Bayesian inference of kinetic schemes for ion channels by Kalman filtering

Jan L. Münch[1*], **Fabian Paul**[2], **Ralf Schmauder**[1], **Klaus Benndorf**[1*]

***For correspondence:**
jan.muench@med.uni-jena.de (JM);
klaus.benndorf@med.uni-jena.de (KB)

[1]Institut für Physiologie II, Universitätsklinikum Jena, Friedrich-Schiller-Universität Jena, 07740 Jena, Germany; [2]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637-1454, U.S.A.

Abstract Inferring adequate kinetic schemes for ion channel gating from ensemble currents is a daunting task due to limited information in the data. We address this problem by using a parallelized Bayesian filter to specify hidden Markov models for current and fluorescence data. We demonstrate the flexibility of this algorithm by including different noise distributions. Our generalized Kalman filter outperforms both a classical Kalman filter and a rate equation approach when applied to patch-clamp data exhibiting realistic open-channel noise. The derived generalization also enables inclusion of orthogonal fluorescence data, making unidentifiable parameters identifiable and increasing the accuracy of the parameter estimates by an order of magnitude. By using Bayesian highest credibility volumes, we found that our approach, in contrast to the rate equation approach, yields a realistic quantification of uncertainty. Furthermore, the Bayesian filter delivers unbiased estimates for a wider range of data quality and identifies parameters which the rate equation approach does not identify. These results also demonstrate the power of assessing the validity of algorithms by Bayesian credibility volumes in general. Finally, we show that our Bayesian filter is more robust against errors induced by either analog filtering before analog-to-digital conversion or by limited time resolution of fluorescence data than a rate equation approach.

Introduction

Ion channels are essential proteins for the homeostasis of an organism. Disturbance of their function by mutations often causes severe diseases, such as epilepsy ***Oyrer et al.*** (***2018***); ***Goldschen-Ohm et al.*** (***2010***), sudden cardiac death ***Clancy and Rudy*** (***2001***) or sick sinus syndrome ***Verkerk and Wilders*** (***2014***) indicating a medical need ***Goldschen-Ohm et al.*** (***2010***) to gain further insight into the biophysics of ion channels. The gating of ion channels is usually interpreted by kinetic schemes which are inferred either from macroscopic currents with rate equations (REs) ***Colquhoun and Hawkes*** (***(1995***) or from single-channel currents using dwell time distributions ***Neher and Sakmann*** (***1976***); ***Colquhoun et al.*** (***1981***); ***Horn and Lange*** (***1983***); ***Epstein et al.*** (***2016***); ***Siekmann et al.*** (***2016***) or hidden Markov models (HMMs) ***Chung et al.*** (***1990***); ***Fredkin and Rice*** (***1992***); ***Qin et al.*** (***2000***); ***Venkataramanan and Sigworth*** (***2002***). A HMM consists of a discrete set of metastable states. Changes of their occupation occur as random events over time. Each state is characterized by transition probabilities, related to transition rates, and a probability distribution of the observed signal ***Rabiner*** (***1989***). It is becoming increasingly clear that the use of Bayesian statistics in HMM estimation constitutes a major advantage ***Ball F. G. and A.*** (***1999***); ***de Gunst et al.*** (***2001***); ***Rosales***

et al. (**2001**); **Rosales** (**2004**); **Gin et al.** (**2009**); **Siekmann et al.** (**2011**, 2012); **Hines et al.** (**2015**). In ensemble patches, simultaneous orthogonal fluorescence measurement of either conformational changes **Zheng and Zagotta** (**2000**); **Taraska and Zagotta** (**2007**); **Taraska et al.** (**2009**); **Bruening-Wright et al.** (**2007**); **Kalstrup and Blunck** (**2013**, 2018); **Wulf and Pless** (**2018**) or ligand binding itself **Biskup et al.** (**2007**); **Kusch et al.** (**2010**, 2011); **Wu et al.** (**2011**) has increased insight into the complexity of channel activation.

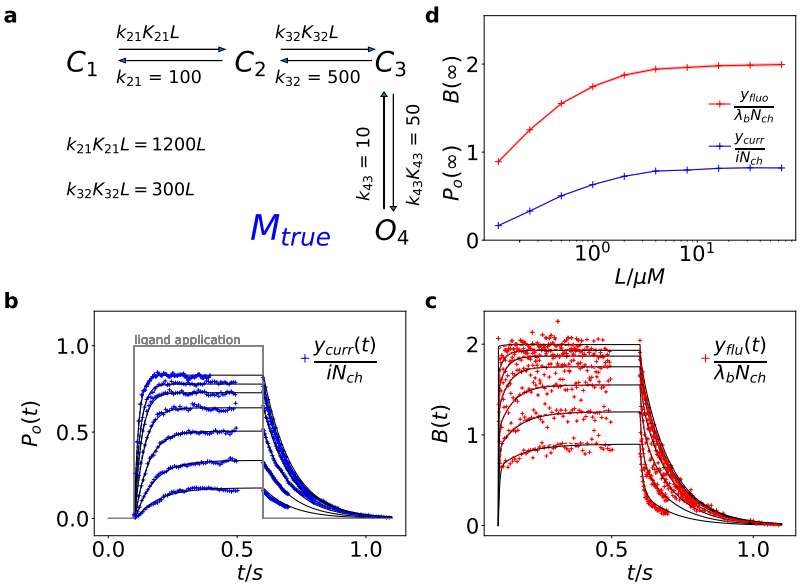


Figure 1. HMM used for the simulations. a, The kinetic scheme M_{true} used for simulating data with 10 kHz and 100 kHz sampling. The Markov state model (kinetic scheme) consists of two binding steps and one opening step. The rate matrix $\mathbf{K}$ is parametrized by the absolute rates k_{ij}, the ratios K_{ij} between on and off rates (i.e. equilibrium constants) and L, the ligand concentration in the solution. The units of the rates are s^{-1} and $\mu\text{M}^{-1}\text{s}^{-1}$ respectively. The liganded states are C_2, C_3, O_4. The open state O_4 conducts a mean single-channel current $i = 1$. Note, that absolute magnitude of the single channel current is irrelevant regarding this study what matters is its relative magnitude compared with σ_{op} and σ_{ex}. The 100 kHz data is used for investigating two related questions regarding the influence on the parameter inference: On the one hand we use 100 kHz data to mimic an analog current signal to investigate the effect of Bessel-filtering before the analog-to-digital conversion. On the other hand, we explore the influence of the finite integration time of the fluorescence signal to create one digital data point. Note, throughout this study the KF analysis frequency f_{ana} for cPCF data is in the range of $(200 - 500)$ Hz while pure current data is analyzed at $(2 - 5)$ kHz. **b-c**, Normalized time traces of simulated relaxation experiments of ligand concentration jumps with $N_{\text{ch}} = 10^3$ channels, $\lambda_{\text{b}} = 0.375$ mean photons per bound ligand per frame and single-channel current $i = 1$. The current y_{curr} and fluorescence y_{flu} time courses are calculated from the same simulation run to mimic the experiment. For visualization, the signals are normalized by the respective median estimates of the KF. The fluctuation of the current traces is due to gating noise, instrumental noise with the variance $\sigma_{\text{m}}^2 = i^2$ and open-channel noise with $\sigma_{\text{op}}^2 = 0.1i^2$. The fluctuation of fluorescence is caused by stochastic binding and Poisson counting noise of photons. The black lines are the theoretical open probabilities $P_{\text{o}}(t)$ and the average binding per channel $B(t)$ for $N_{\text{ch}} \to \infty$ of the used model. Typically we used 10 ligand concentrations which are $(0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 64)$ μM. **d**, Equilibrium binding and open probability as function of the ligand concentration L.

