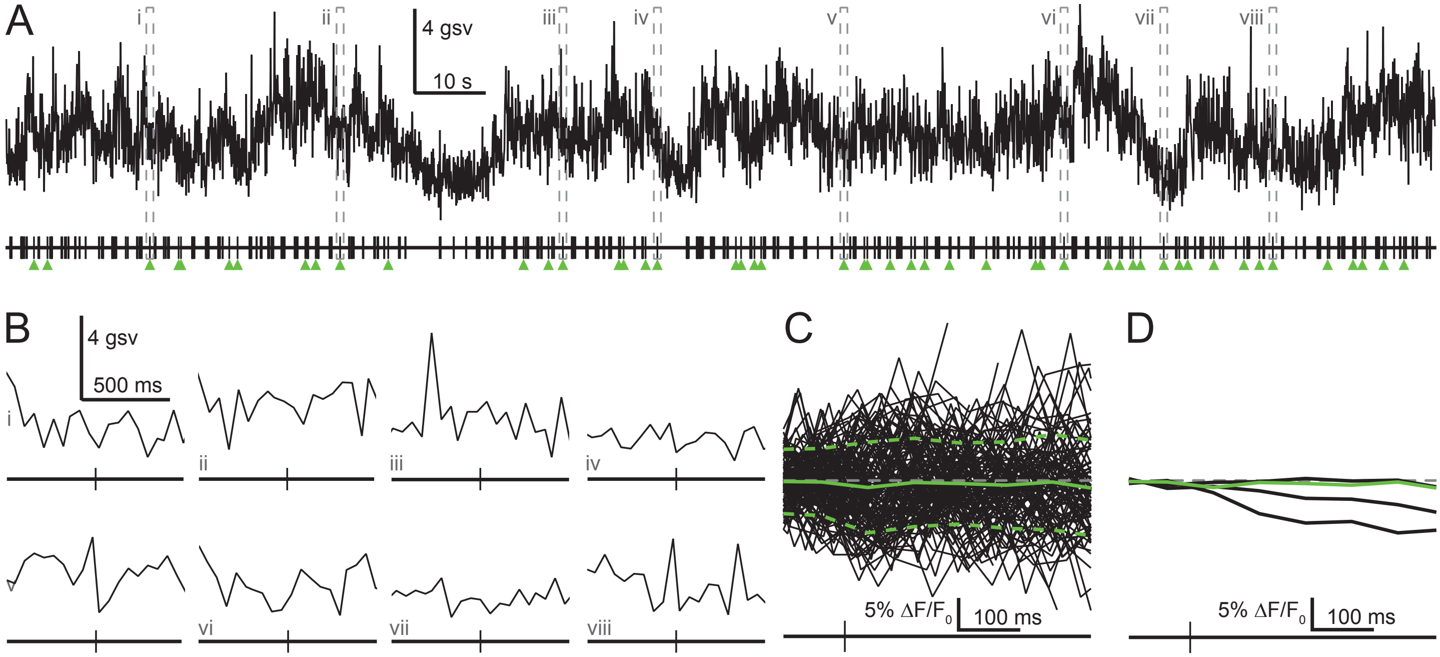

Figure 1–Figure supplement 3. Single APs do not evoke fluorescence increases in interneurons. **(A)** GCaMP6s fluorescence (upper) and simultaneous electrical recording of APs (lower) from an interneuron in L2/3 mouse visual cortex that discharged APs at 2.6 Hz for the data shown and 2.4 Hz overall. Based on this neuron's high firing rate, we defined its isolated single APs (green arrows) as those not preceded by other APs for 300 ms, instead of the 5.5 s required for pyramidal neurons. **(B)** Fluorescence signals recorded during 8 isolated APs within the data shown in (A). **(C)** Fluorescence change relative to baseline during all isolated APs (n = 131) for the neuron shown in (A-B). Solid and dashed green lines indicate mean and standard deviation over events; gray line indicates 0% $\Delta F/F_0$. To calculate $\Delta F/F_0 = (F - F_{\mathrm{BL}})/F_{\mathrm{BL}}$ we estimated baseline fluorescence using the median of fluorescence values over a 60 second window around each time point, since our usual procedure of interpolating fluorescence values from periods at least 5.5 seconds after AP discharge (see methods) could not be applied due to the lack of silent periods. We also subtracted the mean $\Delta F/F_0$ value 0-200 ms before the isolated AP. While it is possible that this procedure could overestimate baseline for frequently firing neurons, resulting in incorrectly scaled $\Delta F/F_0$ values, it nonetheless shows that fluorescence does not show an average increase after single APs. **(D)** Average fluorescence change as in (C), but for all 4 interneurons (additional firing rates 12.5, 7.6 and 8.8 Hz). The green curve indicates the neuron from (A-C). Note that for the neurons with the highest spontaneous firing rates, isolated single APs occurred only after a pause in AP discharge after long periods of frequent AP discharge. As a result, fluorescence decreased after the AP due to decay from the previous period of activity.

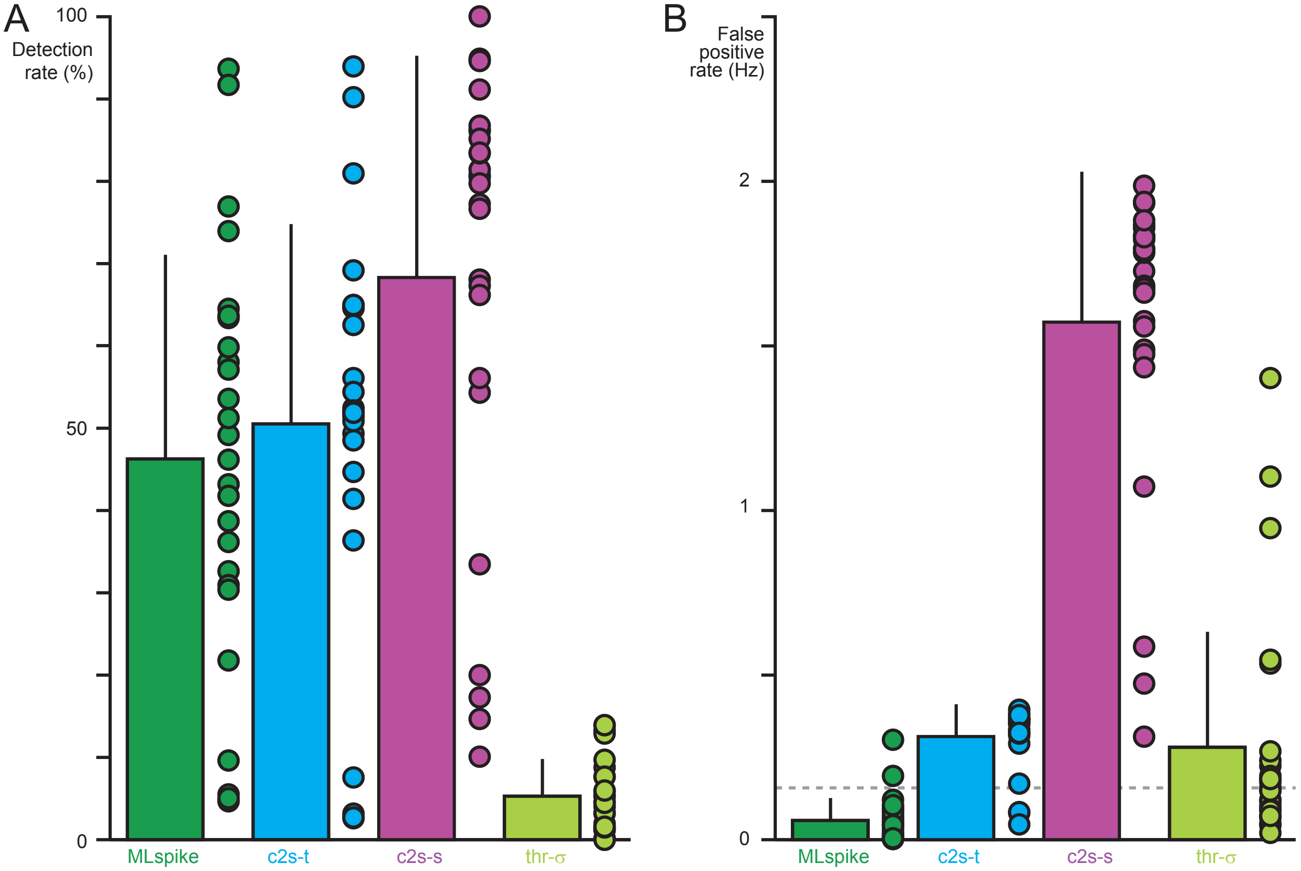

Figure 1–Figure supplement 4. (A) AP detection rates for MLspike (dark green), c2s standard model (magenta), c2s re-trained (cyan) and thr-σ (light green). The detection rate is defined as the fraction of true APs that were successfully identified by each inference method. True and inferred APs within 100 ms were considered as matching. Bar graph shows mean and stadard deviation, while circles denote the detection rate for individual neurons. **(B)** False positive rates for the same inference algorithms. Dashed line indicates the median true firing rate (n = 26).

GCaMP6s

$$\frac{d[\mathrm{GCaMP6s}]}{dt} = -\,k_1^+[\mathrm{Ca}^{2+}][\mathrm{GCaMP6s}] + k_1^-[\mathrm{CaGCaMP6s}]$$

$$\frac{d[\mathrm{CaGCaMP6s}]}{dt} = k_1^+[\mathrm{Ca}^{2+}][\mathrm{GCaMP6s}] - k_1^-[\mathrm{CaGCaMP6s}] - k_2^+[\mathrm{Ca}^{2+}][\mathrm{CaGCaMP6s}] + k_2^-[\mathrm{Ca_2GCaMP6s}]$$

$$\frac{d[\mathrm{Ca_2GCaMP6s}]}{dt} = k_2^+[\mathrm{Ca}^{2+}][\mathrm{CaGCaMP6s}] - k_2^-[\mathrm{Ca_2GCaMP6s}] - k_3^+[\mathrm{Ca}^{2+}][\mathrm{Ca_2GCaMP6s}] + k_3^-[\mathrm{Ca_3GCaMP6s}]$$

$$\frac{d[\mathrm{Ca_3GCaMP6s}]}{dt} = k_3^+[\mathrm{Ca}^{2+}][\mathrm{Ca_2GCaMP6s}] - k_3^-[\mathrm{Ca_3GCaMP6s}] - k_4^+[\mathrm{Ca}^{2+}][\mathrm{Ca_3GCaMP6s}] + k_4^-[\mathrm{Ca_4GCaMP6s}]$$

$$\frac{d[\mathrm{Ca_4GCaMP6s}]}{dt} = k_4^+[\mathrm{Ca}^{2+}][\mathrm{Ca_3GCaMP6s}] - k_4^-[\mathrm{Ca_4GCaMP6s}]$$

Endogenous buffers

$$\frac{d[\mathrm{CaB_1}]}{dt} = b_1^+[\mathrm{Ca}^{2+}]([\mathrm{B}]_1^{\mathrm{total}} - [\mathrm{CaB_1}]) - b_1^-[\mathrm{CaB_1}]$$

$$\frac{d[\mathrm{CaB_2}]}{dt} = b_2^+[\mathrm{Ca}^{2+}]([\mathrm{B}]_2^{\mathrm{total}} - [\mathrm{CaB_2}]) - b_2^-[\mathrm{CaB_2}]$$

Free calcium

$$\frac{d[\mathrm{Ca}^{2+}]}{dt} = \underbrace{\sum_{j=1}^{4}\Big(k_j^-[\mathrm{Ca}_j\mathrm{GCaMP6s}] - k_j^+[\mathrm{Ca}^{2+}][\mathrm{Ca}_{j-1}\mathrm{GCaMP6s}]\Big)}_{\text{Binding to / release from GECI}} + \underbrace{\sum_{k=1}^{2}\Big(b_k^-[\mathrm{CaB}_k] - b_k^+[\mathrm{Ca}^{2+}]([\mathrm{B}]_k^{\mathrm{total}} - [\mathrm{CaB}_k])\Big)}_{\text{Binding to / release from endogenous buffers}} - \underbrace{\frac{[\mathrm{Ca}^{2+}] - [\mathrm{Ca}^{2+}]_{\mathrm{rest}}}{\tau_{\mathrm{ex}}}}_{\text{Extrusion}}$$

Figure 2–Figure supplement 1. Global rate equation describing mass action kinetics and extrusion. Time derivatives of the concentrations of GCaMP6s binding states, endogenous buffer binding states and free calcium are given as instantaneous functions of the current concentrations.

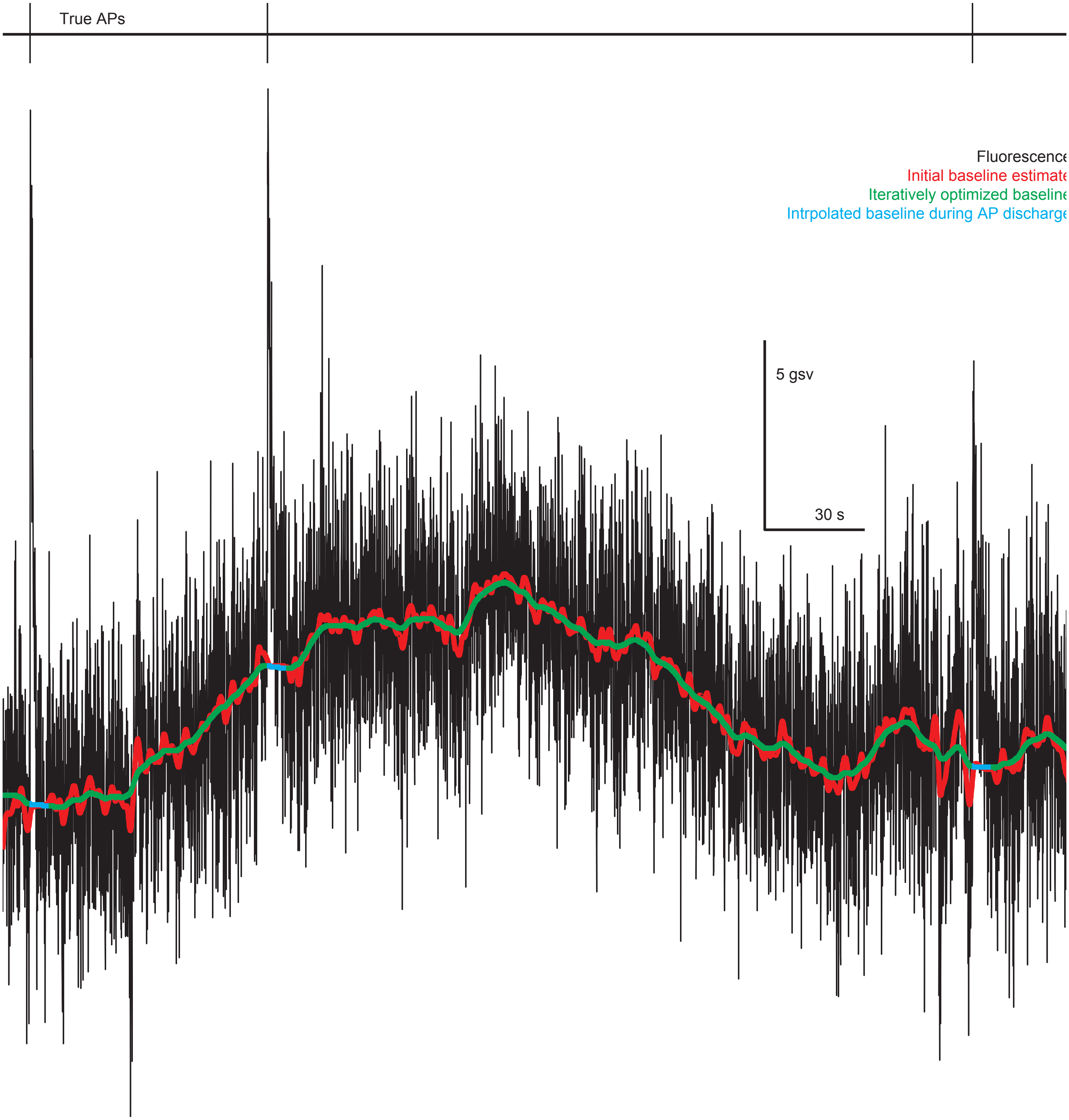

Figure 3–Figure supplement 1. Estimation of drifting baseline fluorescence from fluorescence signals and known AP times acquired *in vivo* (full details in methods). Starting from APs (upper) and fluorescence (lower, black), an initial estimate (red) is computed by Gaussian filtering of the fluorescence data values, with fluorescence values occurring less than 5.5 seconds after an AP excluded (see methods). An iterative procedure then computes a final estimate of baseline fluorescence for periods not following an AP (green), which is interpolated to fill in the missing values following each AP (cyan).