Currently, a Bayesian estimator that can collect information from cross-correlations and time correlations inherent in multi-dimensional signals of ensembles of ion channels is still missing. Traditionally, macroscopic currents are analyzed with solutions of REs which yield a point estimate of the rate matrix or its eigenvalues **Colquhoun et al.** (**1997**); **Sakmann** (**2013**); **Alcantara et al.** (**2002**); **Milescu et al.** (**2005**); **Wang et al.** (**2012**) if they are fitted to the data. The RE approach is based on a deterministic differential equation derived by averaging the chemical master equation (CME) for the underlying kinetic scheme **Kurtz** (**1972**); **Van Kampen** (**1992**); **Jahnke and Huisinga** (**2007a**). Its accuracy can be improved by processing the information contained in the intrinsic noise (stochas-

tic gating and binding) ***Milescu et al.*** (***2005***); ***Munsky et al.*** (***2009***). Nevertheless, all deterministic approaches do not use the information of the time- and cross-correlations of the intrinsic noise. These deterministic approaches are asymptotically valid for an infinite number of channels. Thus, a time trace with a finite number of channels contains, strictly speaking, only one independent data point.

Previous rigorous attempts to incorporate the autocorrelation of the intrinsic noise of current data into the estimation ***Celentano and Hawkes*** (***2004***) suffer from cubic computational complexity ***Stepanyuk and Borisyuk*** (***2011***) in the amount of data points, rendering the algorithm non-optimal or even impractical for a Bayesian analysis of larger data set. To understand this, note, that a maximum likelihood optimization (ML) usually takes several orders of magnitude fewer likelihood evaluations to converge compared to the number of posterior evaluations when one samples the posterior. One Monte Carlo iteration ***Betancourt*** (***2017***) evaluates the posterior distribution and its derivatives many times to propose one sample from the posterior. Stepanyuk suggested an algorithm ***Stepanyuk and Borisyuk*** (***2011***); ***Stepanyuk et al.*** (***2014***) which derives from the algorithm of ***Celentano and Hawkes*** (***2004***) but evaluates the likelihood quicker. Under certain conditions Stepanyuk's algorithm can be faster than the Kalman filter ***Moffatt*** (***2007***). The algorithm by ***Milescu et al.*** (***2005***) achieves its superior computation time efficiency at the cost of ignoring the time correlations of the fluctuations.

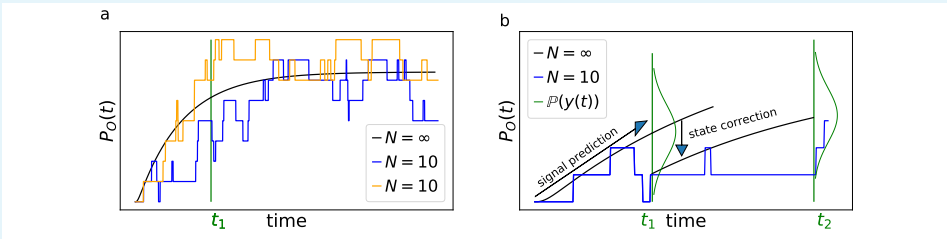
A further argument for our approach, independent from the Bayesian context, is investigated in this paper: The KF is the minimal variance filter ***Anderson and Moore*** (***2012***). Thus instead of excessive analog filtering of currents with the inevitable signal distortions ***Silberberg and Magleby*** (***1993***) one can apply the KF with higher analysing frequency on less filtered data.

On the one hand a complete HMM analysis would deliver the most exact likelihood of macroscopic data. On the other hand the computational complexity of HMMs limits this type of analysis in ensemble patches to no more than a few hundred channels per time trace ***Moffatt*** (***2007***). To tame the computational complexity ***Jahnke and Huisinga*** (***2007b***), we approximate the solution of the CME with a Kalman filter (KF), thereby remaining in a stochastic framework ***Kalman*** (***1960***). This allows us to explicitly model the time evolution of the first two moments (mean value and covariance matrix) of the probability distribution of the hidden channel states. Notably, for linear (first or pseudo first order) Gaussian system dynamics the KF is optimal in producing a minimal prediction error for the mean state. Our approach generalizes the work of Moffat ***Moffatt*** (***2007***) by including state-dependent fluctuations such as open-channel noise and Poisson noise in additional fluorescence data. A central technical difficulty which we solved is that due to the state-dependent noise the central Bayesian update equation loses its analytical solution. We derived an approximation which is correct for the first two moments of the probability distributions.

Stochastic rather than deterministic modeling is generally preferable for small systems or nonlinear dynamics ***Van Kampen*** (***1992***); ***Gillespie and Golightly*** (***2012***). However, even with simulated data of unrealistic high numbers of channels per patch, the KF outperforms the deterministic approach in estimating the model parameters. Moffatt ***Moffatt*** (***2007***) already demonstrated the advantage of the KF to learn absolute rates from time traces at equilibrium. Like all algorithms that estimate the variance and the mean ***Milescu et al.*** (***2005***) the KF can infer the number of channels N_{ch} for each time trace, the single-channel current i and analogous in optical recordings the mean number λ_{b} of photons from bound ligands per recorded frame.

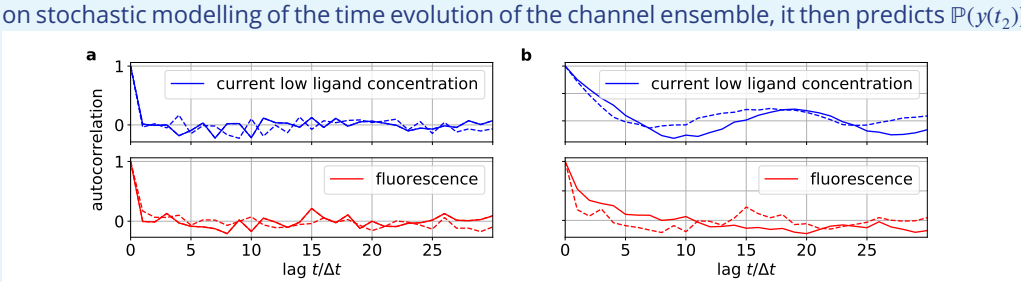
To select models and to identify parameters, stochastic models are formulated within the framework of Bayesian statistics where parameters are assigned uncertainties by treating them as random variables ***Hines*** (***2015***); ***Ball*** (***2016***). In contrast, previous work on ensemble currents combined the KF only with ML estimation ***Moffatt*** (***2007***). Difficulties in treating simple stochastic models by ML approaches in combination with the KF ***Auger-Methe Marie et al.*** (***2016***), especially with non-observable dynamics, justify the computational burden of Bayesian statistics. Bayesian statistics has an intuitive way to incorporate soft or hard constrains from diverse sources of prior information. Those sources include mathematical prerequisites, other experiments, simulations or the-

Box 1. Phenomenological difference between an RE approach and our Bayesian filter



Box 1 Figure 1. a, Idealized patch-clamp (PC) data in the absence of instrumental noise for either ten (colored) or an infinite number of channels generating the mean time trace (black). The fluctuations from the mean time trace (black) reveal autocorrelation **b,** Conceptual idea of the Kalman Filter (KF): the stochastic evolution of the ensemble signal is predicted and the prediction model updated recursively.

Two major problems for parameter inference for the dynamics of the ion channel ensemble $\mathbf{n}(t)$ are: **(I)** that currents are only low-dimensional observations (e.g. one dimension for patch-clamp or two for cPCF) of a high-dimensional process (dimension being the number of model states) blurred by noise and **(II)** the fluctuations due to the stochastic gating and binding process cause autocorrelation in the signal. Traditional analyses for macroscopic PC data (and also for related fluorescence data) by the RE approach, e.g. ***Milescu et al.*** (***2005***) ignores the long-lasting autocorrelations of the deviations (Box 1 Fig. 1 **a**) blue and orange curves from the mean time trace (black) that occur in real data measured from a *finite* ensemble. To account for the autocorrelation in the signal, an optimal prediction (Box 1 Fig. 1 **b**) of the future signal distribution $\mathbb{P}(y(t_2))$ should use the measurement $y(t_1)$ from the current time step t_1 to update the belief about the underlying $\mathbf{n}(t_1)$. Based on stochastic modelling of the time evolution of the channel ensemble, it then predicts $\mathbb{P}(y(t_2))$.



Box 1 Figure 2. a, Autocorrelation of the residuals r_I of two ligand concentrations of currents (blue) and of the fluorescence (red) after the data have been analysed with the KF approach. **b,** Autocorrelation of r_I after analysing with he RE approach.

To demonstrate the difference how the two algorithms analyse the data, we compute the autocorrelation of the residuals of the data. After the analysis with either the RE approach or the KF we can construct from the model with the mean predicted signal $\mathbf{H}\mathbb{E}[\mathbf{n}(t_i)]$ (see, Eq. 4 for the definition of $\mathbf{H}$) and the predicted standard deviation $\sqrt{\mathrm{var}[y(t_i)]}$ the normalized residual time trace of the data which are defined as

$$r(t_i) := \frac{y(t_i) - \left(\mathbf{H}\mathbb{E}[\mathbf{n}(t_i)]\right)}{\sqrt{\mathrm{var}[y(t_i)]}}. \tag{1}$$

Filtering (fitting) with the KF (given the true kinetic scheme) one expects to find a white-noise process for the residuals. Plots of the autocorrelation function of both signal components (Box 2 Fig. 2 **a)**) confirms our expectation. The estimated autocorrelation vanishes after one multiple of the lag time (the interval between sampling points), which means that the residuals are indeed a white-noise process. In contrast, the residuals derived from the RE approach (Box 1 Fig. 2 **a**) display long lasting periodic autocorrelations.

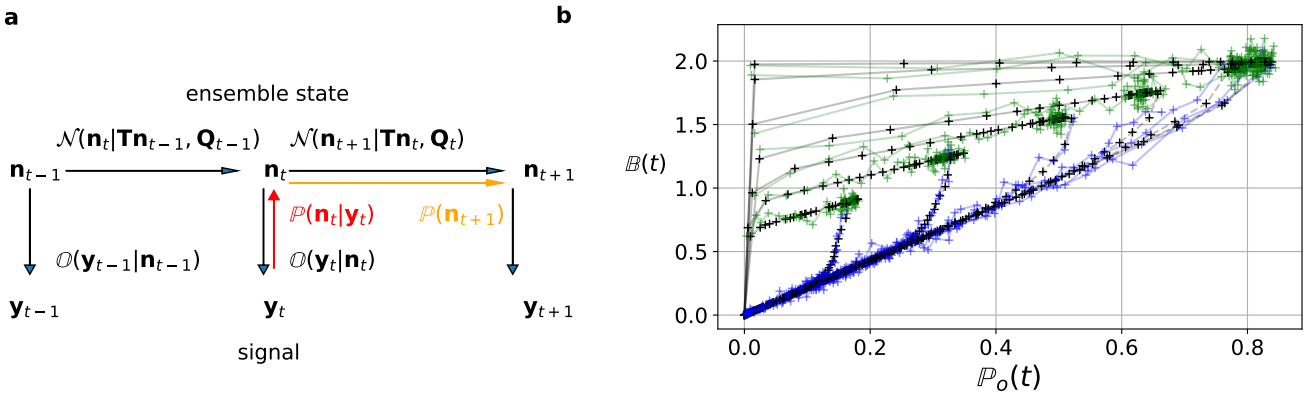


Figure 2. KF as Bayesian filter. a, Graphical model of the conditional dependencies of the stochastic process. Horizontal black arrows represent the conditional multivariate normal transition probability $\mathcal{N}(\mathbf{n}_{t+1}|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t)$ of a continuous state Markov process. Notably, it is $\mathbf{n}(t)$ which is treated as the Markov state by the KF. The transition matrix $\mathbf{T}$ and the time-dependent covariance $\mathbf{Q}_t = \mathbf{Q}(\mathbf{T}, \mathbf{n}_t)$ characterise the single-channel dynamics. The vertical black arrows represent the conditional observation distribution $\mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)$. The observation distribution summarizes the noise of the experiment, which in the KF is assumed to be multivariate normal. Given a set of model parameters and a data point $\mathbf{y}_t$, the Bayesian theorem allows to calculate in the correction step $\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t)$ (red arrow). The posterior is propagated linearly in time by the model, predicting a state distribution $\mathbb{P}(\mathbf{n}_{t+1})$ (orange arrow). The propagated posterior predicts together with the observation distribution the mean and covariance of the next observation. Thus, it creates a multivariate normal likelihood for each data point in the observation space. **b,** Observation space trajectories of the predictions and data of the binding per channel vs. open probability. The curves are normalized by the median estimates of λ_b, i and N_{ch} and the ratio of open-channels $\frac{y_{curr}}{N_{ch} i}$ which approximates the open probability $P_o(t)$. The black crosses represent $\mathbf{H}\mathbf{n}_{t+1}$, the mean of the parameter samples of the predicted signal for each data point of the KF. The green and blue trajectories represent the part of the time traces with a non-zero ligand concentration and a zero ligand concentration in the bulk, respectively.

oretical assumptions. They are applied as additional model assumptions by a prior probability distribution over the possible parameter space. Hence, knowledge of the model parameters prior to the experiment are correctly accounted for in the analyses of the new data. Alternatively, some of these benefits of prior knowledge can be incorporated by penalized maximum likelihood ***Salari et al.*** (***2018***); ***Navarro et al.*** (***2018***). Bayesian inference provides outmatching tools for modeling over point estimates: First, the Bayesian approach is still applicable in situations where parameters are not identifiable ***Hines et al.*** (***2014***); ***Middendorf and Aldrich*** (***2017***) or posteriors are non-Gaussian, whereas ML fitting ceases to be valid ***Calderhead et al.*** (***2013***); ***Watanabe*** (***2007***). Second, a Bayesian approach provides superior model selection tools for singular models such as HMMs or KFs ***Gelman et al.*** (***2014***). Third, Bayesian statistics has a correct uncertainty quantification based on the data and the prior for the statistical problem. In contrast, ML or maximum posterior approaches lack uncertainty quantification based on one data set ***Joshi et al.*** (***2006***). Only under optimal conditions their uncertainty quantification becomes equivalent to Bayesian credibility volumes ***Jaynes and Kempthorne*** (***1976***). This study focuses on the effects on the posterior due to formulating the likelihood via a KF instead of an RE approach and the benefits of adding a second dimension of observation.