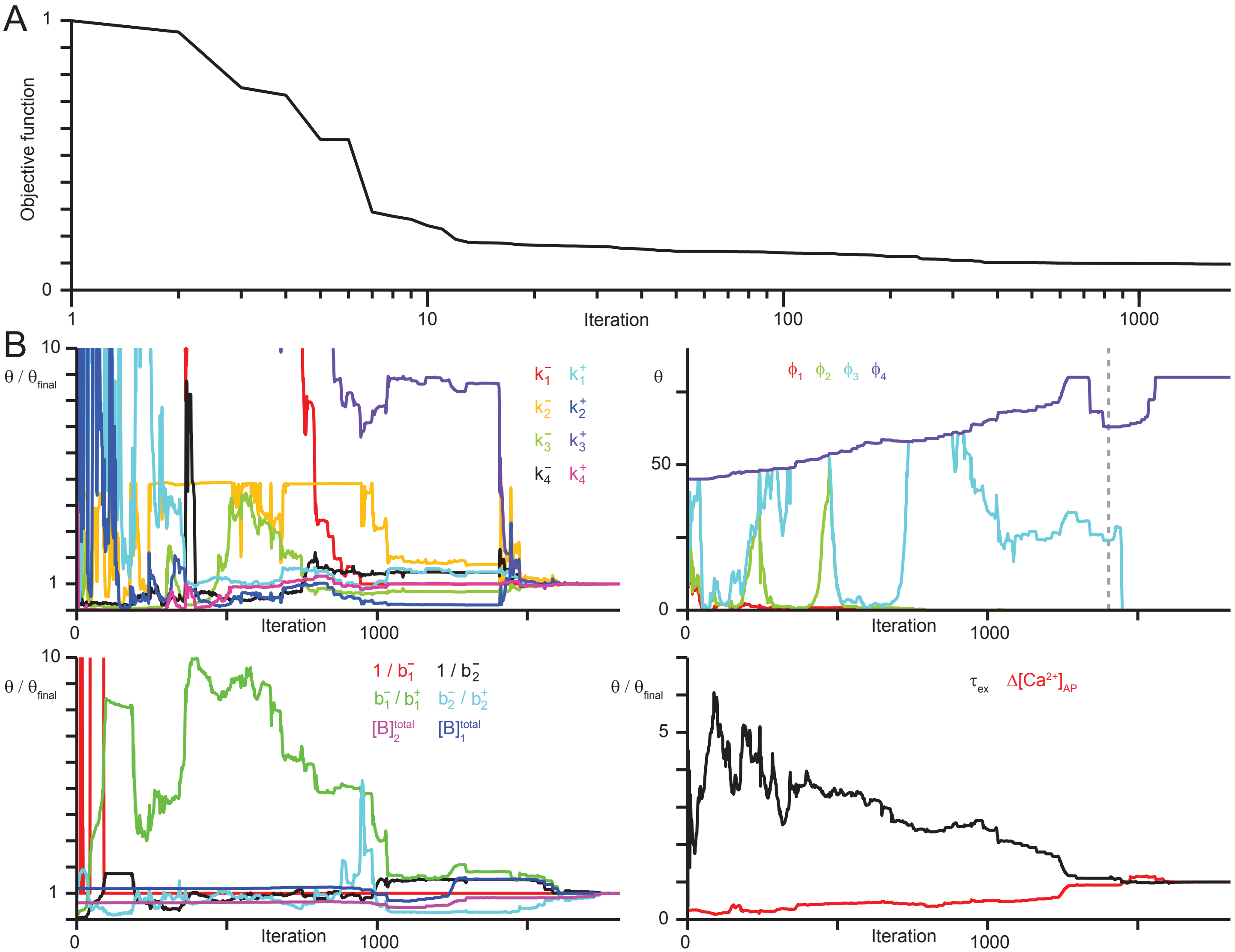

Figure 3–Figure supplement 2. Convergence of SBM parameters while fitting *in vivo* data. **(A)** Values of the objective function e_{vivo} over successive iterations while fitting the SBM to full set of optical/electrical recordings (n = 22 pyramidal neurons). **(B)** Convergence of SBM parameters during optimization. *Upper left*: rate constants normalized to their final values. *Upper right*: values of ϕ, denoting brightness values for each binding state relative to the calcium free state. Lower left: time constants, dissociation constants and total concentrations of the two endogenous buffers, normalized to their final values. Lower right: AP-evoked calcium influx and extrusion time constant, normalized to their final values.

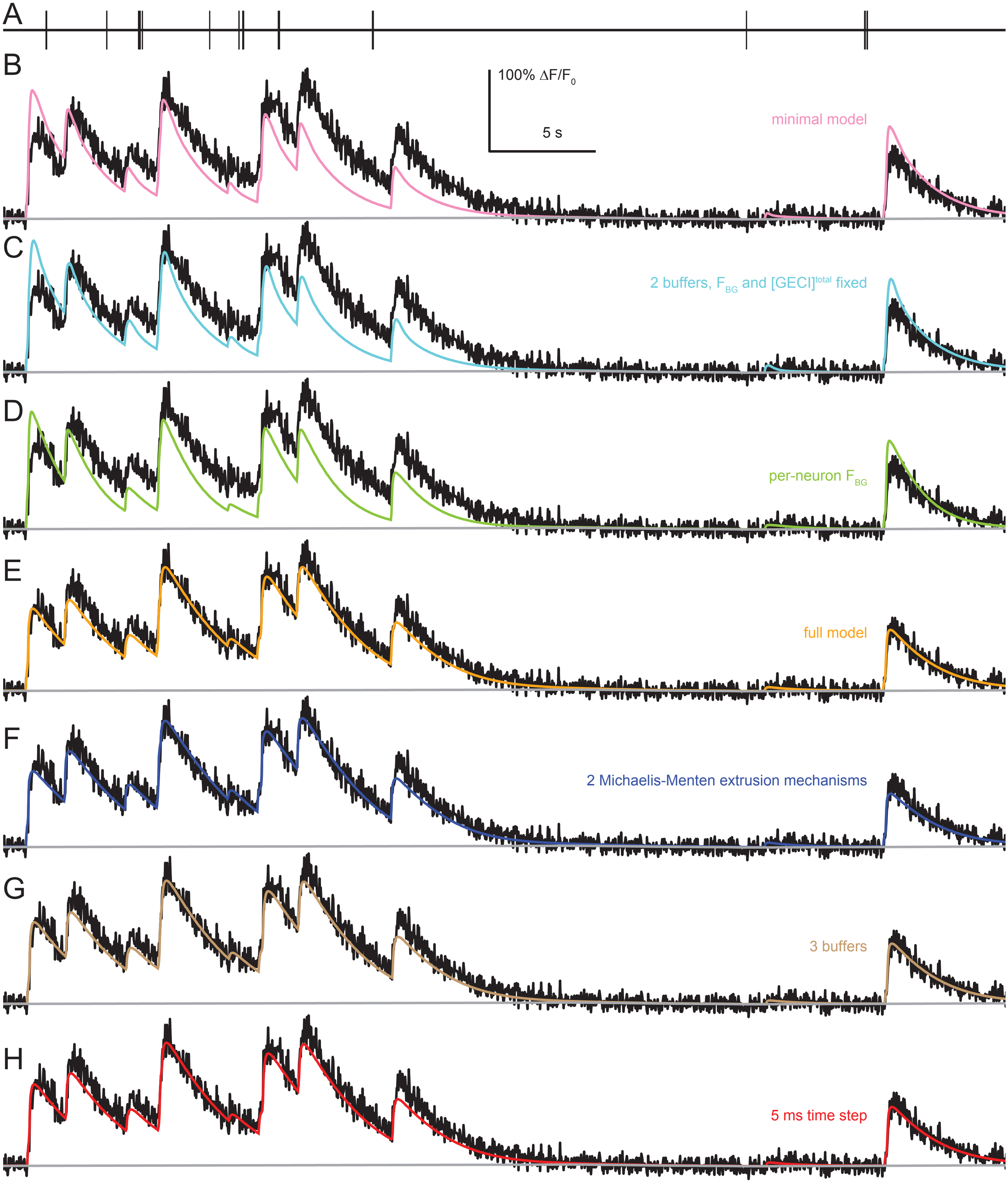

Figure 3–Figure supplement 3. Comparison of SBM fits to data with increasing model complexity. **(A)** AP sequence recorded from a L2/3 mouse visual cortical pyramidal neuron *in vivo*. **(B)** Fluorescence (black) compared to SBM fit using a reduced minimal' model without endogenous buffers or variation in F_{BG} or $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ over neurons. **(C)** SBM fit when including 2 endogenous calcium buffers but still without variation in F_{BG} or $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ over neurons. **(D)** SBM fit including variation in F_{BG} across neurons, allowing the amplitude but not the shape of AP-evoked fluorescence to be different for each neuron. **(E)** SBM fit when allowing both F_{BG} and $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ to vary over neurons; this is the SBM variant used throughout the present study unless otherwise noted. **(F)** Fit for extended SBM, consisting of the full SBM in (E) along with 2 Michaelis-Menten extrusion mechanisms, allowing for calcium-dependent modulation of extrusion rate. **(G)** Fit for extended SBM, consisting of the full SBM in (E) but with 3 endogenous buffers instead of 2. **(H)** Fit for the full SBM shown in (E), but when solving the rate equation with a maximum time step of 5 instead of 10 ms.

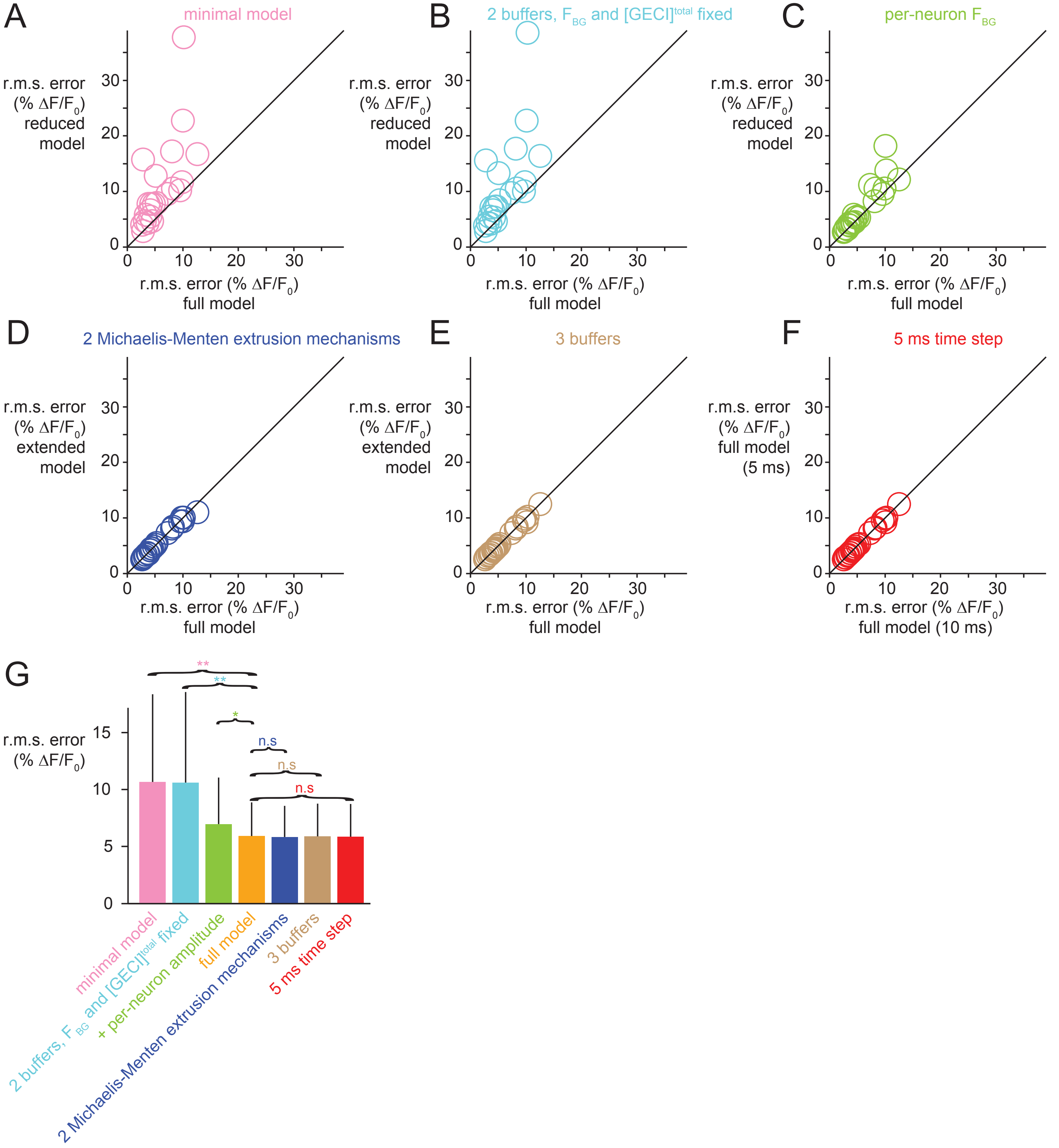

Figure 3–Figure supplement 4. Comparison of SBM variants over pyramidal neurons (n = 22) shows that the standard SBM fits the data better than simplified versions, but using more complicated variants do not improve fit quality. **(A)** Root-mean-square error for the standard SBM compared to a reduced 'minimal' version without buffering or variation in F_{BG} or [GCaMP6s]$^{\mathrm{total}}$ over neurons (Figure 3 - figure supplement 2b). **(B)** Comparison of standard SBM to reduced version with 2 endogenous calcium buffers but no variation in F_{BG} or [GCaMP6s]$^{\mathrm{total}}$ over neurons (Figure 3 - figure supplement 2c). **(C)** Comparison of SBM to reduced version which also includes variation in F_{BG} but not [GCaMP6s]$^{\mathrm{total}}$ over neurons (Figure 3 - figure supplement 2d). **(D)** Comparison of standard SBM to an extended version using two Michaelis-Menten extrusion mechanisms, each with its own Michaelis constant and maximum rate (see methods). **(E)** Comparison o f standard SBM to an extended version with 3 endogenous buffers (each with its own concentration, affinity and time constant). **(F)** Comparison of standard SBMs with maximum time steps of 5 vs. 10 ms. **(G)** Mean and standard deviation of root-mean-square error over neurons for each SBM variant. Single asterisk indicates $p < 0.01$, double asterisks indicate $p <$ 1e-6.

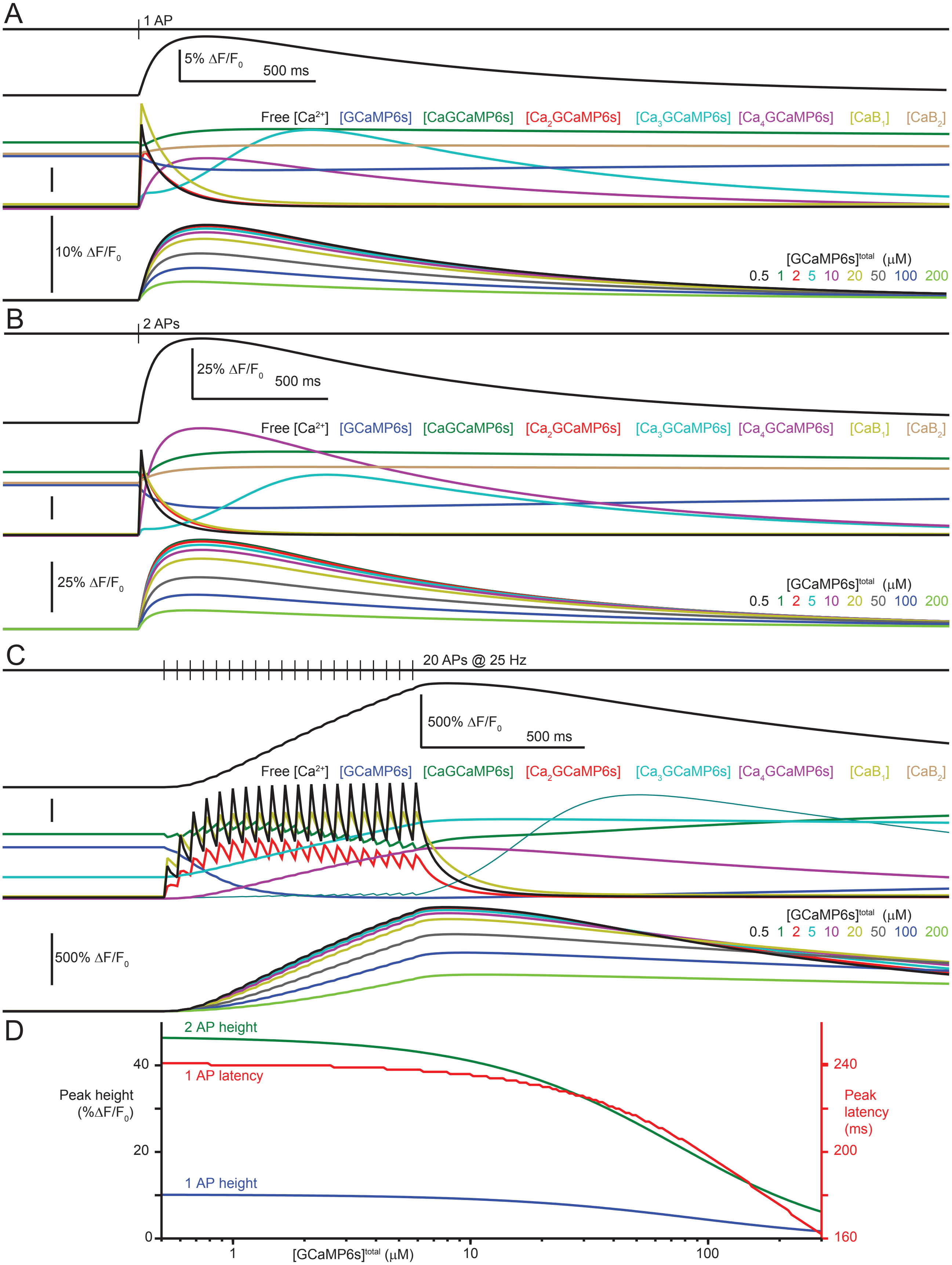

Figure 3–Figure supplement 5. SBM simulations using parameters fit to *in vivo* data ($F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}} = 2.5$). **(A)** Fluorescence response (upper) to one AP using $[\mathrm{GCaMP6s}]^{\mathrm{total}} = 5\ \mu\mathrm{M}$, with binding state concentrations over time (middle, scalebar: 400 nM Ca^{2+}, 1 μM GCaMP6s, 1 μM CaGCaMP6s, 100 nM $\mathrm{Ca_2GCaMP6s}$, 0.4 nM $\mathrm{Ca_3GCaMP6s}$, 10 nM $\mathrm{Ca_4GCaMP6s}$, 4 μM $\mathrm{B_1}$, 4 μM $\mathrm{B_2}$). Fluorescence response to one AP as a function of $[\mathrm{GCaMP6s}]^{\mathrm{total}}$ (lower). **(B)** As in (A) but for a 2-AP burst (scalebar: 1 μM Ca^{2+}, 1 μM GCaMP6s, 1 μM CaGCaMP6s, 200 nM $\mathrm{Ca_2GCaMP6s}$, 2 nM $\mathrm{Ca_3GCaMP6s}$, 20 nM $\mathrm{Ca_4GCaMP6s}$, 10 μM $\mathrm{B_1}$, 4 μM $\mathrm{B_2}$). **(C)** As in (A-B) but for 20 APs at 25 Hz (scalebar: 1 μM Ca^{2+}, 1 μM GCaMP6s, 1 μM CaGCaMP6s, 400 nM $\mathrm{Ca_2GCaMP6s}$, 20 nM $\mathrm{Ca_3GCaMP6s}$, 1 μM $\mathrm{Ca_4GCaMP6s}$, 10 μM $\mathrm{B_1}$, 10 μM $\mathrm{B_2}$). **(D)** Simulated peak fluorescence evoked by one (blue) and two APs (green) and single-AP peak latency (red) as functions of $[\mathrm{GCaMP6s}]^{\mathrm{total}}$.