Results and Discussion

Simulation of ligand-gated ion-channel data

Here we treat an exemplary ligand-gated channel with two ligand binding steps and one open-closed isomerization described by an HMM (see Fig. 1a). For this model confocal patch-clamp fluorometry (cPCF) data were simulated: time courses of ligand binding and channel current upon concentration jumps were generated, see Appendix 5 and Methods section . Idealized example data with added white noise are shown in Fig. 1b-d. We added realistic instrumental noise to the simulated data (see Appendix 5). A qualitative description of the statistical problem that needs to be addressed when modeling time series data such as the simulated is outlined in Box. 1.

Kalman filter derived from a Bayesian filter

Here and in the Methods section, we derive the mathematical tools to account correctly for the stochastic Markov dynamics of single molecules in the fluctuations of macroscopic signals. The KF is a Bayesian filter (see Methods), i.e. a continuous state HMM with a multivariate normal transition probability ***Ghahramani*** (***1997***) (Fig. 2**a**). We define the hidden ensemble state vector

$$\mathbf{n}(t) := (n_1(t), n_2(t), n_3(t), n_4(t))^\top = \sum_{i=1}^{N_{\mathrm{ch}}} \mathbf{s}_i(t), \tag{2}$$

which counts the number of channels in each state $\mathbf{s}$ (see Methods). To make use of the KF, we assume the following general form of the dynamic model: The evolution of $\mathbf{n}(t)$ is determined by a linear model that is parametrized by the state evolution matrix $\mathbf{T}$

$$\mathbf{n}_{t+1} = \mathbf{T}\mathbf{n}_t + \boldsymbol{\omega}_t \sim \mathcal{N}(\cdot|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t), \tag{3}$$

where $\sim$ means *sampled from* and $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a shorthand for the multivariate normal distribution, with the mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$. The state evolution matrix (transition matrix) is related to the rate matrix $\mathbf{K}$ by the matrix exponential $\mathbf{T} = \exp(\mathbf{K}\Delta t)$. The mean of the hidden state evolves according to the equation $\mathbb{E}[\mathbf{n}_{t+1}|\mathbf{n}_t] = \mathbf{T}\mathbf{n}_t$. It is perturbed by normally-distributed white process noise $\boldsymbol{\omega}$ with the following properties: The mean value of the noise fulfills $\mathbb{E}[\boldsymbol{\omega}_t] = 0$ and the variance-covariance matrix of the noise is $\mathrm{cov}[\boldsymbol{\omega}_t, \boldsymbol{\omega}_t] = \mathbf{Q}(\mathbf{T}, \mathbf{n}_{t-1})$ (see Methods Eq. 38d). In short, Eq. 3 defines a Gaussian Markov process.

The observations $\mathbf{y}_t$ depend linearly on the hidden state $\mathbf{n}_t$. The linear map is determined by an observation matrix $\mathbf{H}$.

$$\mathbf{y}_t = \mathbf{H}\mathbf{n}_t + \boldsymbol{\nu}_t \sim \mathbb{O}(\cdot|\mathbf{H}\mathbf{n}_t) := \mathcal{N}(\cdot|\mathbf{H}\mathbf{n}_t, \boldsymbol{\Sigma}_t) \tag{4}$$

The noise of the measurement setup (Appendix 5 and Eq. 43) is modeled as a random perturbation of the mean observation vector. The noise fulfills $\mathbb{E}[\boldsymbol{\nu}] = 0$ and $\mathrm{cov}[\boldsymbol{\nu}_t, \boldsymbol{\nu}_t] = \boldsymbol{\Sigma}_t$. Eq. 4 defines the state-conditioned observation distribution $\mathbb{O}$ (Fig. 2**a**). If the system strictly obeys Eq. 3 and Eq. 4 then the KF is optimal in the sense that it is the minimum variance filter of that system ***Anderson and Moore*** (***2012***). If the distributions of ν and ω are not normal the KF is still the minimum variance filter in the class of all linear filters but there might be better non-linear filters. In case of colored noise ν and ω the filtering equations (see Methods) can be reformulated by state augmentation or measurement-time-difference approach techniques ***Chang*** (***2014***).

For each element in a sequence of hidden states $\{\mathbf{n}_t : 0 < t < T\}$ and for a fixed set of parameters $\boldsymbol{\theta}$, an algorithm based on a Bayesian filter (Fig. 2**a**), exploits the conditional dependencies of the assumed stochastic process. A Bayesian filter recursively predicts prior distributions for the next $\mathbf{n}_t$

$$\mathbb{P}(\mathbf{n}_t) = \int \mathbb{P}(\mathbf{n}_t|\mathbf{n}_{t-1})\mathbb{P}(\mathbf{n}_{t-1}|\mathbf{y}_{t-1})\,\mathrm{d}\mathbf{n}_{t-1}, \tag{5}$$

given what is known about y at time $t-1$. The KF as a special Bayesian filter assumes that the transition probability is multivariate normal according to Eq. 3

$$\mathbb{P}(\mathbf{n}_t) = \int \mathcal{N}(\mathbf{n}_t|\mathbf{T}\mathbf{n}_{t-1}, \mathbf{Q}_{t-1})\mathbb{P}(\mathbf{n}_{t-1}|\mathbf{y}_{t-1})\,\mathrm{d}\mathbf{n}_{t-1} \tag{6}$$

Note, that Eq. 6 is a central approximation of the KF. While the exact transition distribution of an ensemble of ion channels is the generalized-multinomial distribution (Methods Eq. 32), the quality of normal approximations to multinomial ***Milescu et al.*** (***2005***) or generalized-multinomial ***Moffatt*** (***2007***) distributions depends on the number of ion channels N_{ch} in the patch and on the position of the probability vector in the simplex space. The difference between the log-likelihoods of the true generalized-multinomial dynamics and Eq. 6 type approximation scales as $1/N_{\mathrm{ch}}$ ***Moffatt*** (***2007***). As a rule of thumb one should be careful with both algorithms for time traces with $N_{\mathrm{ch}} \in [10^1, 10^2]$. Below or even inside this interval there are more qualified concepts such as the forward algorithm

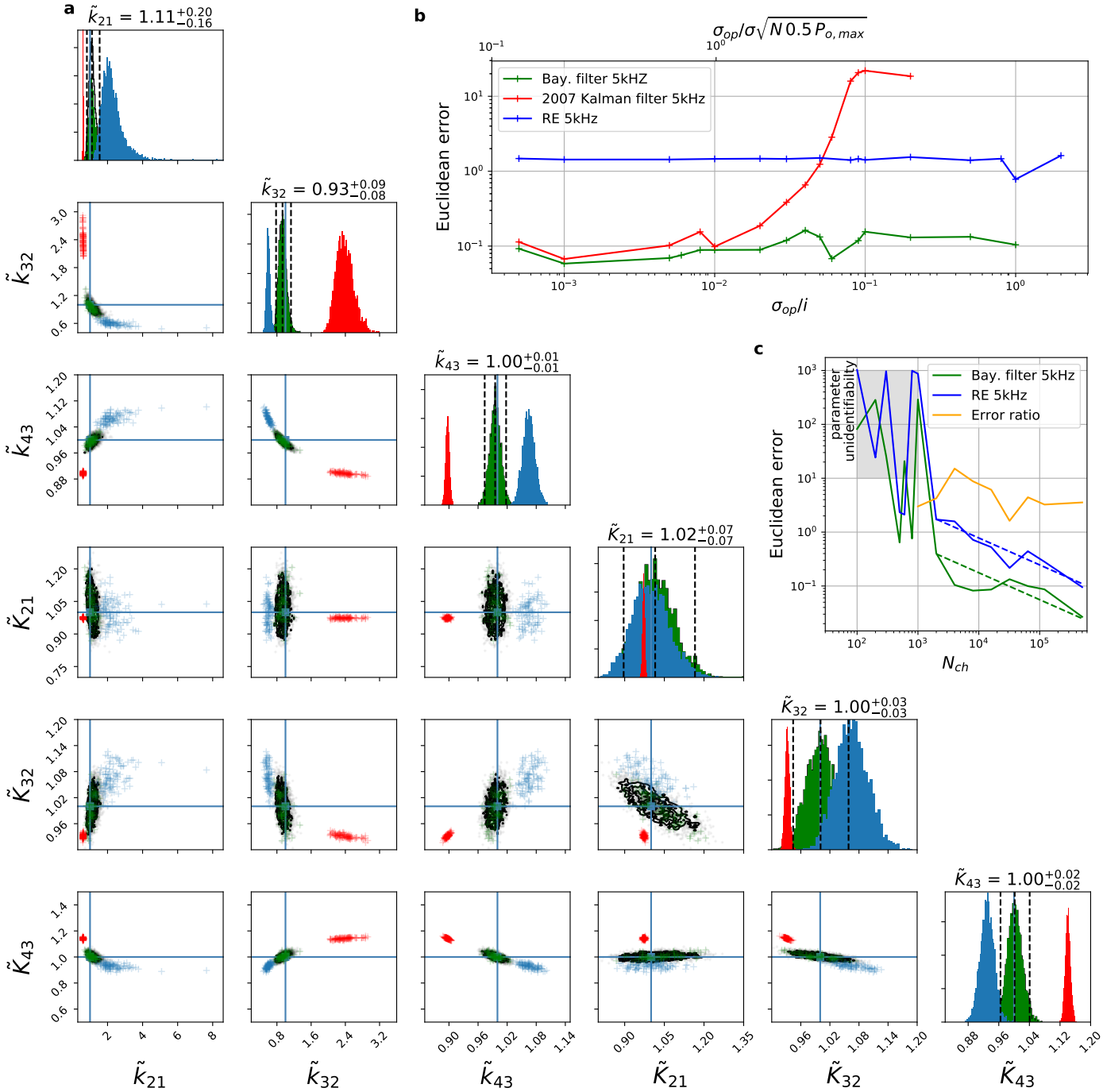


Figure 3. Comparison between a RE approach, a classical Kalman filter and the Bayesian filter for patch-clamp data and its scaling with $N_{\mathbf{ch}}$ and open-channel noise. For all data sets in the figure the analysing frequencies f_{ana} ranges within $(2-5)$ kHz. **a,** Posterior of the rate matrix of the classical deterministic RE (blue), 2007 Kalman filter (red) and our Bayesian filter (green) for a dataset with $N_{\text{ch}} = 5 \cdot 10^3$ and $\sigma_{\text{op}}/i = 0.05$. The blue crosses indicate the true values. All samples are normalized by their true values which is indicated by the ˜above the parameters. **b,** The Euclidean error is plotted vs. σ_{op}/i (low axis) for the Bayesian filter, the classical Kalman filter algorithm and the rate RE approach. **c,** Scaling of the Euclidean error vs. N_{ch} follows $\sim (N_{\text{ch}})^{-0.5}$for $N_{\text{ch}} > 2 \cdot 10^3$ for the RE and the Bayesian filter approach. The data indicates a constant error ratio (orange) for large N_{ch}. For $N_{\text{ch}} < 2 \cdot 10^3$ samples of the posteriors for many data sets suggest an improper posterior. For the instrumental noise we used $\sigma_{\text{ex}}/i = 1\ \sigma_{\text{op}}/i = 0.01$.

or even particle filters which avoid the normal approximation.
Each prediction of $\mathbf{n}_t$ (Eq. 6) is followed by a correction step,

$$\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t) = \frac{\mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)}{\int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)\,\mathrm{d}\mathbf{n}_t}, \tag{7}$$

that allows to incorporate the current data point into the estimate, based on the Bayesian theorem ***Chen et al.*** (***2003***). Additionally, the KF assumes ***Anderson and Moore*** (***2012***); ***Moffatt*** (***2007***) a multivariate normal observation distribution

$$\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t) = \frac{\mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbf{n}_t, \mathbf{\Sigma}_t)\mathbb{P}(\mathbf{n}_t)}{\int \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbf{n}_t, \mathbf{\Sigma}_t)\mathbb{P}(\mathbf{n}_t)\,\mathrm{d}\mathbf{n}_t}, \tag{8}$$

If the initial prior distribution is multivariate normal then due to the mathematical properties of the normal distributions the prior and posterior $\mathbb{P}(\cdot)$ in Eq. 8 become multivariate normal ***Chen et al.*** (***2003***) for each time step. In this case one can derive algebraic equations for the prediction (Methods Eq. 37 and 38d) and correction (Methods Eq. 58 and Eq. 58) of the mean and covariance. The algebraic equations originate from the fact that a normal prior is the conjugated prior for the mean value of a normal likelihood. Due to the recursiveness of its equations, the KF has a time complexity that is linear in the number of data points, allowing a fast algorithm. The denominator of Eq. 8 is the normal distributed marginal likelihood $\mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})$ for each data point, which constructs by

$$\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \prod_{t=2}^{N_T} \mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta}) = \prod_{t=2}^{N_T} \int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})\,\mathrm{d}\mathbf{n}_t = \prod_{t=2}^{N_T} \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}[\mathbf{n}_t], \mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \mathbf{\Sigma}_t), \tag{9}$$

a product marginal likelihood of normal distributions of the whole time trace $\mathcal{Y}_\mathrm{T} = \{\mathbf{y}_1, \dots, \mathbf{y}_{N_\mathrm{T}}\}$ of length N_T for the KF. For the derivation of $\mathbf{P}_t$ and $\mathbf{\Sigma}_t$ see Methods Eq. 38d and Eq. 43. The likelihood for the data allows to ascribe a probability to the parameters $\boldsymbol{\theta}$, given the observed data (Methods Eq. 20). An illustration for the operation of the KF on the observation space is given in Fig. 2**b**. The predicted mean signal $\mathbf{H}\mathbb{E}[\mathbf{n}(t)]$ corresponds to binding degree $B(t) = \frac{\mathbf{H}\mathbb{E}[\mathbf{n}(t)]_1}{N_\mathrm{ch}}$ and open probability $P_O(t) = \frac{\mathbf{H}\mathbb{E}[\mathbf{n}(t)]_2}{N_\mathrm{ch}}$. These values are plotted as vector trajectories.
The standard KF ***Moffatt*** (***2007***); ***Anderson and Moore*** (***2012***); ***Chen et al.*** (***2003***) has additive constant noise $\Sigma_t = const$ in the observation model. Thus, in this case a constant variance term is added, in Eq. 9 to the variance $\mathbf{H}\mathbf{P}_t\mathbf{H}^\top$ which originates (Eq. 38d) from the fact that we do not know the true system state $\mathbf{n}(t)$. For signals with Poisson-distributed photon counting or open-channel noise we need to generalize the noise model to account for additional white-noise fluctuations with $\mathbf{n}(t)$-dependent *variance*. For instance, in single-channel currents additional noise is often observed whose variance is referred to by σ^2_op. In macroscopic currents this additional noise can be modeled by a term $\sigma^2_\mathrm{op} n_4(t)$, causing state-dependency of our noise model.