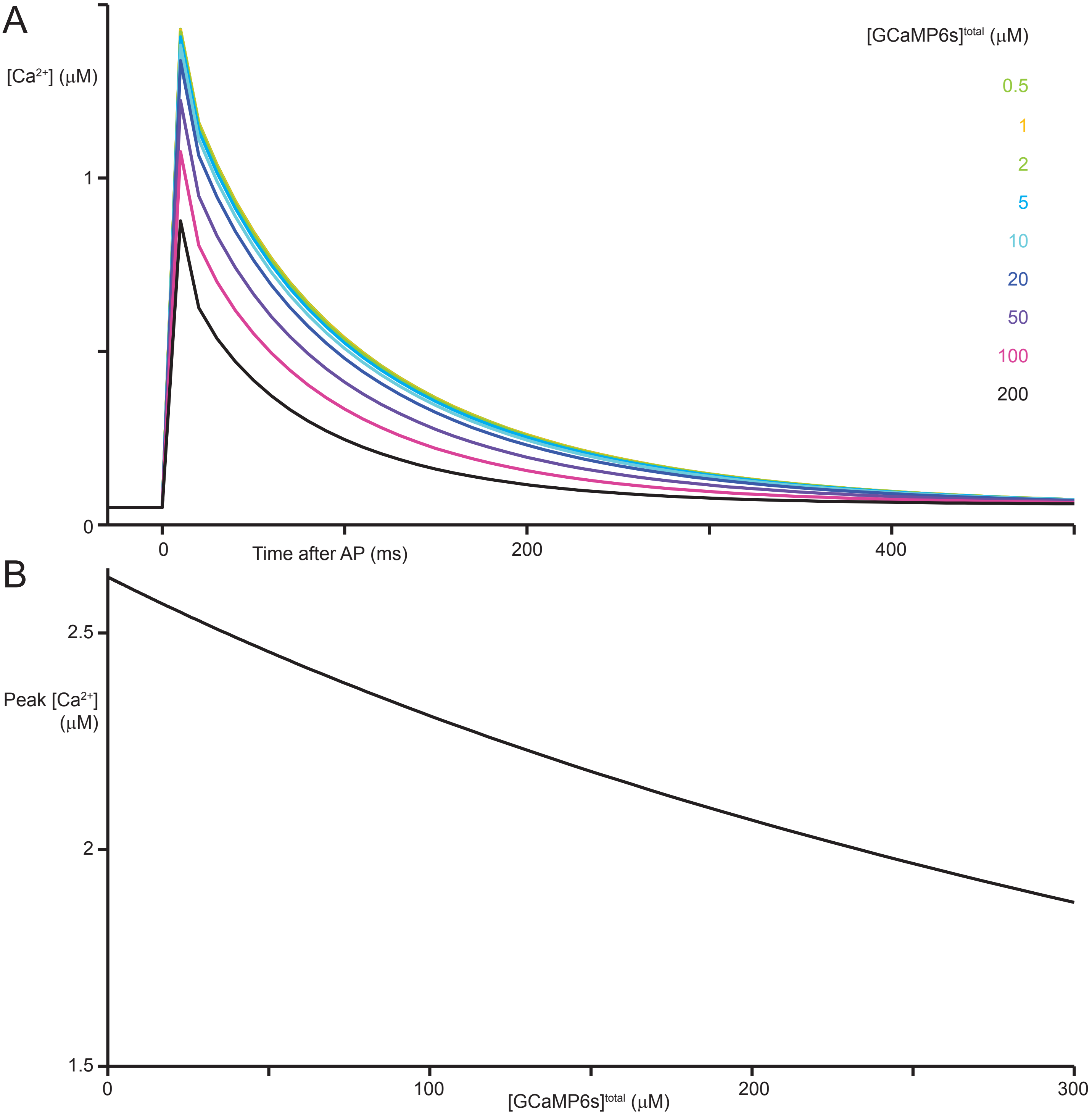

Figure 3–Figure supplement 6. The concentration of free calcium over time after AP discharge depends on total GCaMP6s concentration in SBM simulations. **(A)** SBM simulation of calcium concentration after a single AP as a function of total GCaMP6s concentration. At low concentrations GCaMP6s does not contribute significantly to calcium buffering compared to endogenous buffers, while at higher concentrations it acts as a buffer, reducing the free calcium concentration in the cytosol. **(B)** Peak [Ca^{2+}] one time step ($\delta \leq 10$ ms) after a single AP as a function of total GCaMP6s concentration.

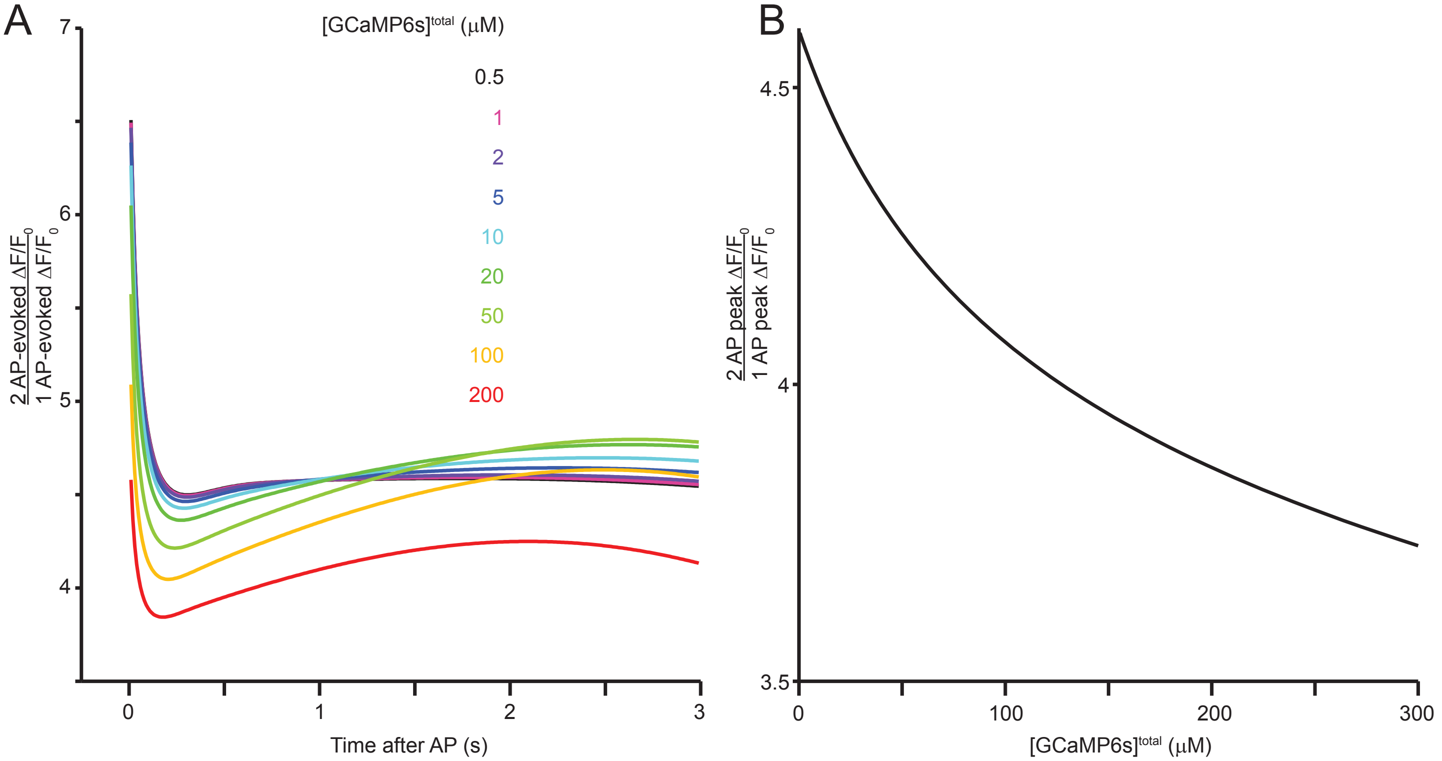

Figure 3–Figure supplement 7. Nonlinearity of AP-evoked fluorescence depends on total GCaMP6s concentration in SBM simulations ($F_{\mathrm{BG}}/F_{\mathrm{cyt}}^{\mathrm{eq}} = 2.5$). **(A)** 2-AP fluorescence response divided by 1-AP response, as a function of time after AP discharge and for a range of total GCaMP6s concentrations. Note that the ratio of the 2-AP response to the 1-AP response is not constant over time for any GCaMP6s concentration, and that the shape of the ratio as a function of time depends on the GCaMP6s concentration. **(B)** Ratio of peak 2-AP evoked fluorescence to peak 1-AP fluorescence as a function of GCaMP6s concentration.

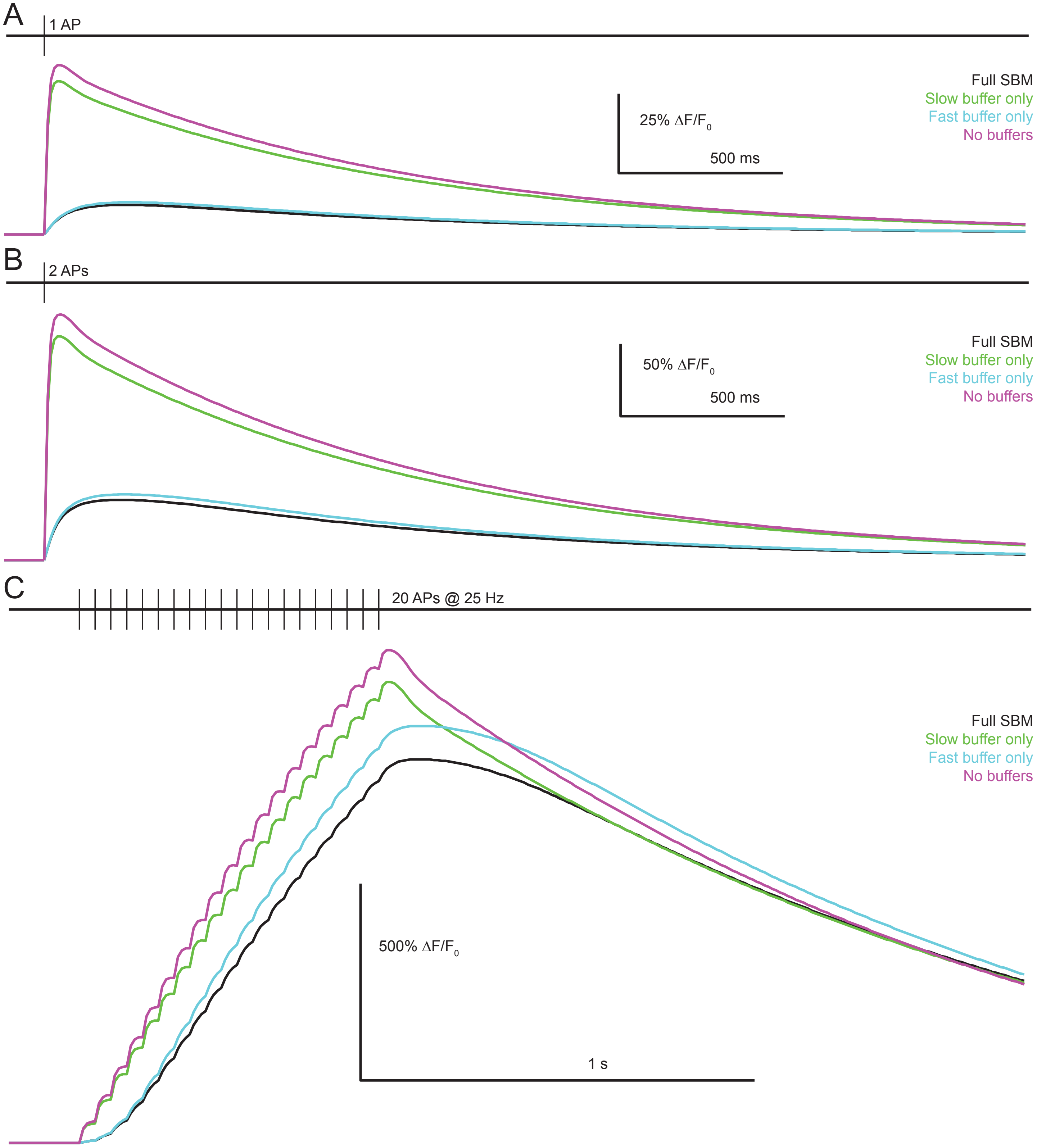

Figure 3–Figure supplement 8. Role of endogenous buffers in shaping AP-evoked fluorescence in SBM simulations ($[\text{GCaMP6s}]^{\text{total}} = 5$ µM, $F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}} = 2.5$). **(A)** Fluorescence arising from discharge of one AP in simulations from the full SBM (black) or when incorporating only the slow buffer (green), only the fast buffer (cyan) or neither buffer (magenta). **(B)** As in (A), but for a burst of 2 APs. **(C)** As in (A-B), but for a train of 20 APs at 50 Hz. Note that the slow buffer plays only a minor role in determining fluorescence responses to discharge of 1 or 2 APs, while the fast buffer has no effect on the decay phase following the train of 20 APs.

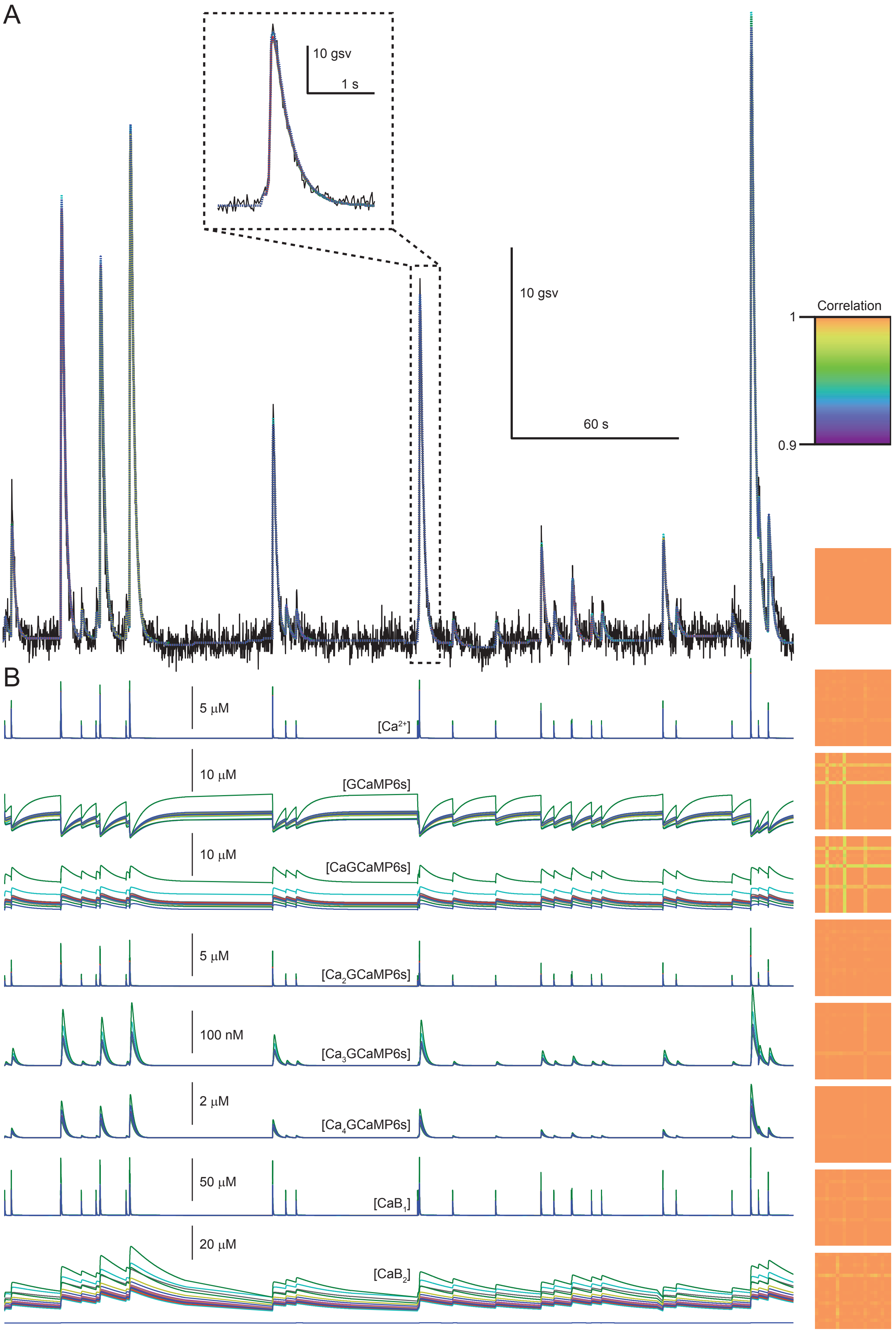

Figure 3–Figure supplement 9. Variation of SBM fluorescence predictions over multiple parameter sets. **(A)** *Left*: Fluorescence (black, same data as Figure 4a-c) with predictions from each parameter set (colored dashed lines). Each set of SBM parameters was obtained by fitting the *in vivo* dataset while excluding a single neuron. *Right*: Correlation matrix showing correlation over time of fluorescence predictions from the 22 parameter sets. **(B)** Binding state concentrations from the fits in (A) (as in Figure 3b-c), with correlation over time of binding state concentrations from the 22 parameter sets.

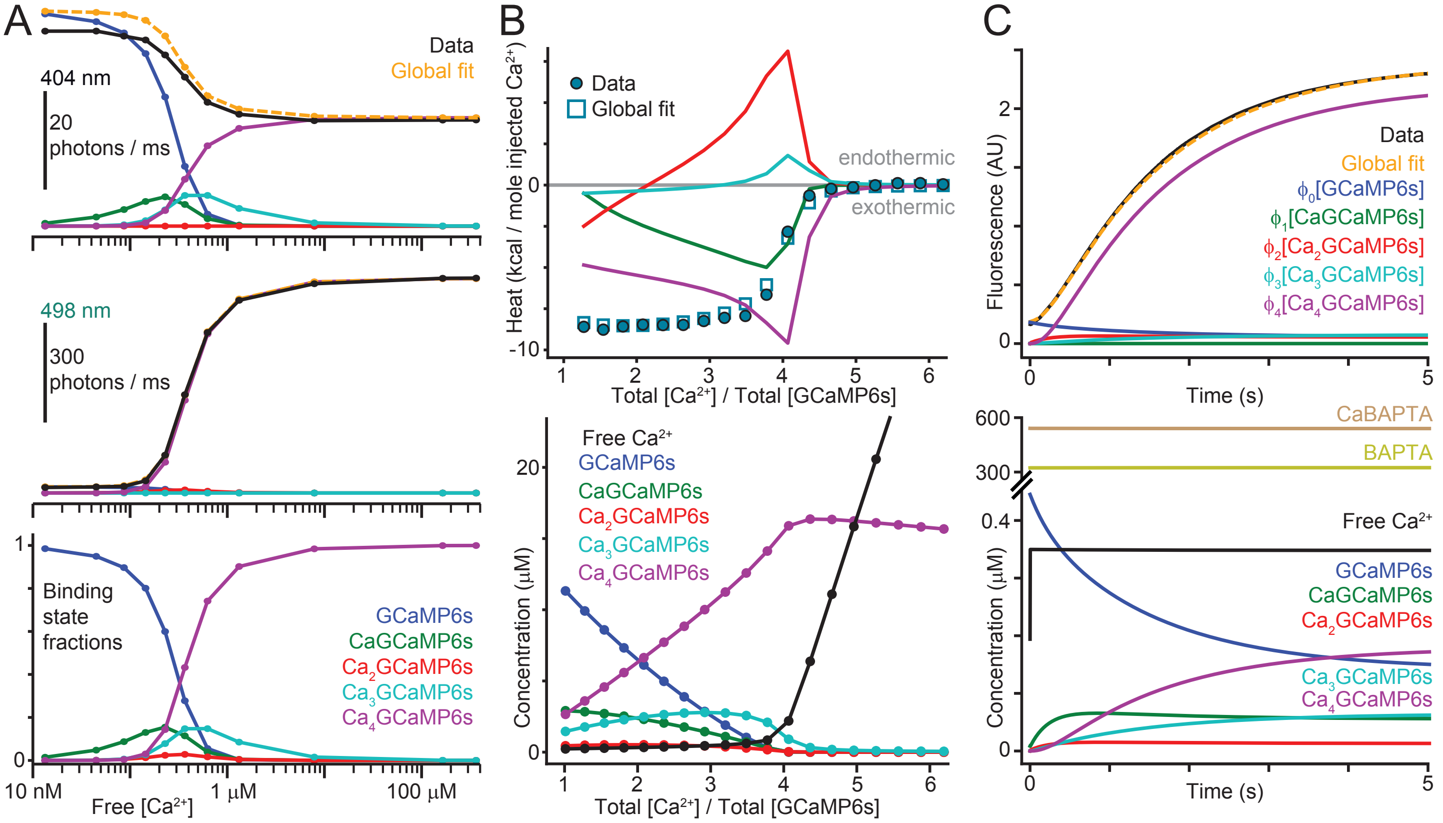

Figure 4–Figure supplement 1. Decomposition of *in vitro* binding assay data into contributions from each GCaMP6s binding state. **(A)** Fluorescence (black) excited at 404 nm (upper) and 498 nm (middle) as a function of free calcium concentration at each step of a titration used to measure excitation spectra (Figure 4a,d), and prediction fluorescence from SBM global fit (orange). Additional colors show contributions of each binding state to total predicted fluorescence; excitation wavelengths are shown as arrowheads in Figure 4c. Modeled binding state fractions (lower) as a function of free calcium for the same data. **(B)** Integrated peak heats (circles, upper) for the data shown in turqouise in Figure 4b,d along with global fit (squares) and contributions to enthalpy changes arising from changes in each binding state concentration (colored curves). Modeled concentrations of free Ca^{2+} (black, lower) and each GCaMP6s binding state are shown for the same data. Free Ca^{2+} increases slowly until the four binding sites of GCaMP6s are saturated, after which it increases rapidly. The concentration of the saturated state Ca$_4$GCaMP6s initially increases due to binding, then decreases due to dilution and perfusion. **(C)** Upper graph: Fluorescence signals acquired during stopped-flow experiment (black) and global fit (orange) for the data in Figure 4f showing binding kinetics after a transition from 17 nM to 348 nM free Ca^{2+}. Additional colors show the modeled fluorescence contributions of each GCaMP6s binding state. Lower graph: modeled concentrations over time for all molecular species.

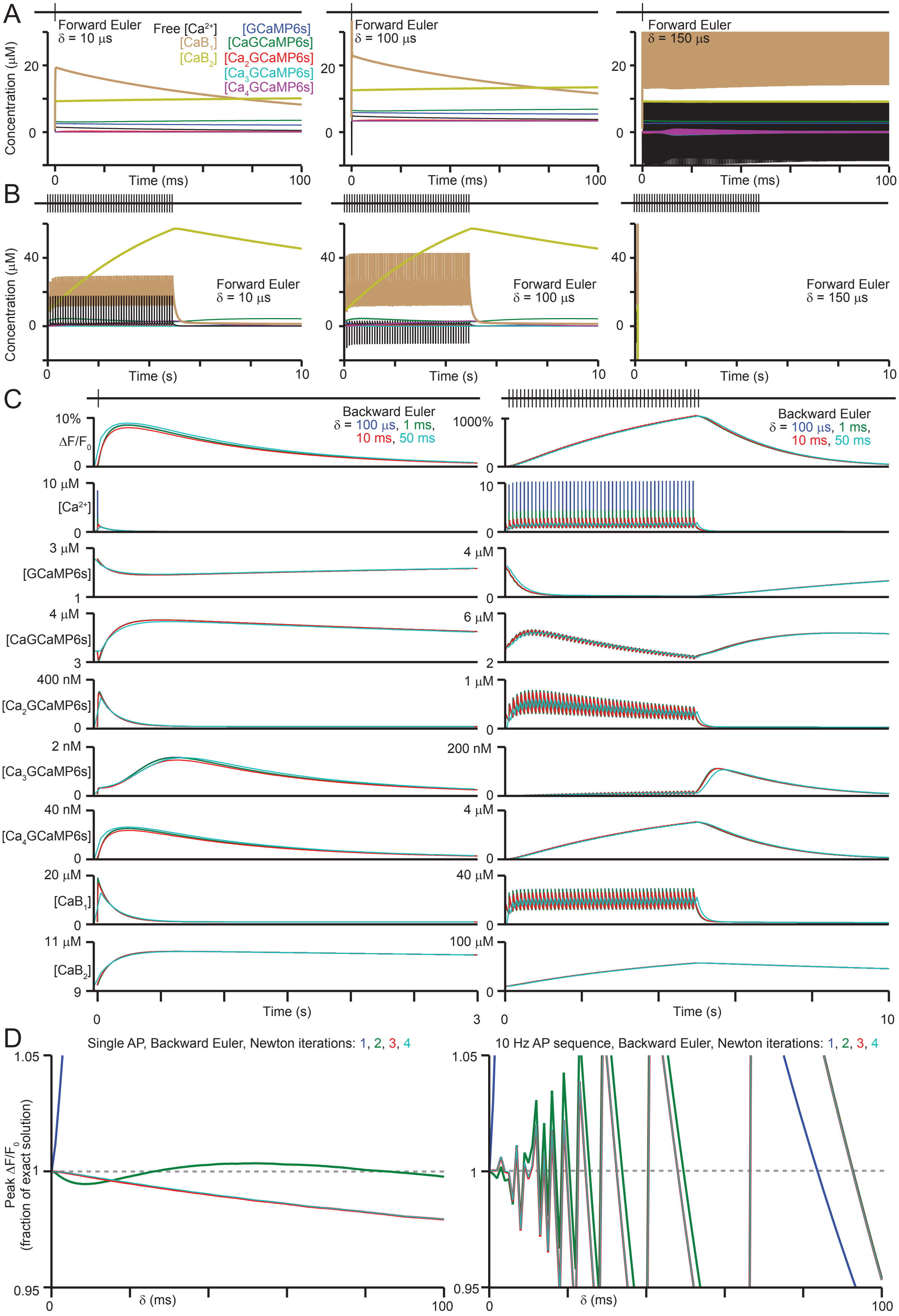

Figure 5–Figure supplement 1. (A) Forward (explicit) Euler integration of the SBM rate equation for 1 AP, with $\delta = 10$ (left), 100 (center) or 150 µs. Steps >50 µs resulted in negative concentrations. **(B)** As in (A), but for 50 APs at 10 Hz. Steps >100 µs caused divergence to infinity. **(C)** Backward Euler integration (see methods) for the AP sequences in (A-B) with 3 Newton iterations and $\delta = 0.1$, 1, 10 and 50 ms. Note the close agreement for $\delta \leq 10$ ms, and that transient differences in [Ca^{2+}] for $\delta \leq 10$ ms do not lead to differences in binding state concentrations or predicted fluorescence. **(D)** Peak $\Delta F/F_0$ values for the AP sequences in (A-C), relative to the exact solution and as a function of δ for backward Euler integration, with 1-4 Newton iterations. Note the nearly identical solutions for 3 (the default value for model fitting and AP inference) vs. 4 Newton iterations.

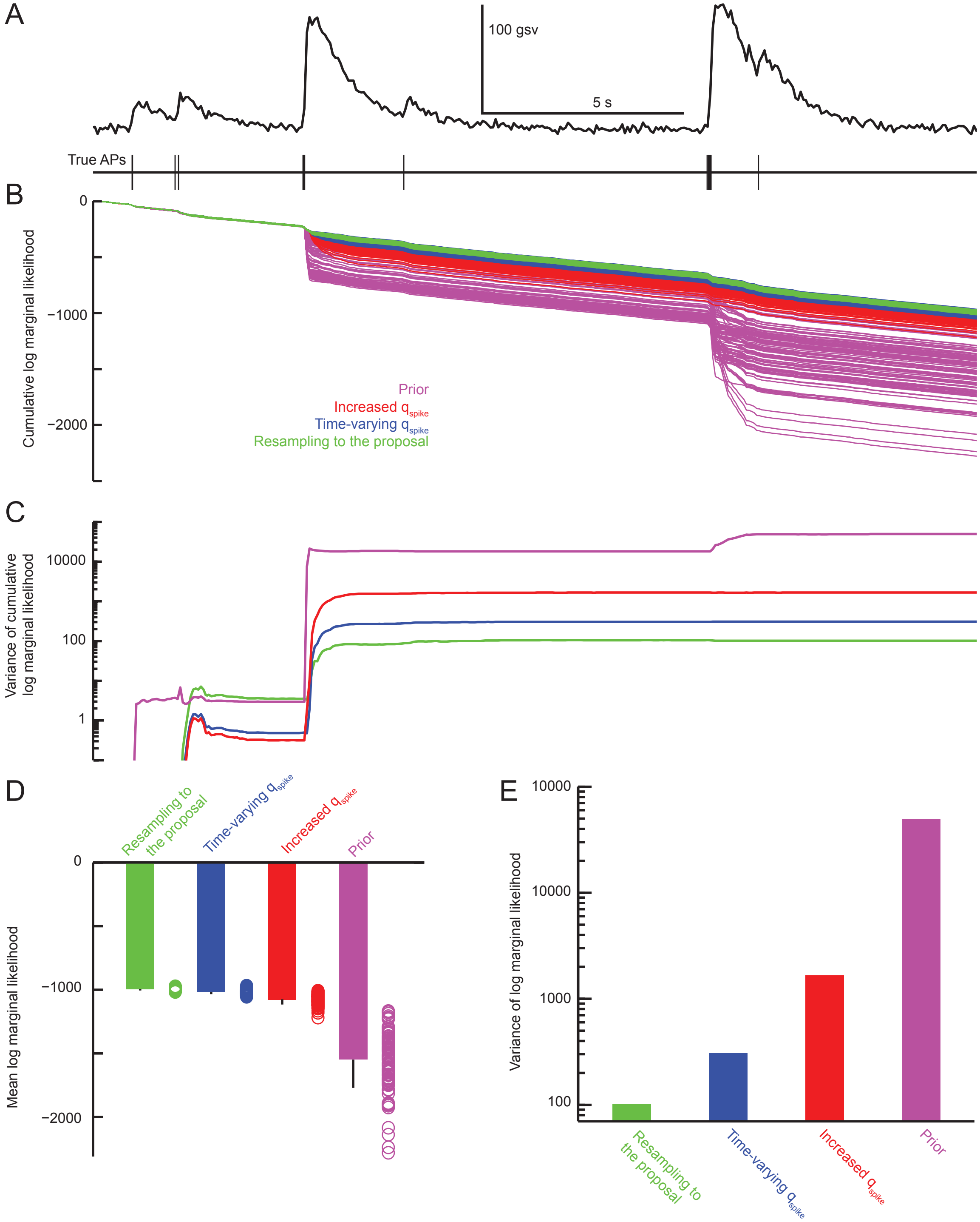

Figure 5–Figure supplement 2. Sampling and resampling techniques for reducing the variability of the SMC algorithm's output. **(A)** Fluorescence and APs recorded for 22 seconds from a pyramidal neuron in L2/3 mouse visual cortex. **(B)** Cumulative log marginal likelihood calculated for the data in (A). At each time point the SMC technique (see methods) is used to calculate the likelihood given the model parameters of all observed data up to the current time point. Results are shown for 100 independent runs using the same parameters and data, with sampling of AP sequences from the prior distribution (magenta), a proposal distribution with increased AP discharge probability (red), a time varying proposal in a two-round scheme (blue) or a time varying proposal with resampling to the proposal (green). **(C)** Variance over runs of the cumulative log marginal likelihoods shown in (B). **(D)** Distribution over runs of marginal likelihood at the end of the data in (A-C) (circles). Bar graph shows means and standard deviations. **(E)** Variance over runs for the distributions in (D).

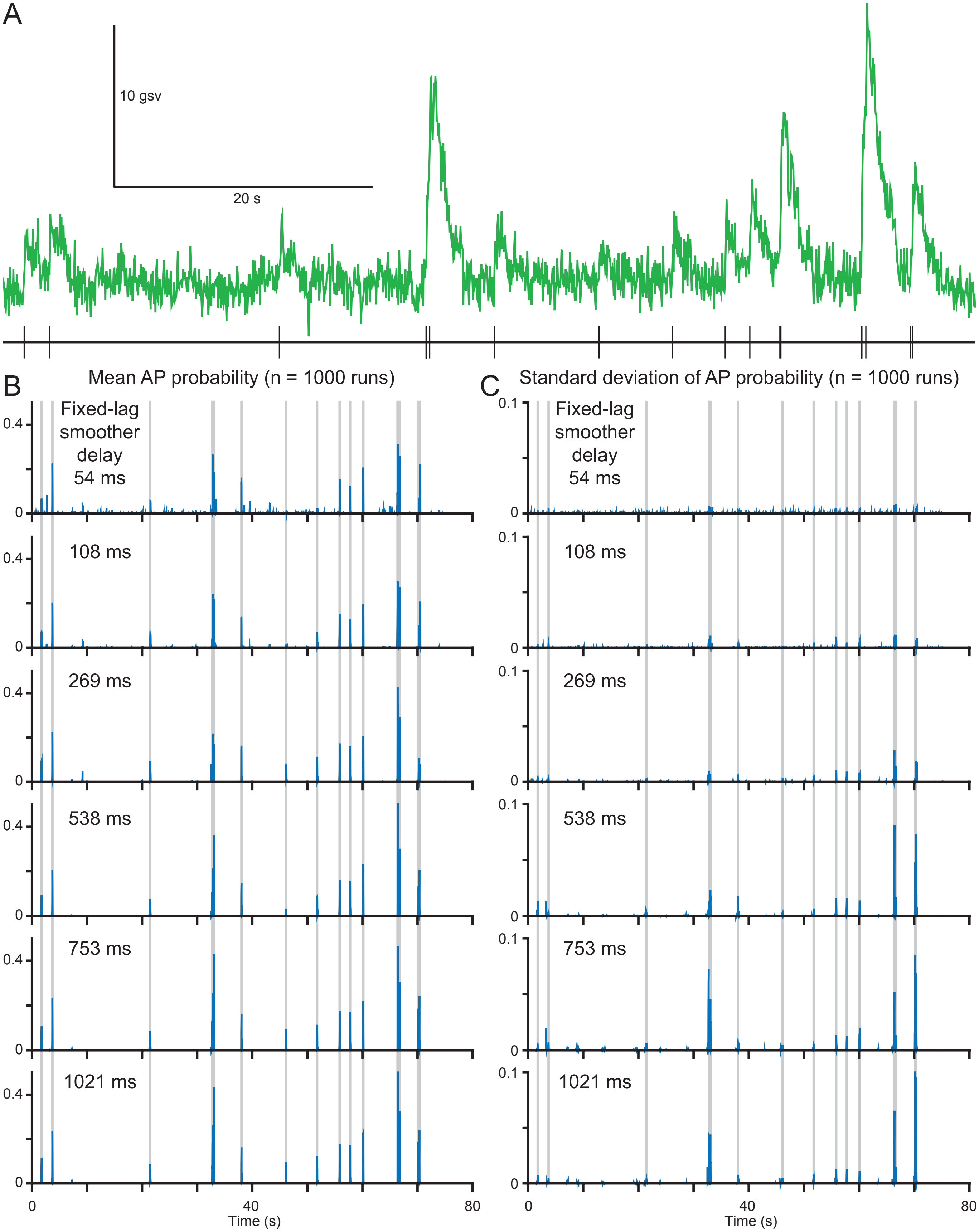

Figure 5–Figure supplement 3. Effect of the fixed-lag smoother delay on inferred AP discharge probability using the SMC algorithm with the SBM. **(A)** Fluorescence signals (upper) and electrically recorded APs (lower) from a L2/3 mouse visual cortical pyramidal neuron. **(B)** Probability of AP discharge every 10 ms, inferred using the filter smoother and averaged over 1000 runs of the algorithm. The SMC algorithm was used with 102400 particles. Results are shown for fixed-lag smoother delays 54 (top), 108, 269, 538, 753 and 1021 ms (bottom). Same data as in (A); gray lines indicate true AP times. **(C)** As in (B), but showing the standard deviation over 1000 runs of the algorithm.