$$y(t) = \mathbf{H}\mathbf{n}(t) + v_\mathrm{m}(t) + v_\mathrm{op}(t) \Leftrightarrow y \sim \mathrm{O}(y|\mathbf{n}) = \mathcal{N}\left(y|\mathbf{H}\mathbf{n}(t), \sigma^2_\mathrm{m} + n_4(t)\sigma^2_\mathrm{op}\right) = \mathcal{N}(y|\mathbf{H}\mathbf{n}(t), \Sigma_t) \tag{10}$$

The second noise term v_op is defined in terms of the first two moments $\mathbb{E}(v_\mathrm{op}) = 0$ and $\mathrm{var}(v_\mathrm{op}) = \mathbb{E}(v^2_\mathrm{op}) = \sigma^2_\mathrm{op} n_4(t)$. To the best of our knowledge such a state-dependent noise makes the integration of the denominator of Eq. 8 (which is also the incremental likelihood) intractable

$$\mathbb{P}(y(t)) = \int \mathcal{N}(y|\mathbf{H}\mathbf{n}, \sigma^2_\mathrm{m} + n_4\sigma^2_\mathrm{op})\ \mathcal{N}(\mathbf{n}|\overline{\mathbf{n}}(t), \mathbf{P}(t))\,\mathrm{d}n \tag{11a}$$

$$= \frac{1}{const} \int \exp\left(\frac{(y - \mathbf{H}\mathbf{n})^2}{2(\sigma^2_\mathrm{m} + n_4\sigma^2_\mathrm{op})}\right) \exp\left(\frac{1}{2}(\mathbf{n} - \overline{\mathbf{n}}(t))\mathbf{P}^{-1}(\mathbf{n} - \overline{\mathbf{n}}(t))^\top\right)\,\mathrm{d}n \tag{11b}$$

This is because the state distribution $\mathcal{N}(\mathbf{n}|\overline{\mathbf{n}}(t), \mathbf{P}(t))$ as the prior also influences the variance parameter of the likelihood which means that the conjugacy property is lost. While a normal distribution is the conjugated prior of the mean of a normal likelihood, it is not the conjugated prior for the variance. However, by applying the theorem of total variance decomposition Eq. 46 we deduce

a normal approximation to Eq. 8 and to the related problem of Poisson-distributed noise in fluorescence Eq. 55c and 57 data. By computing the mean and the variance or covariance matrix of the signal we can reformulate the noise model to fit the form of the traditional KF framework. Note, that the derived equations for the covariance matrix are still exact for the more general noise model. Mean and covariance just do not form a set of sufficient statistics anymore.

Our derivation is not limited to ligand-gated ion channels. For example, when investigating voltage-gated channels, the corresponding noise model can be easily adapted. This holds also when using the P/n protocol for which the noise model resembles that of the additional variance in the fluorescence signal. The additional variance is induced because the mean signal from the ligands swimming in the bulk (Methods Eq. 43 Appendix 5) is eliminated by subtracting scaled mean reference signal which itself has an error. This manipulation adds additional variance to the resulting signal comparable to P/n protocol. Other experimental challenges, as e.g. series resistance compensation promoting oscillatory behavior of the amplifier, deserve certainly advanced treatment. Nevertheless, for voltage-clamp experiments with a rate equation approach it also becomes clear ***Lei et al.*** (***2020***) that modeling of the actual experimental limitations, including series resistance, membrane and pipette capacitance, voltage offsets, imperfect compensations by the amplifier, and leak currents are necessary for consistent kinetic scheme inference.

The Bayesian posterior distribution

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) = \frac{\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\int \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{12}$$

encodes all information from model assumptions and experimental data used during model training (see Methods). A full Bayesian inference is usually not an optimization (finding the global maximum or mode of the posterior or likelihood) but calculates all sorts of quantities derived from the posterior distribution such as mean values of any function f including the mean value or covariance matrix of the parameters themselves or even the likelihood of the data.

$$\mathbb{E}[f] = \int f(\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T)d\boldsymbol{\theta} \tag{13}$$

Besides the covariance matrix of the parameter to express parameter uncertainty, the posterior allows to calculate a credibility volume. The smallest volume V_P that encloses a probability mass P of

$$P = \int_{V_P} \mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T)d\boldsymbol{\theta}. \tag{14}$$

is called the Highest Density Credibility Volume/Interval (HDCV/HDCI). Those credibility volumes should not be confused with confidence volumes although under certain conditions they can become equivalent. Given that our model sufficiently captures the true process, the true values $\boldsymbol{\theta}_{\text{true}}$ will be inside that volume with a probability P. Unfortunately, typically there is no analytical solution to Eq. 12. However, it can be solved numerically with Monte Carlo techniques, enabling to calculate all quantities related to Eq. 13 and Eq. 14. Our algorithm uses automatic differentiation of the statistical model to sample from the posterior (Appendix Fig. 1**a**) via Hamiltonian Monte Carlo (HMC) ***Betancourt*** (***2017***),see Appendix 1, as provided by the Stan software ***Hoffman and Gelman*** (***2014***); ***Gelman et al.*** (***2015***).

Benchmark for PC data against the gold standard algorithms

We compare the posterior distribution (Fig. 3**a**) of our algorithm against the deterministic ***Milescu et al.*** (***2005***) and stochastic ***Moffatt*** (***2007***) gold standard algorithms for patch-clamp data (Fig. 3). Both former algorithms are far away from the true parameter values with their maxima and also fail to cover the true values within reasonable tails of their posterior. Note that in Fig. 3**a** parameter unidentifiability by heavy tails/ multiple maxima of the posterior distribution or (anti-) correlation is easily visible as non axial symmetric patterns. To assess the location of the posterior conditioned