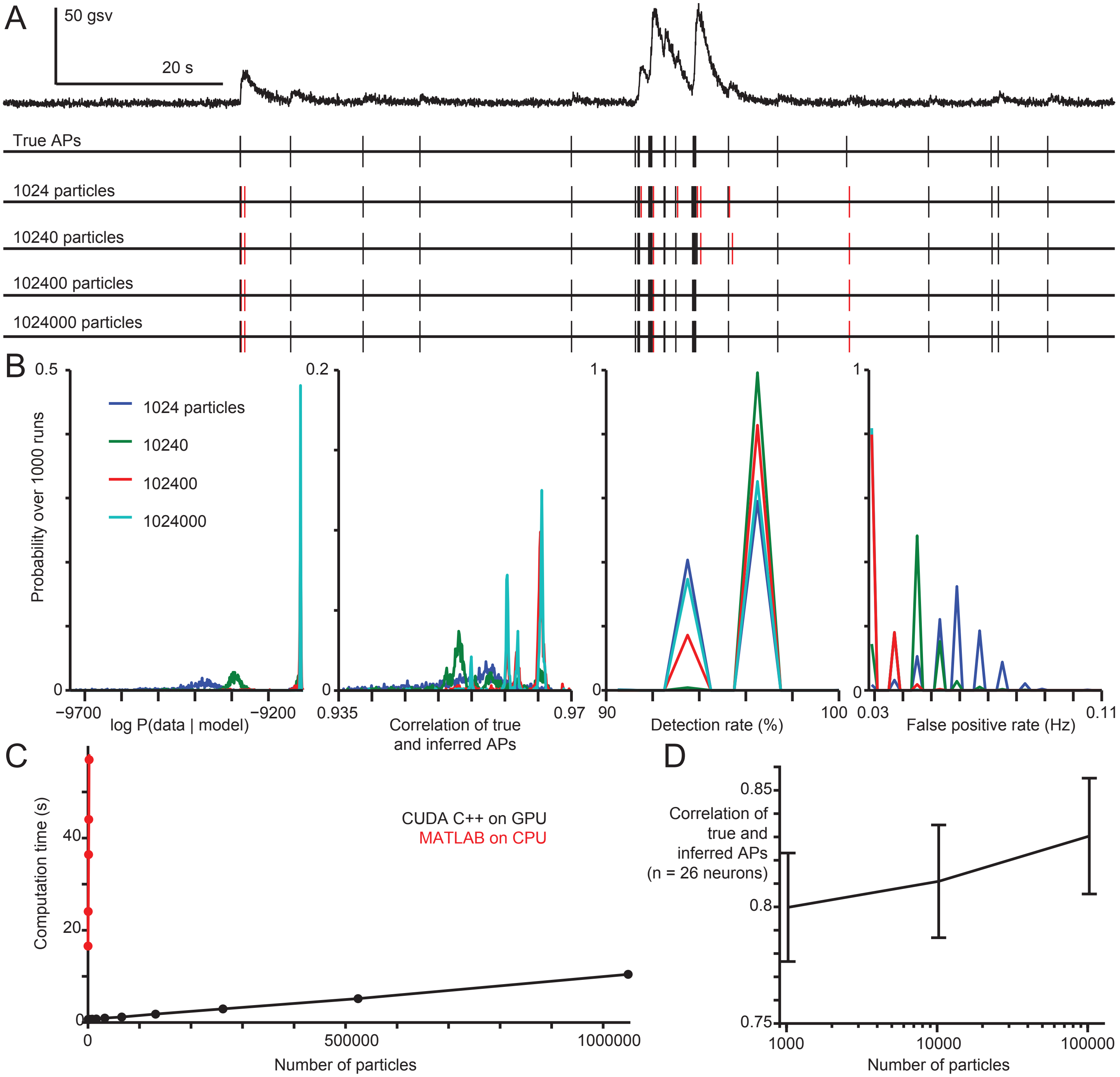

Figure 5–Figure supplement 4. Accuracy and speed of SMC/SBM-based AP inference as a function of particle count. **(A)** 134 seconds of fluorescence signals (upper) and simultaneously recorded APs for a L2/3 mouse visual cortical pyramidal neuron. SBM/SMC-based AP inference results are shown for a single run of the algorithm with 1024, 10240, 102400 and 1024000 particles. False positives produced by the algorithm (time difference from true APs $>$ 200 ms) are shown in red. **(B)** Probability distributions over 1000 runs for the marginal data likelihood (see methods), correlation of true and inferred AP (smoothing σ 200 ms), detection rate and false positive rate (timing tolerance 200 ms). Results are shown for the same particle counts and data as in (A). **(C)** Computation time for a single SMC forward pass with calculation of posterior probability of AP discharge every 10 ms, as a function of the number of particles and for the same data as in (a-b). Results are shown for a CUDA-C++ implementation running on a Geforce GTX 1080 TI (nVidia) or a Matlab implementation running on an Ryzen 7 1800X 8 core processor (AMD). **(D)** Correlation of true and inferred APs over our entire *in vivo* dataset (n = 26 neurons) as a function of particle count. Error bars show standard error of the mean over 1000 runs.

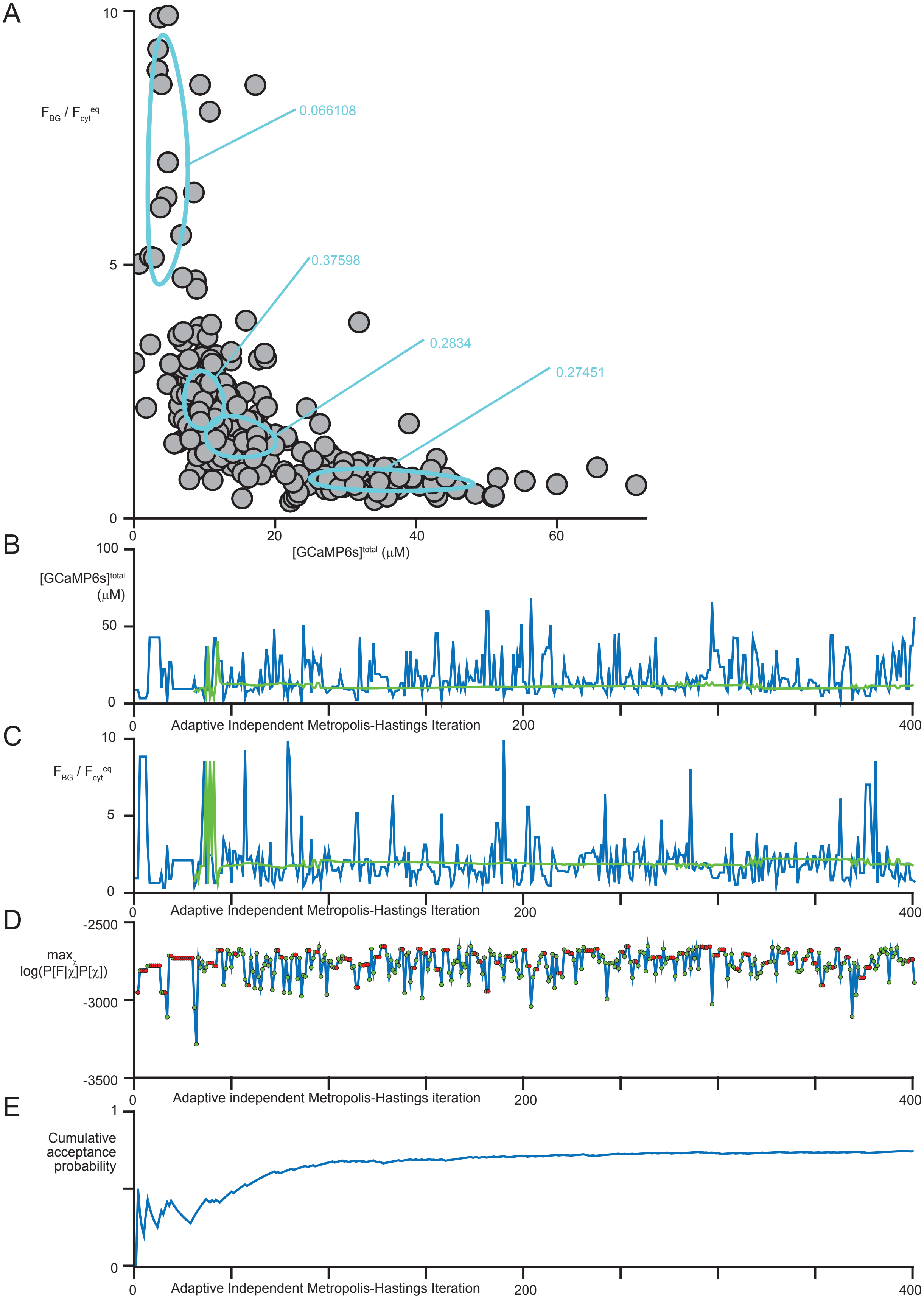

Figure 5–Figure supplement 5. Identification of per-neuron parameters from fluorescence data alone. **(A)** Scatter plot of [GCaMP6s]$^{\text{total}}$ vs. $F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}}$ for adaptive independent Metropolis-Hasting procedure samples (see methods). Contours show 1-s.d. isoclines from each mixture component; numbers indicate component probabilities. **(B)** [GCaMP6s]$^{\text{total}}$ concentration for each sample in (A) (blue) and mode of the current iteration's lognormal mixture fit (green). **(C)** As in (B), but for $F_{\text{BG}}/F_{\text{cyt}}^{\text{eq}}$. **(D)** Log marginal data likelihood times the prior probability, after optimization with respect to the firing rate, for the samples in (A-C). Green dots indicate accepted samples, while red dots indicate rejections with repetition of the previous sample. **(E)** Overall fraction of samples that have been accepted as a function of the number iterations completed.

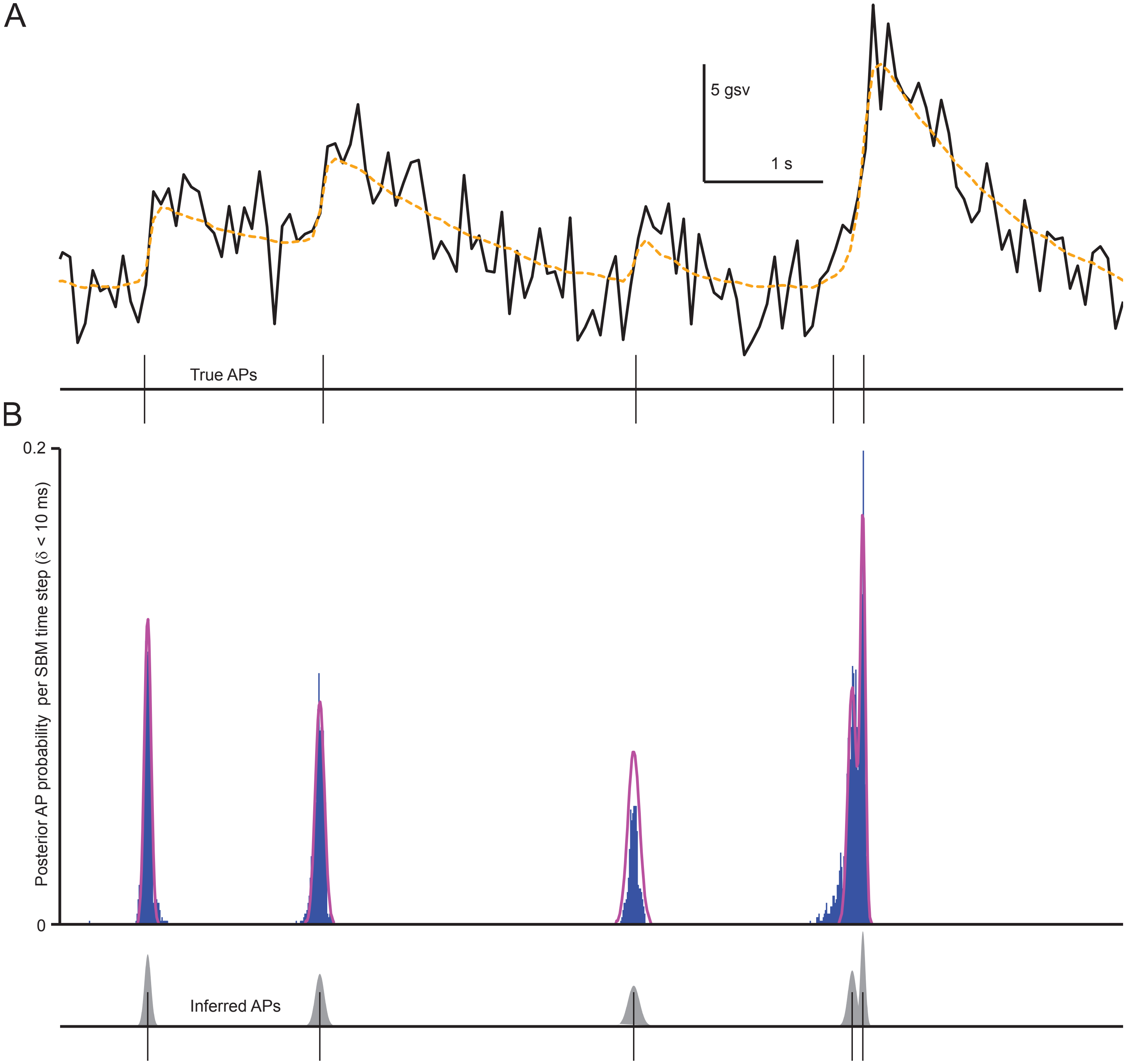

Figure 5–Figure supplement 6. A single AP sequence is fit to posterior probability of AP discharge every time step (δ ≤10 ms) given fluorescence data. **(A)** Neuronal fluorescence (upper, black) with the SBM-inferred posterior mean of denoised fluorescence (orange). Simultaneously recorded APs (lower). **(B)** Posterior probability of AP discharge every 10 ms given the fluorescence data (upper, blue) output by the SMC algorithm. Sum of Gaussian functions (magenta) fit to the posterior AP discharge probabilities. Note that the posterior probability of AP discharge never approaches 1 due to temporal uncertainty, but the total increase associated with each inferred AP sums to around one. By fitting a sum of unweighted Gaussian functions to the posterior probabilities, the SMC output is quantized, so that each period of increase in the posterior is interpreted as a discrete number of APs. The result (lower) is a set of Gaussian means that we interpret as AP times and standard deviations (gray) that we interpret as temporal uncertainties for each AP.

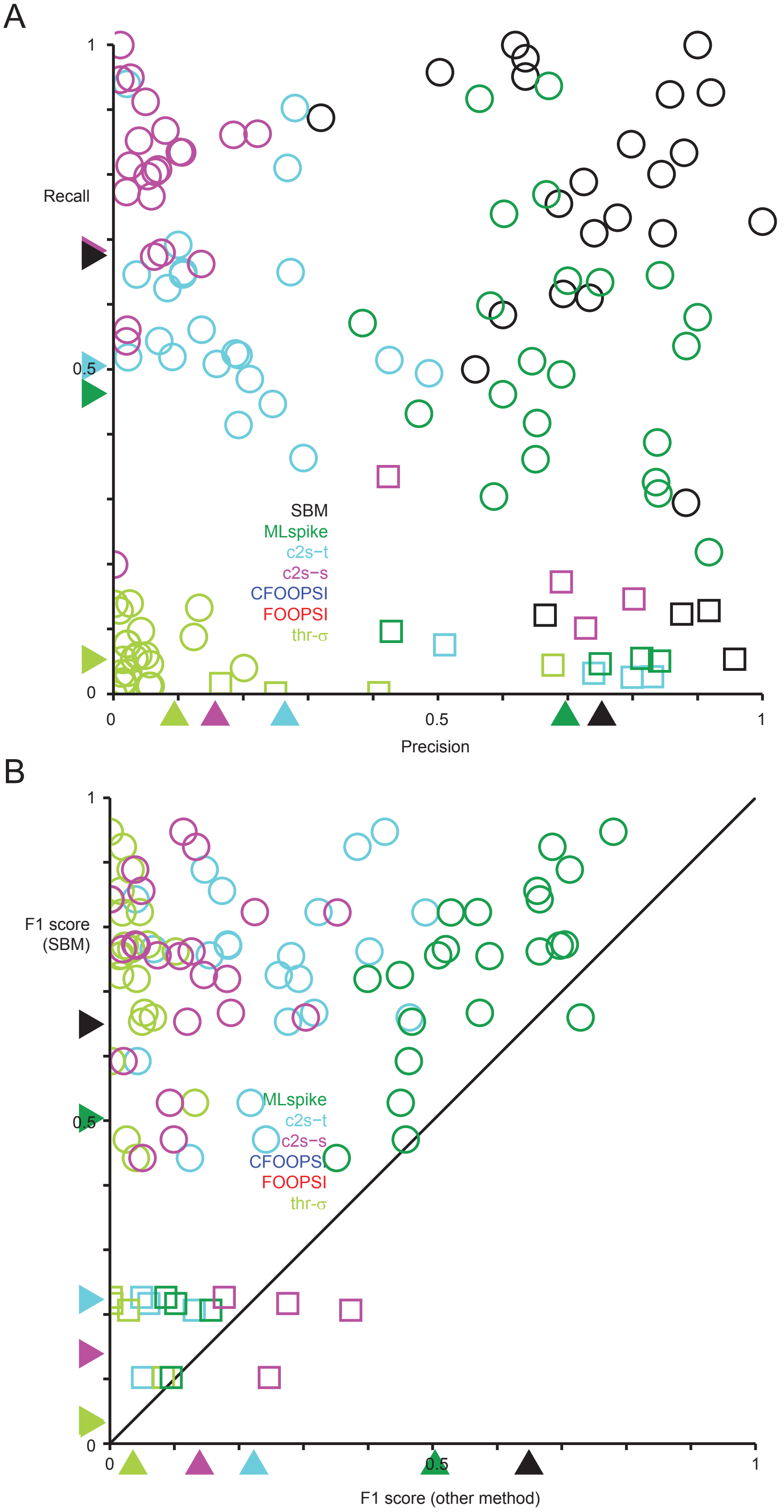

Figure 6–Figure supplement 1. (A) Precision vs. recall for each neuron and AP inference algorithm. Recall is defined as the fraction of true (electrically detected) APs that are detected by the inference algorithm, and is referred to as detection rate in the main text. Precision is defined as the fraction of inferred APs that actually occurred. Both precision and recall have been computed within a maximum allowed time difference of 100 ms when matching true and inferred APs. Squares indicate interneurons. **(B)** Scatter plot comparing the performance of the SBM approach to other algorithms using the F1 score. Squares indicate interneurons. The F1 score is defined as the harmonic rate of recall and precision: $2 \cdot \text{recall} \cdot \text{precision}/(\text{recall} + \text{precision})$.