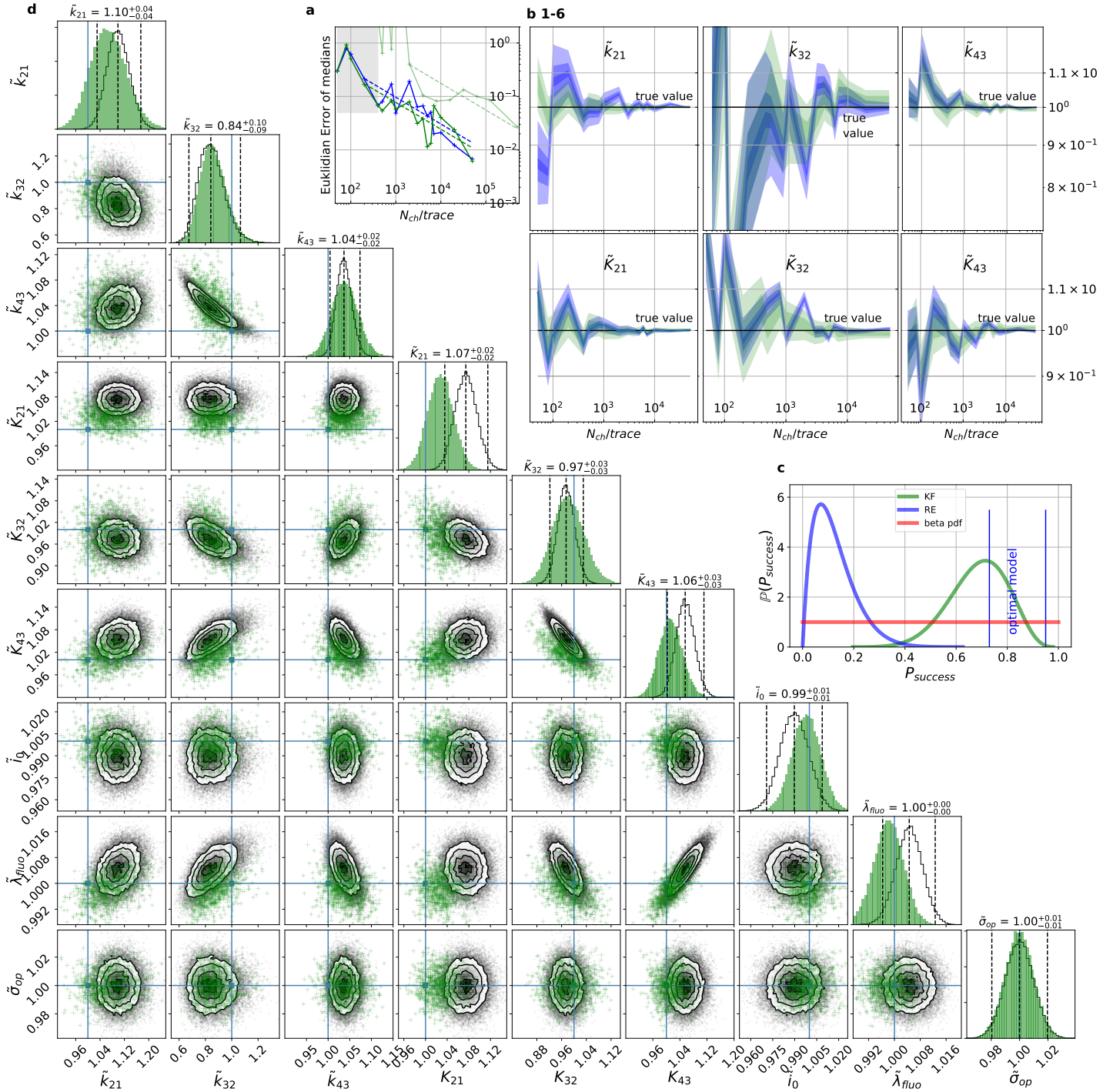


Figure 4. Benchmark of the Bayesian KF against Bayesian REs for cPCF data By $\tilde{\theta}_i$ we indicate that the samples from the posterior are normalized by their true value. **a,** Euclidian errors of the maximum for the rate k_{ij} and equilibrium constants K_i obtained by the KF (green) and from the REs (blue) are plotted against N_{ch} for $\sigma_{\mathrm{ex}}^2/i = 0.25$, $\sigma_{\mathrm{op}}^2/i = 0.01$ and $\lambda_b = 0.375$. Both algorithms scale like $1/\sqrt{N_{\mathrm{ch}}}$ for larger N_{ch} which is the expected scaling since the intrinsic noise divided by the mean signal scales in the same way. With lighter green we indicate the Euclidean error of patch-clamp data set. For smaller $N_{\mathrm{ch}} < 500$ the error is roughly the same indicating that limitations of the normal approximation to the multinomial distribution dominate the overall error in this regime. **b,** HDCI and the mode of the 3 k_{ij} and 3 K_i plotted vs. N_{ch} revealing that the maximum is a consistent estimator (converges in distribution to the true value with increasing data quality). While the KF (green) 0.95-HDCI includes usually the true value, the RE HDCI (blue) is too narrow and, thus, the real values are frequently not included. **c,** Bayesian estimation of true success probability for the event that all 6 HDCI include the respective true values at the same time by a binomial likelihood. Since the data sets have different N_{ch} and the model approximations become better with increasing N_{ch}, we use a cut-off for $N_{\mathrm{ch}} = 200$. The two posteriors of the true success probability were updated from $beta(1,1)$ prior (red) which is a uniform prior on the open interval $(0,1)$. **d,** Comparison of 1-D and all combinations of 2-D marginal posteriors of both algorithms calculated from a $N_{\mathrm{ch}} = 10^3$ simulation. For visual clarity we suppressed all other inferred parameters. Blue lines indicate the true value. We depict that in two dimensions the disproportion of the deviation of the mode and the spread of RE (grey) approach is worsened while KF (green) posterior includes the true values with more reasonable probability mass.

on N_{ch}, we select the median vector $\boldsymbol{\theta}$ of the marginal posteriors and calculate its Euclidean distance to the true values by:

$$\text{Euclidean Error} = \sqrt{\sum_i [\theta_i/\theta_{i,\mathrm{true}} - 1]^2} \tag{15}$$

This defines a single value to judge the overall accuracy of the posterior. Varying σ_{op}/i reveals the range of the validity (Fig. 3**b**) of the algorithm from ***Moffatt*** (***2007***). While both stochastic approaches are nearly equivalent for low open-channel noise, the RE performs consistently poorer. It may seem surprising that even for $\sigma_{\mathrm{op}}/i < 0.01$ the two stochastic algorithms start to produce different results. But considering the scaling (Methods Eq. 46) of the total open-channel noise (top axis) from currents of an ensemble patch $\propto (N_{\mathrm{ch}} P_{\mathrm{open,max}} 0.5)^{0.5} \sigma_{\mathrm{open}}$ one sees that if $\propto (N_{\mathrm{ch}} P_{\mathrm{open,max}} 0.5)^{0.5} \sigma_{\mathrm{open}}$ approaches σ the traditional KF suffers from ignoring state dependent noise contributions. The lower scale changes with experiments (e.g. N_{ch} and σ_{op}). In contrast, the upper scale is largely independent of the particular measurements. The two different normalizations indicate an experimental intuition: "Why should I consider the extra noise from the open state of the single channel if only $\sigma_{\mathrm{op}}/i = \sigma_{\mathrm{op}}/\sigma \approx 0.01$" is misleading. The small advantage of our algorithm for small σ_{op}/i over ***Moffatt*** (***2007***) is due to the fact that we could apply an informative prior in the formulation of the inference problem on $\sigma_{\mathrm{exp}} \sim \mathrm{normal}(\sigma^2_{\mathrm{exp,true}}, \sigma^2_{\mathrm{exp,true}} \cdot 0.01)$ by taking advantage of our Bayesian filter. For an increasing data quality, which in our benchmark is an increasing N_{ch} per trace, we show (Fig. 3**c**) that the deterministic RE and our Bayesian filter are consistent, i.e. they converge in distribution to the true parameter values with their posterior maxima or median for increasing data quality. The scaling of the RE approach (blue) and our Bayesian filter (green) vs. N_{ch} shows that for large N_{ch} both algorithms seem to have a constant error ratio relative to each other. They are both well described by $\mathrm{error}(N_{\mathrm{ch}}) \propto a/\sqrt{N_{\mathrm{ch}}}$ with an error ratio computed from the fit of 4.4. Thus, although our statistical model is singular, its asymptotic learning behaviour ***Watanabe*** (***2007***) is similar to a regular model (Fig. 4**c**) which, however, means that the euclidean error from both algorithms stays different also for large N_{ch}. For data with $N_{\mathrm{ch}} < 2 \cdot 10^3$ the samples from the posterior typically indicate that the posterior is improper which is defined as

$$\int \mathbb{P}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \infty \tag{16}$$

We consider this case as unidentified parameters. In that case more data sets need to be collected or better priors to be applied.