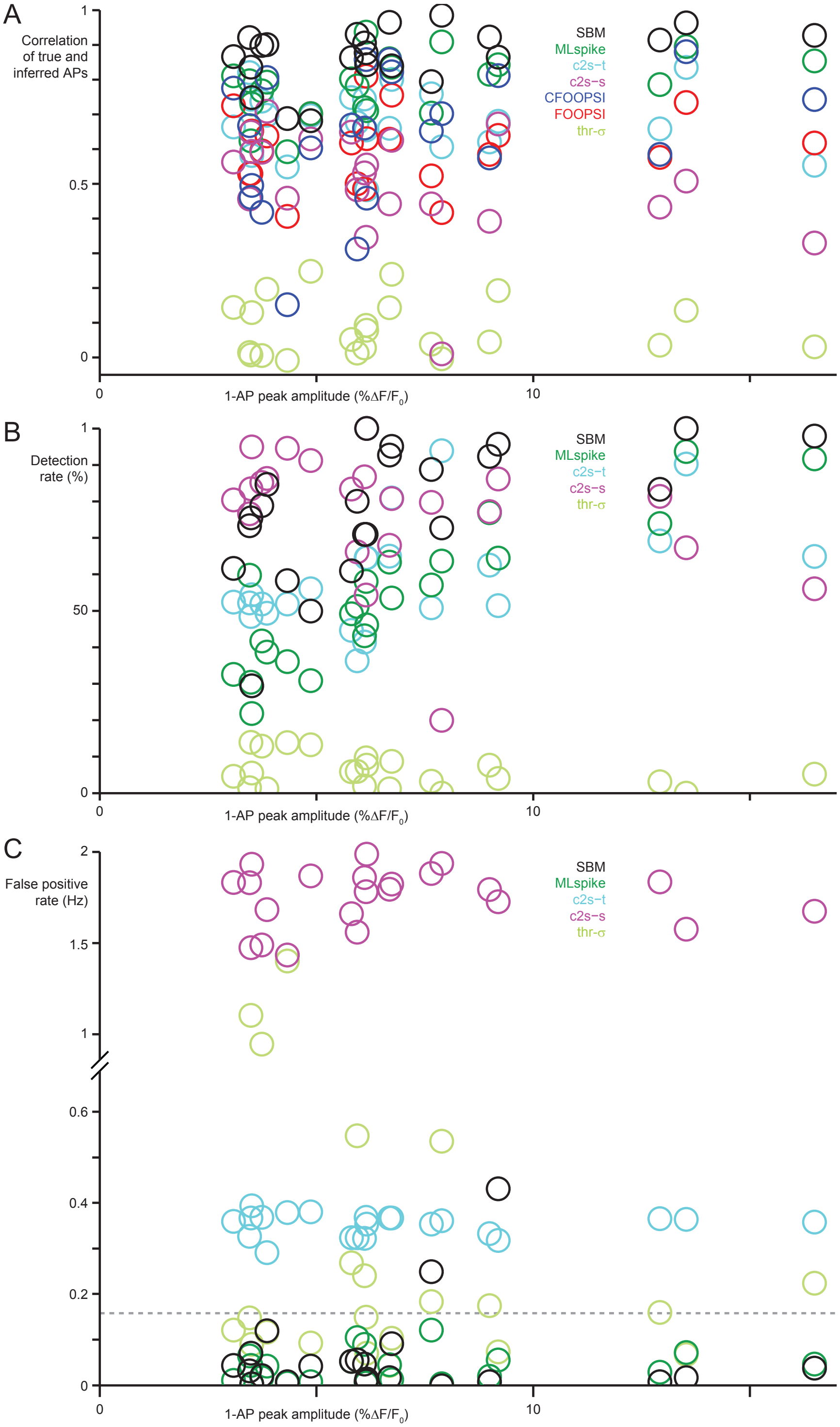

Figure 6–Figure supplement 2. AP inference accuracy as a function of peak 1-AP fluorescence amplitude. **(A)** Peak mean fluorescence evoked by single APs in each pyramidal neuron compared to the correlation between true and inferred AP sequences for each algorithm (n = 22). **(B)** Peak mean fluorescence evoked by single APs for each neuron compared to each algorithm's detection rate. **(C)** Peak mean fluorescence evoked by single APs for each neuron compared to each algorithm's false positive rate. Dashed line indicates median true firing rate.

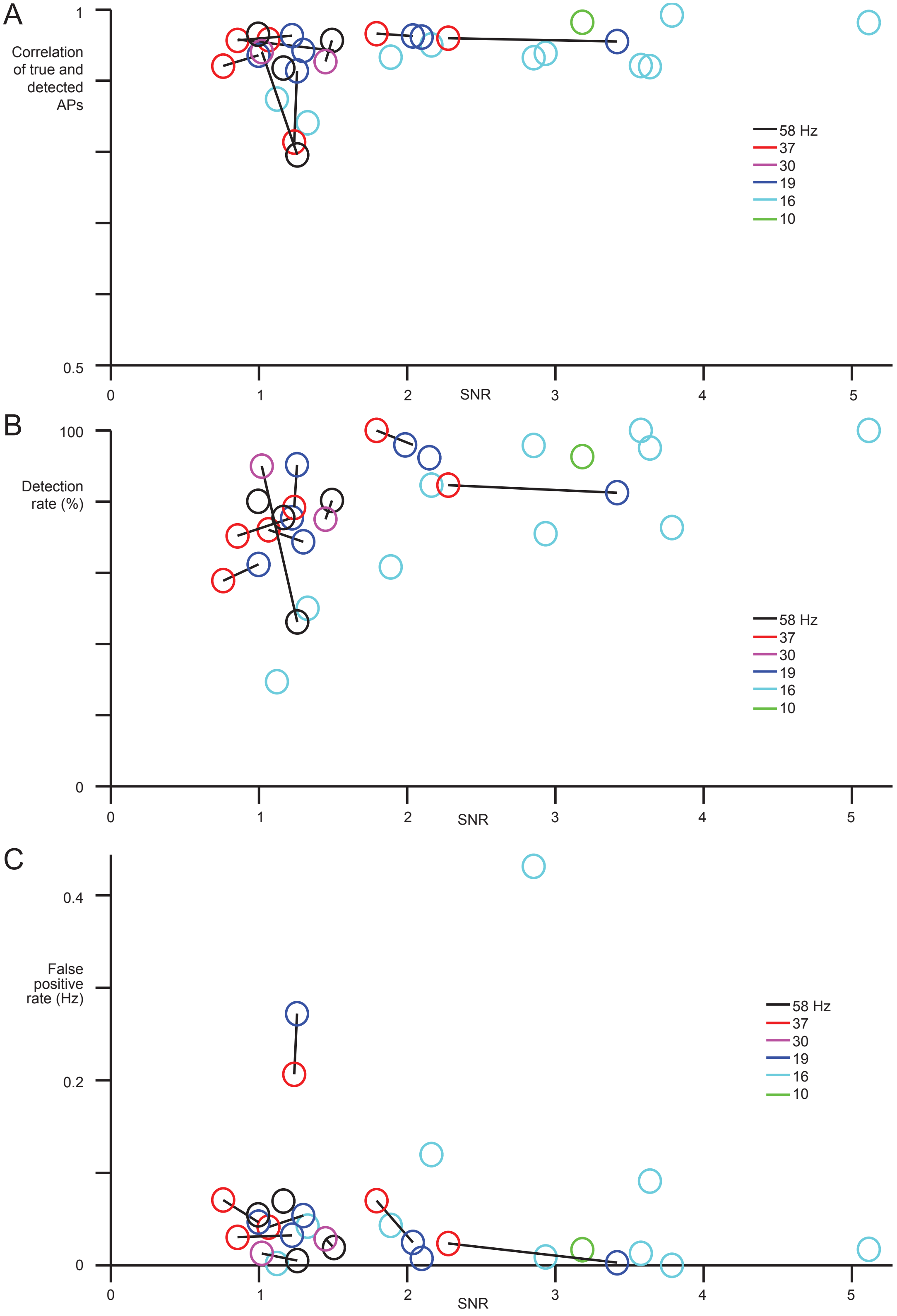

Figure 6–Figure supplement 3. Effect of imaging frame rate and SNR on the accuracy of AP sequences inferred by the SBM. **(A)** Correlation of true and inferred APs as a function of SNR (n = 22 pyramidal neurons). Each color indicates a different imaging frame rate from 10 to 60 Hz; data from the same neuron are connected by line segments. **(B)** As in (A), for the SBM's AP detection rate. **(C)** As in (A-B), for the rate of false positives inferred by the SBM.

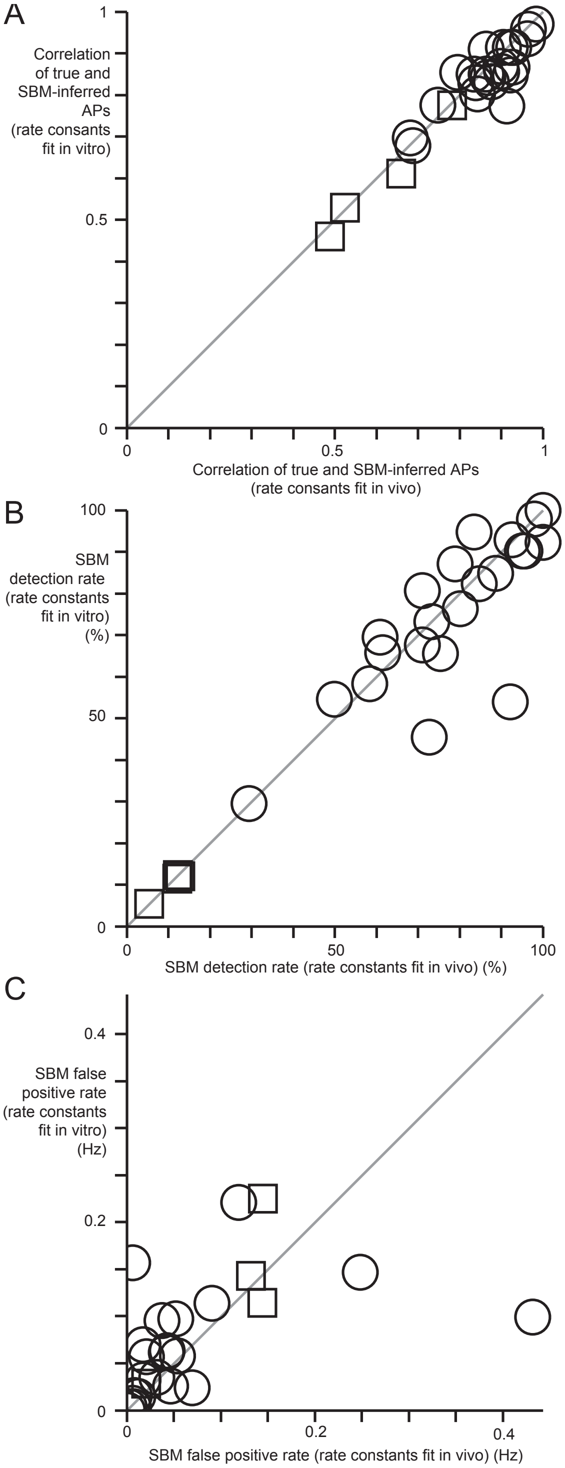

Figure 6–Figure supplement 4. Comparison of SBM-based AP inference accuracy using rate constants fit to *in vivo* data vs. rate constants fit to *in vitro* binding assay data. **(A)** Correlation between true and inferred AP sequences for each neuron for the SBM using rate constants fit from *in vitro* binding assays (x-axis) and rate constants fit from *in vivo* data. Squares indicate interneurons. **(B)** SBM detection rate for rate constants fit *in vitro* vs. *in vivo*. **(C)** SBM false positive rate for rate constants fit *in vitro* vs. *in vivo*.

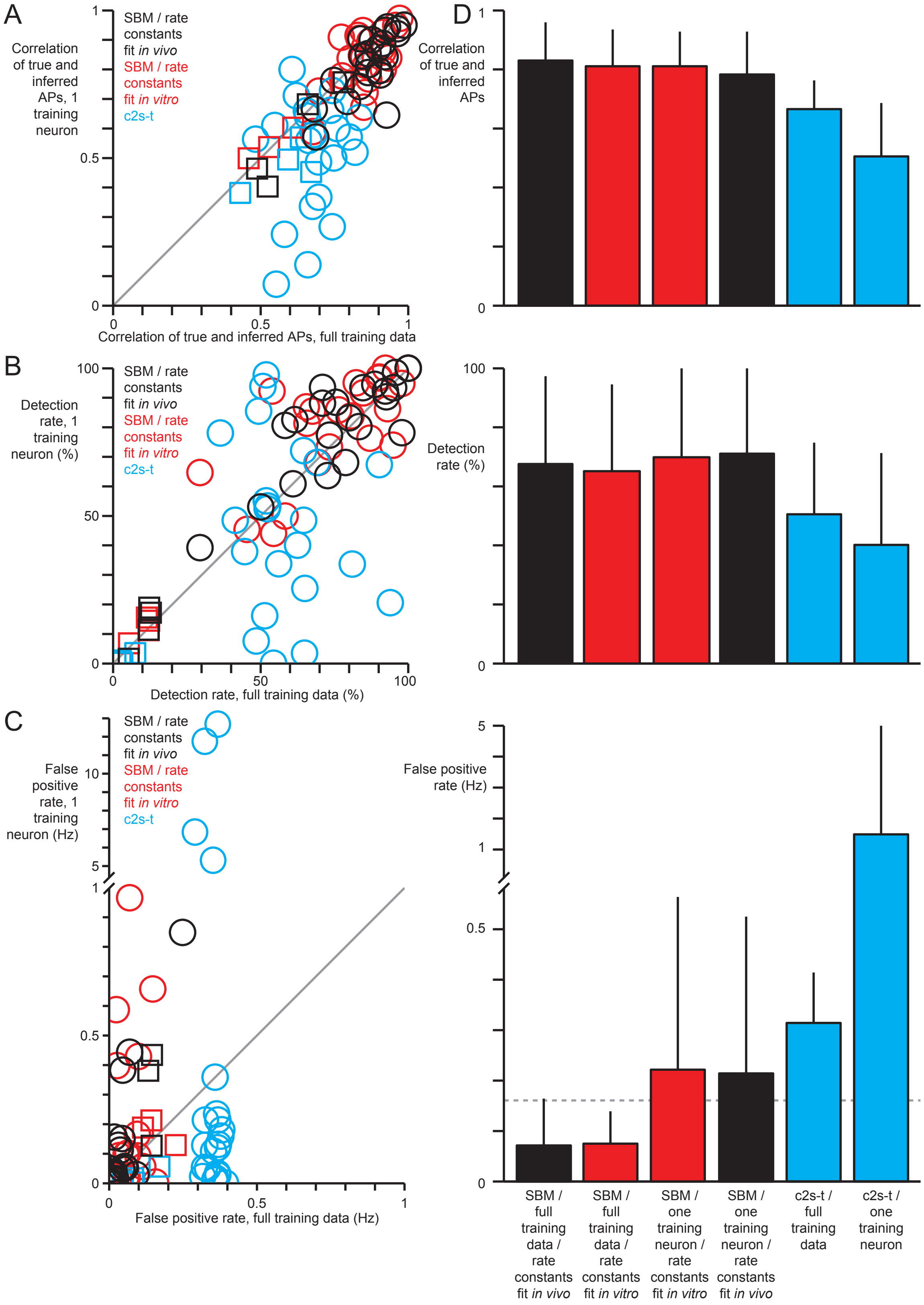

Figure 6–Figure supplement 5. Accuracy of AP inference with one neuron of training data. **(A)** Correlation of true and inferred AP sequences for each neuron (n = 26, squares indicate interneurons) when using full training data (only one neuron at a time held out for cross-validation, x-axis) or a single training neuron (y-axis). Results are shown for the SBM using rate constants fit to *in vivo* data (black), for the SBM using rate constants fit to *in vitro* binding assays (red) and for c2s-t (cyan). **(B)** As in (A), but showing AP detection rates when using training full training data vs. a single training neuron. **(C)** As in (A-B), but showing false positive rates. **(D)** Mean and standard deviations of correlation, detection rate and false positive rate for the data shown in (A-C). Dashed line indicates true median firing rate.

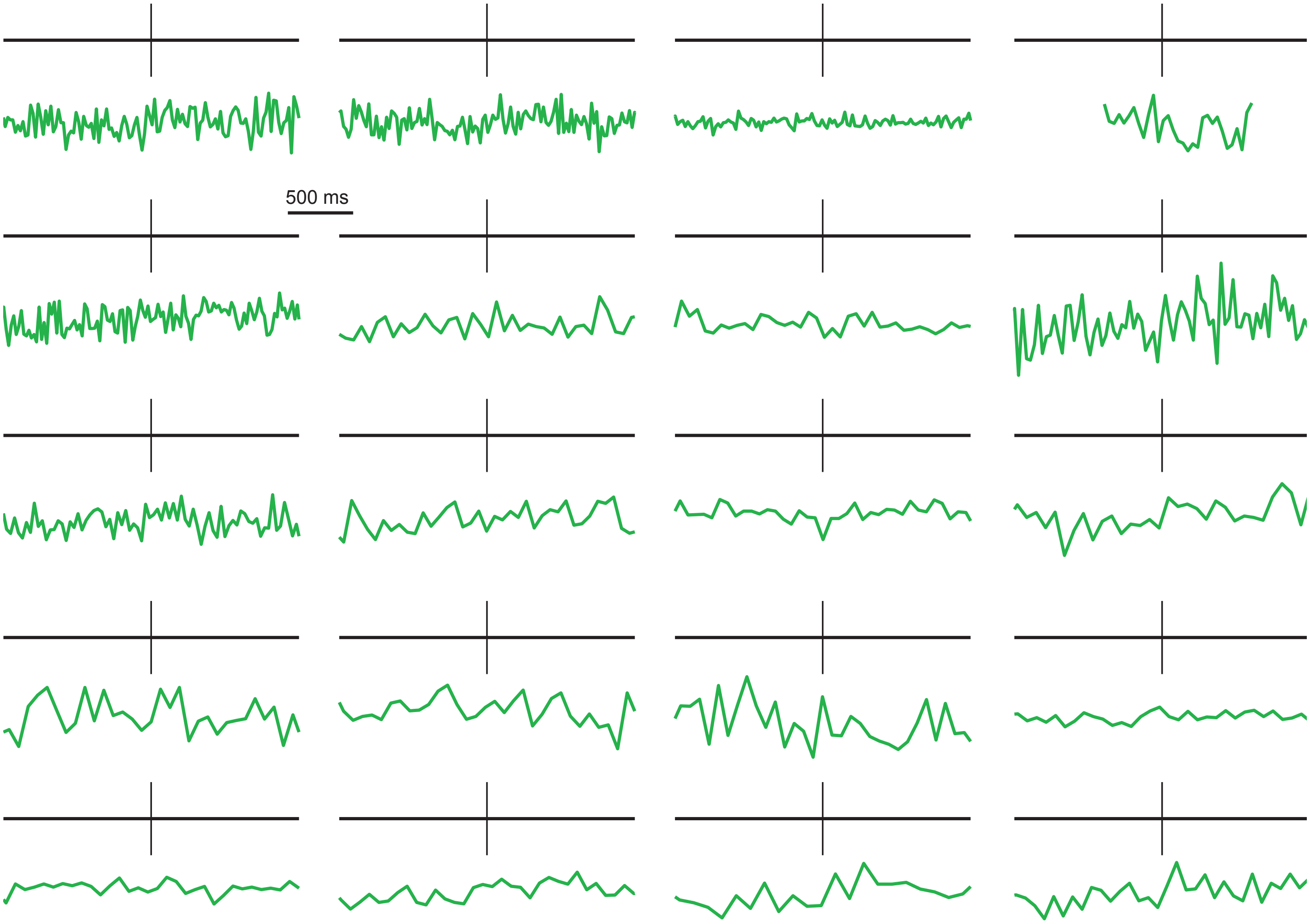

Figure 6–Figure supplement 6. 20 isolated single APs without visually apparent fluorescence increases. Each AP was recorded in a different pyramidal neuron and was chosen based on the lack of accompanying fluorescence increase. In the 2 pyramidal neurons not shown, every isolated single AP evoked a fluorescence increase that could be visually identified.

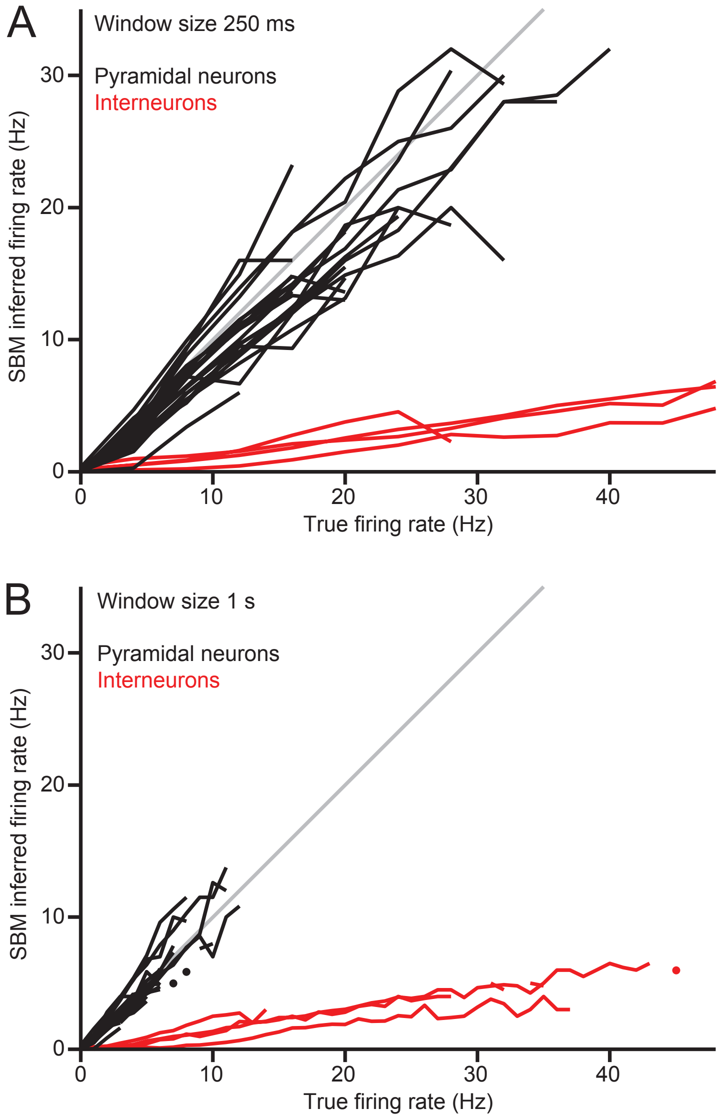

Figure 7–Figure supplement 1. Linearity of SBM-based firing rate inference is robust to the choice of time window size. **(A-B)** As in Figure 7b, but using window sizes of 250 ms and 1s.

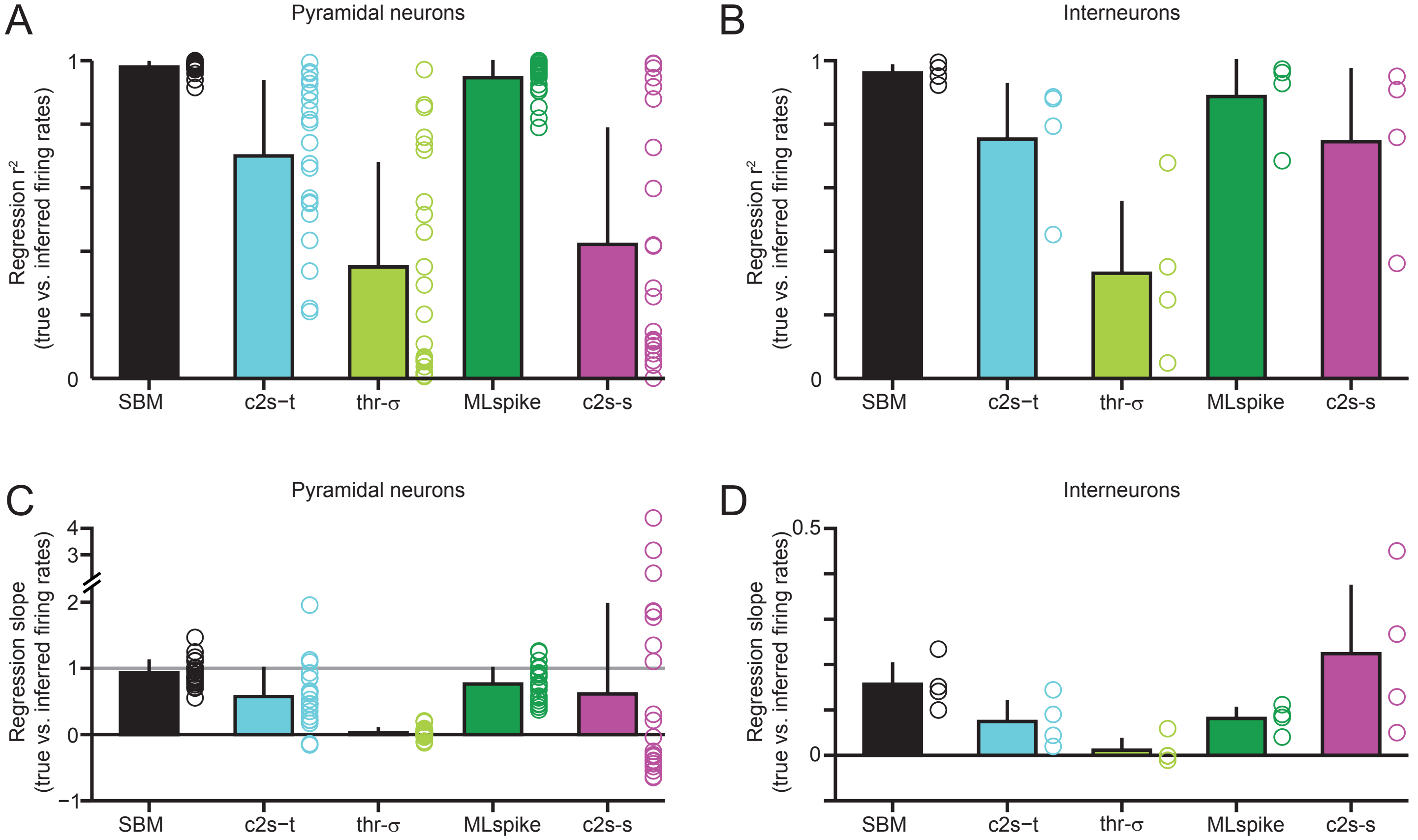

Figure 7–Figure supplement 2. Linearity and slope of inferred firing rate as a function of true firing rate, for all inference methods (500 ms windows). **(A)** Regression r^2 values for inferred firing rate as a function of true firing rate (regressions included non-zero y-intercepts). Each circle represents one pyramidal neuron, bar graph shows mean values and error bars show standard deviations. **(B)** as in (A), but for interneurons. **(C)** Linear regression slopes for inferred firing rate as a function of true firing rate, for all pyramdial neurons and each inference method. Gray line indicates a slope of 1, for which inferred firing rates have the correct units. **(D)** As in (C), but for interneurons.

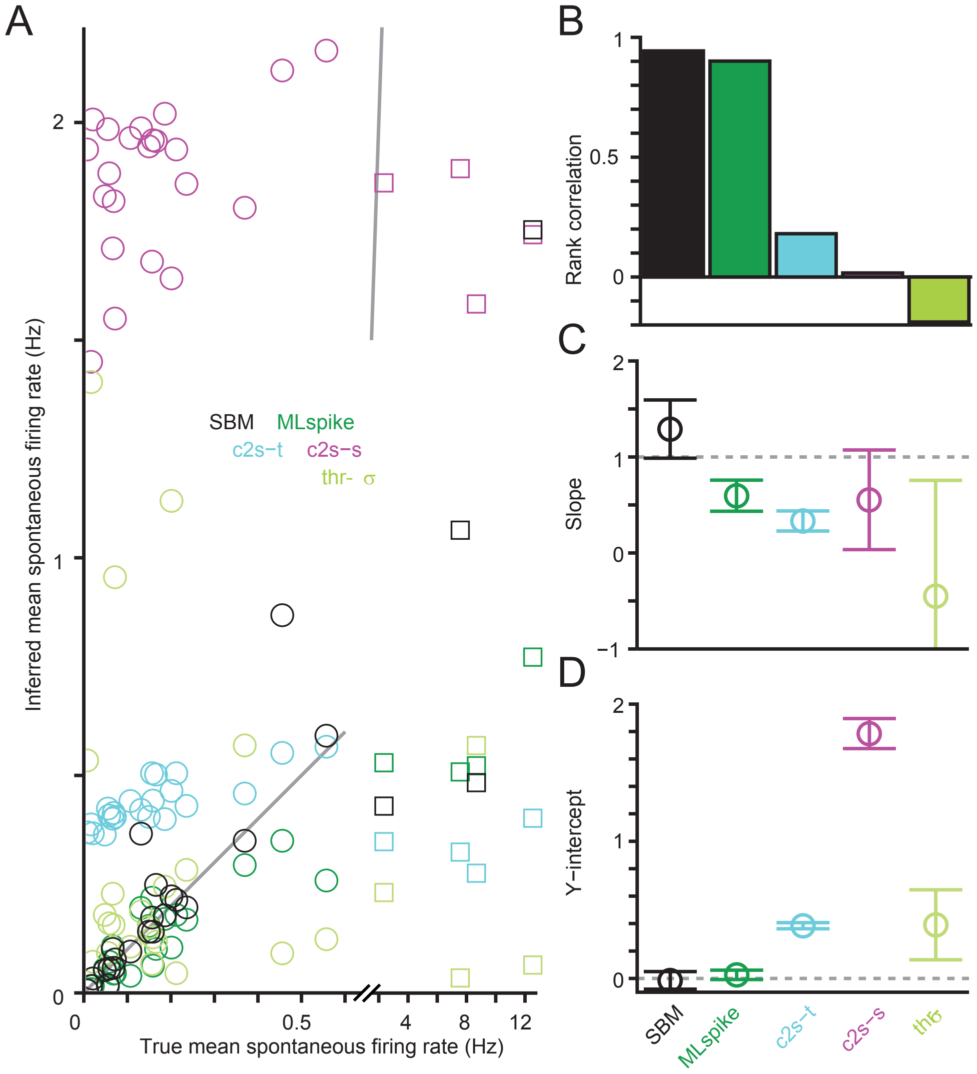

Figure 7–Figure supplement 3. Accuracy of inferred mean spontaneous firing rates. **(A)** True vs. inferred mean spontaneous firing rates for each inference method (n = 26 neurons, squares indicate interneurons). **(B)** Rank correlation of true and estimated firing rates for each method (n = 26 neurons). **(C-D)** Slope and y-intercept of linear regressions for pyramidal neurons (n = 22). Error bars indicate 95% confidence intervals, which include both a slope of 1 and y-intercept of 0 for the SBM, but not for other methods.

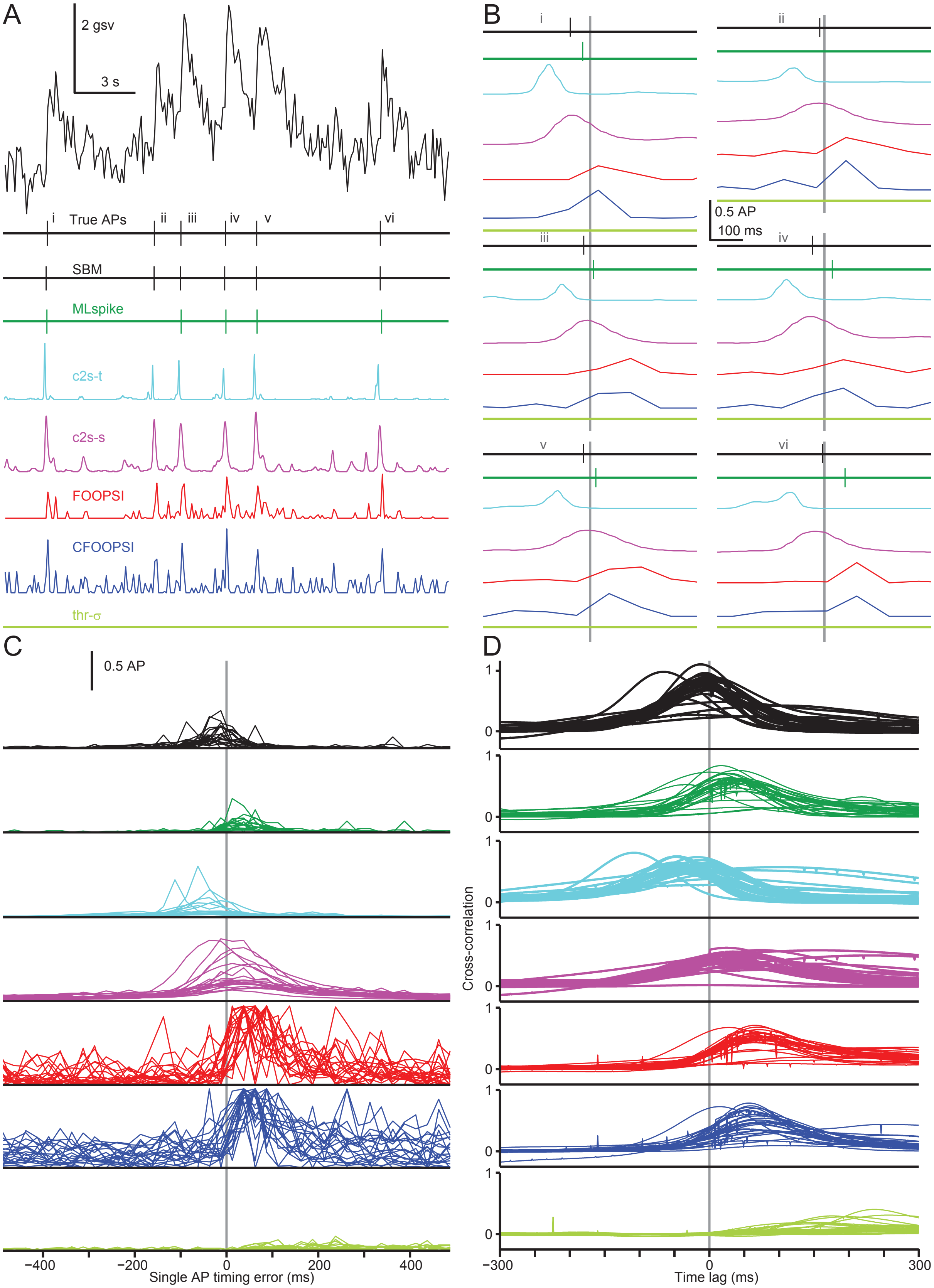

Figure 8–Figure supplement 1. Cross-correlation and single-AP timing accuracy in individual neurons. **(A)** Six electrically recorded single APs and simultaneous fluorescence measurements (upper) with the output of AP inference algorithms (lower). Note that SBM, MLspike and thr-σ output an AP sequence, c2s outputs estimated spiking probabilities over time and CFOOPSI and FOOPSI have unitless outputs. **(B)** Magnified views showing AP inference results relative to electrically detected AP times (gray vertical lines) for the six single APs in (A). **(C)** Mean rates of AP inference relative to true AP times (gray vertical line) for single APs (no other APs for 1 s before and 0.5 s after), for individual neurons. Averaging these results over neurons gives the curves shown in Figure 8G. To be included in this analysis, a neuron's data had to include at least 5 true isolated single APs, and for methods whose outputs were not unitless at least 5 total APs had to be inferred within 0.5 s of these 5 true isolated single APs. The unitless outputs of CFOOPSI and FOOPSI have been rescaled for each neuron to give a maximum average output of 1 for each neuron. **(D)** Cross-correlation between true and inferred APs for each inference method, as a function of time lag, for individual neurons. Averaging these results over neurons gives the curves shown in Figure 8D.

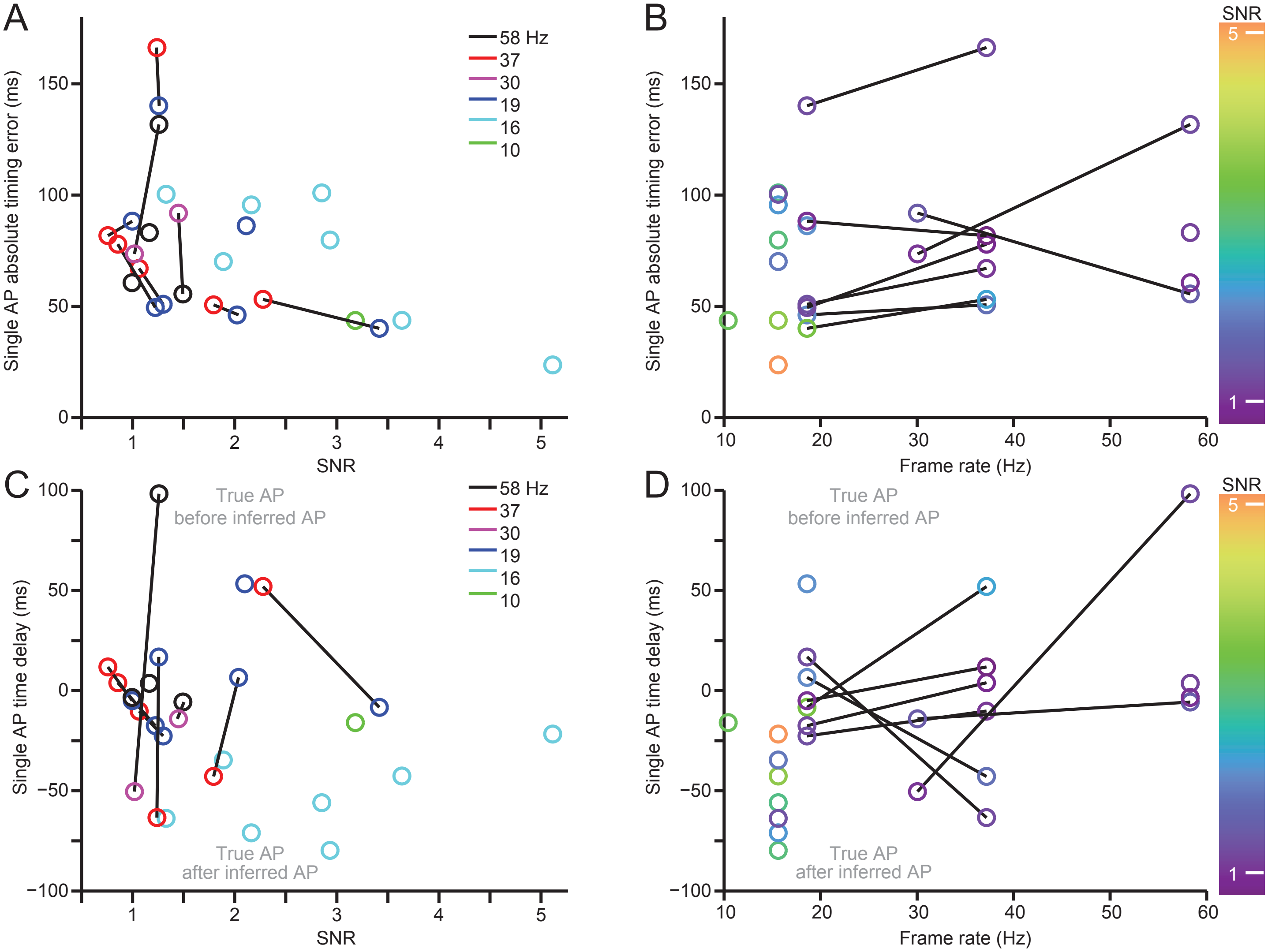

Figure 8–Figure supplement 2. Effect of imaging frame rate and SNR on timing accuracy of APs inferred by the SBM. **(A)** Absolute timing error of isolated single APs inferred by the SBM as a function of SNR. Each color indicates a different imaging frame rate from 10 to 60 Hz; data from the same neuron are connected by line segments. Each data point corresponds to a combination of a specific pyramidal neuron with a specific imaging frame rate, and a combination was included only if it contained at least 5 true isolated single APs with at least 5 APs inferred by the SBM. **(B)** As in (A), but with imaging frame rate shown as the x-coordinate and SNR indicated by color. **(C)** Average delay from true to SBM-inferred times for isolated single APs, as a function of SNR (colors as in (A)). **(D)** As in (C), but with imaging frame rate shown as the x-coordinate and SNR indicated by color.

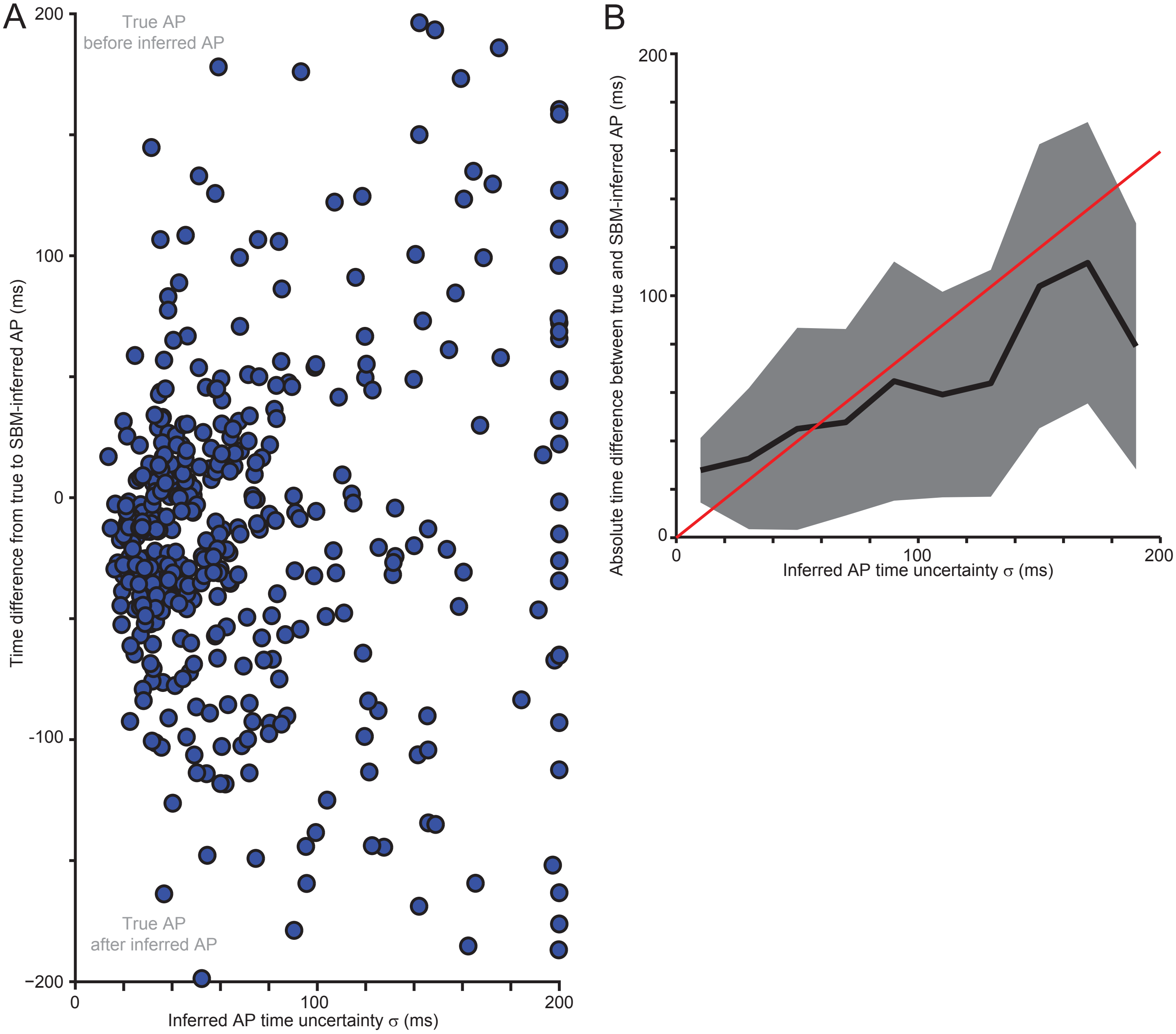

Figure 8–Figure supplement 3. Data-based evaluation of timing uncertainty values inferred by the SBM for isolated single APs. For each AP inferred by the SBM, the algorithm also fits a standard deviation σ to the posterior moments inferred by the particle filter (Figure 5 - figure supplement 6), up to the maximum allowed value of 200 ms. To test whether σ can be validly interpreted as the width of the posterior distribution of the AP time, we compared the inferred σ values to the actual differences between true and inferred AP times. **(A)** Difference between true and inferred AP times as a function of inferred σ values (n = 399 APs from 22 neurons). APs were included in this analysis only if no other APs were present for 5 s before or 500 ms after, the SBM inferred exactly one AP within 200 ms of the true AP time and the SBM inferred no other APs within 500 ms of the true AP time. **(B)** Mean (black) and standard deviation (gray) of the absolute time difference between true and inferred AP times for the data in (A) as a function of the AP timing uncertainty output by the SBM/SMC inference algorithm. The red line shows the relation $y = \sqrt{2/\pi}x$, as would be expected if each σ value correctly describes a Gaussian posterior distribution on an AP time given the fluorescence data.

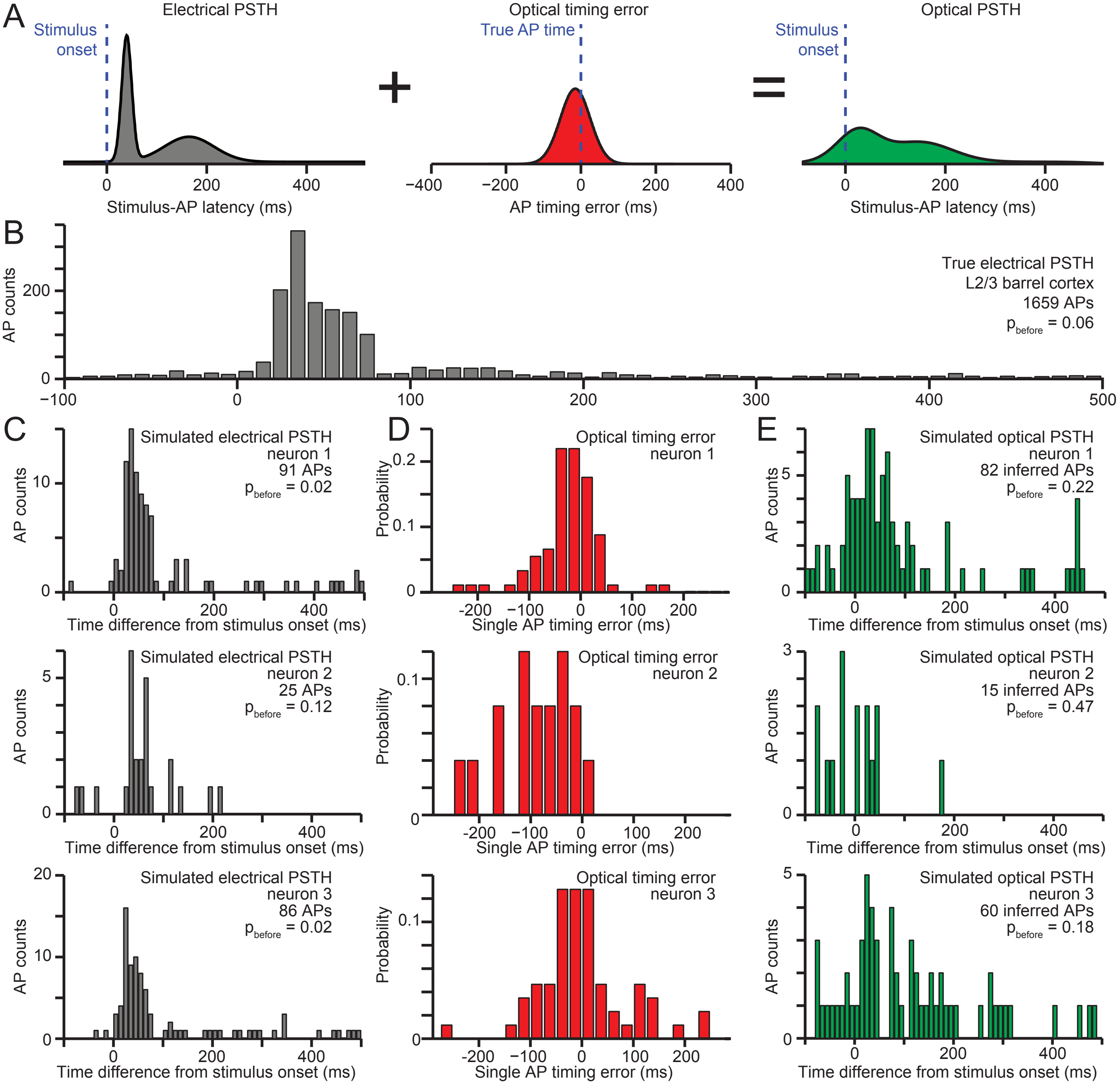

Figure 8–Figure supplement 4. Simulations showing the effect of timing errors in optically detected APs on peri-stimulus time histograms (PSTHs). **(A)** Diagram illustrating simulation of an optical PSTH (right) by adding together stimulus→AP latencies from a true electrical PSTH (left) and timing errors for spontaneous APs detected optically in another neuron (center). **(B)** PSTH showing APs evoked by whisker deflection and recorded electrically in a L2/3 neuron in somatosensory cortex, previously published in [71]. Among APs recorded from 100 ms before the stimulus onset to 500 ms after, the fraction arising from spontaneous activity that occurred before stimulus onset was $p_{\mathrm{before}} = 6\%$. **(C)** Simulated electrical PSTHs for 3 GCaMP6s-expressing L2/3 pyramidal neurons from mouse visual cortex. For each isolated single AP recorded electrically in the neuron, an AP time relative to the simulated stimulus was drawn randomly from the PSTH shown in (B). Due to random sampling of stimulus-AP latencies and a finite number of trials, p_{before} ranges from 2% to 12% in these simulations. **(D)** Distribution of timing errors for isolated single APs, for the 3 neurons in (C). **(E)** Simulated optical PSTHs for the neurons shown in (C-D). For each isolated single AP recorded in the neuron, a true AP time relative to the stimulus was drawn from the PSTH shown in (B), and the results of SBM-based AP inference were used to assign optically detected AP times relative to the true AP time. Since the detection rate for single APs was less than 100%, on some stimulus trials no APs were inferred. Due to AP timing errors for the inferred APs, a greater fraction of APs were assigned to time points before the stimulus for the simulated optical PSTH than for the simulated electrical PSTH, with p_{before} ranging from 13% to 47%.

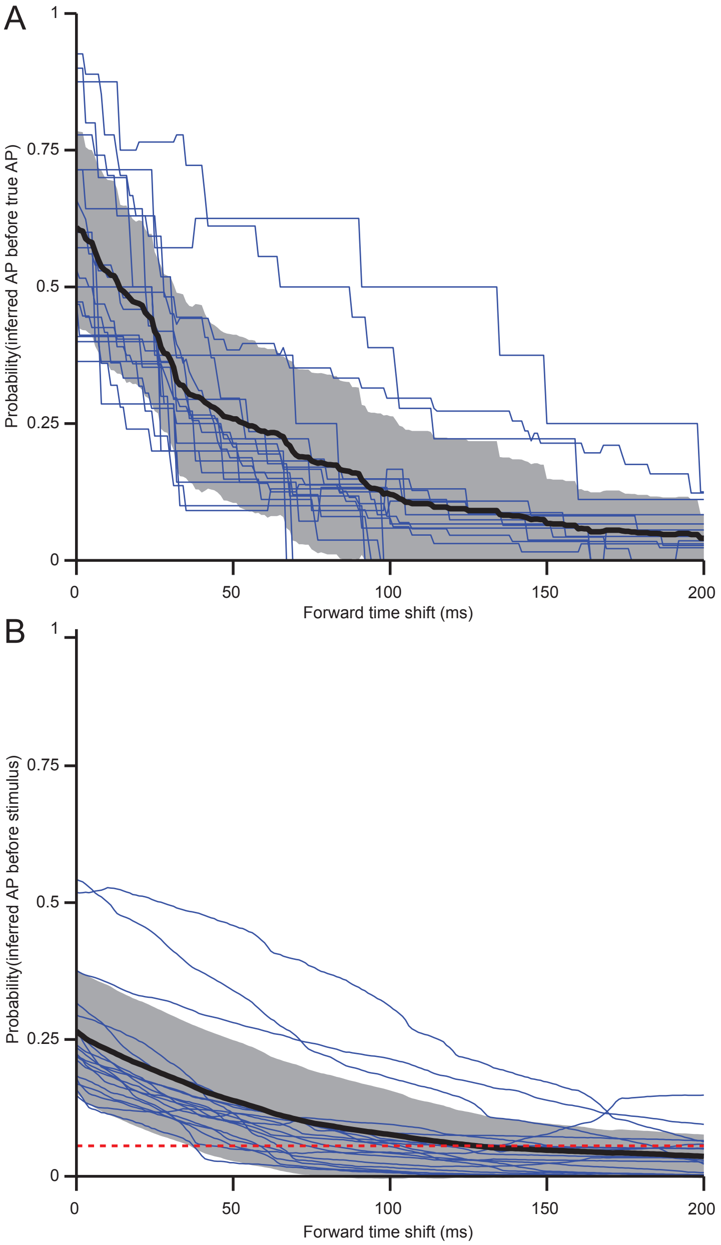

Figure 8–Figure supplement 5. Forward time shifts limit too-early assignment of inferred APs. **(A)** Probability that SBM-inferred APs occur from 0 to 200 ms before true isolated single APs, as a function of the forward time shift applied to all inferred APs. Blue curves shows this relationship for neurons with ≥ 5 true isolated single APs and for which ≥ 5 APs were inferred within 500 ms of true single AP times (n = 18). Black curve and gray shaded region show mean and standard deviation over neurons. To limit the average probability that inferred APs occur before true APs to 0.1, a forward shift of 114 ms is required. **(B)** As in (A), but showing the probability of SBM-inferred APs occurring before a sensory stimulus for a simulated PSTH as in Figure 8 - figure supplement 4. Dashed red line indicates the true probability (0.056) that APs occurred before the stimulus in the electrically recorded PSTH used to carry out the simulations. To obtain the same probability for SBM-inferred APs on average a forward shift of 128 ms is required, while obtaining a probability of 0.1 requires a forward shift of 74 ms.