Benchmarking for cPCF data against the gold standard algorithm

For the simulated time traces with an optimistically high signal-to-noise assumption, the posterior of the KF (from hereon KF denotes our Bayesian Filter) and a RE ***Milescu et al.*** (***2005***) approach are compared for cPCF data (Fig. 4**a-d**). For a brief introduction of the RE approach, see Appendix 3. The failure to analyze PC data with moderate open-channel noise ***Moffatt*** (***2007***) (Fig. 3 **a**) disqualifies the classical KF with its constant noise variance also as a useful algorithm for fluorescence data, because here the Poisson distribution of the signal generates an even stronger state dependency of the signal variance.

By "high signal-to-noise assumption," we refer to an experimental situation with a standard deviation of the current recordings $\sigma_{\mathrm{ex}}/i = 0.5$, a low additional $\sigma_{\mathrm{op}}/i = 0.05$, and a high mean photon rate per bound ligand and frame $\lambda_{\mathrm{b}} = 5$. Additionally, we assume vanishing fluorescence background noise generated by the ligands in the bulk. The benefit of the high signal-to-noise is that the limitations of the two different approximations to the stochastic process of binding and gating can be investigated without running into the risk of being compensated or obscured by the noise from the experimental setup. For these experimental settings (Fig. 4**a**) we calculate the Euclidean distance of the median (Eq.15) for different N_{ch}. For $N_{\mathrm{ch}} < 500$ (grey shaded area in Fig. 4**a**) the error of both algorithms is roughly the same. On the single parameter level (Fig. 4**b**) this can be seen as an onset of correlated deviations from the true value for both algorithms. Each marginal

posterior has for each N_{ch} a similar deviation in magnitude and direction. This indicates that in this regime the approximations to the involved multinomial distributions fail in a similar manner for both algorithms, implying that treating the autocorrelation of the gating and binding becomes irrelevant compared to the error induced by normal approximations. For larger N_{ch} the Euclidean error of the RE is on average 1.6 times larger than the corresponding error of the posterior mode of the KF, which we deduce by fitting the function $\mathrm{error}(N_{\mathrm{ch}}) = \frac{a}{\sqrt{N_{\mathrm{ch}}}}$. On the one hand both algorithms are better in approaching the true values than with patch-clamp data alone. On the other hand the smaller error ratio means, that adding a second observable constrains the posterior, such that much of the overfitting is prevented for the RE approach. By overfitting we define the adaptation of any inference algorithm to the specific details of the used data set due to experimental and intrinsic noise which is aggravated if too complex kinetic schemes are used. Similarly, ***Milescu et al.*** (***2005***) showed that the over fitting tendency of the RE can be reduced if the autocorrelation of the data is eliminated. The lighter green curve derives from PC data. The Euclidean error is roughly an order of magnitude larger for $N_{\mathrm{ch}} > 2000$. Thus, in this regime the cPCF data set is equivalent to 10^2 fold more time traces or 10^2 more N_{ch} in a similar PC data set. For $N_{\mathrm{ch}} < 2000$ only cPCF establishes parameter identifiability (given a data set of 10 ligand concentrations and no other prior information). In (Fig. 4**b 1-6**) we demonstrate the $0.95-$HDCI (Eq. 14) of all parameters and their modes vs. N_{ch}. Even though both algorithms are consistent, the RE approach covers the true values with its $0.95-$HDCI only occasionally. The modeling assumption of the RE approach of treating each data point as if it does not come from a Markov process but from an individual draw from a multinomial distribution with deterministically evolving mean and variance makes the parameter estimates overly confident (Fig. 4**b 1-6**). A likely explanation can be found by analyzing the extreme case where data points are sampled at high frequency relative to the time scales of the channel dynamics. The RE approach treats each data point as a new draw from Eq. 67 while in reality the ion channel ensemble had no time to evolve into a new state. In contrast, the KF updates its information about the ensemble state after incorporating the current data point and then predicts from this updated information the generalised multinomial distribution of the next data point. For $N_{\mathrm{ch}} > 200$, the marginal posterior of the KF usually contains the true value.

To investigate the six one-dimensional $0.95-$HDCIs simultaneously, we declare the analysis of a data set as successful if all $0.95-$HDCIs include the true values. Otherwise we define it as a failure. This enables to determine the true probability at which the probability mass of the two algorithms covers the true values in a binomial setting. The left blue vertical line in Fig. 4**c** indicates $p = 0.95^6 \approx 0.735$ which is the lower limit and which would be the true success probability for an ideal model whose six $0.95-$ HDCIs are drawn from $y \sim \mathrm{binomial}(0.95, 6)$. This is the probability of getting 6 successes in 6 trials. The right blue vertical line equals $p = 0.95$, signifying the upper limit obtained by treating the six $0.95-$ HDCIs as being drawn from $y \sim \mathrm{binomial}(0.95, 1)$ each, which is a rather loose approximation. All marginal distributions are computed from the same high-dimensional posterior which is formed by one data set for each trial. Thus, the six $0.95-$ HDCIs $y \sim \mathrm{binomial}(0.95, 1)$ must have success rates between those two extremes if the algorithm creates an accurate posterior. We next combine the binomial likelihood with the conjugated beta prior ***Hines et al.*** (***2014***) for mathematical convenience. On this occasion, for the sake of the argument, $beta(1, 1)$ seems sufficient. The estimated true success rate by the RE approach (blue) is ≈ 0.15 and therefore far away from the true success probability. In contrast, the posterior (green) of the true success probability of the KF resides with a large probability mass between the lower and upper limit of the success probability of an optimal algorithm (given the correct kinetic scheme).

Exploiting six one-dimensional posterior distributions does not necessarily answer whether the posterior is accurate in 6 dimensions but we can refine the used binomial setting. In Fig. 4**d** $\mathbb{P}(\tilde{k}_{32}, \tilde{K}_{43})$ we see that 2-D marginal distributions can, due to their additional degree of freedom, twist around the true value without covering it with HDCV (Eq. 14) of reasonable size while simultaneously the two 1–D marginal distribution do cover it with a reasonable HDCI. In general, the KF

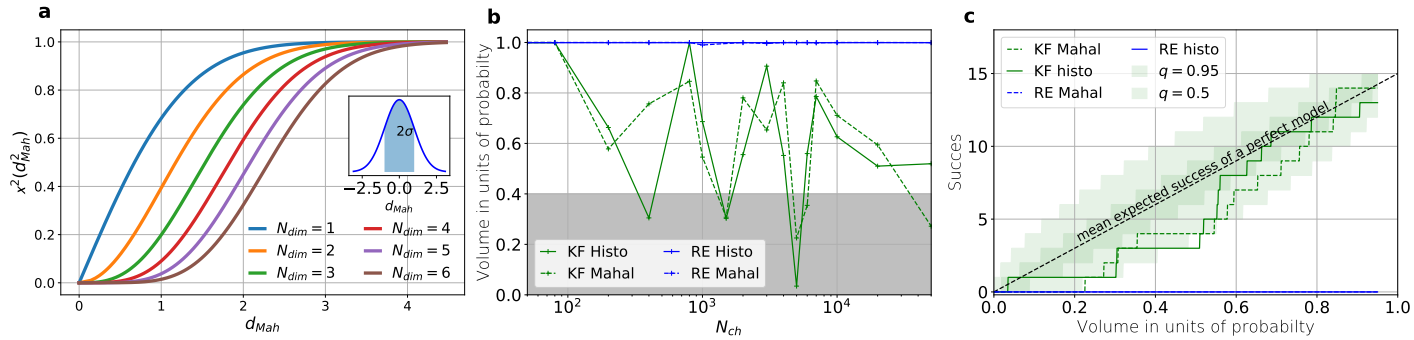


Figure 5. How to evaluate the overall shape of the posterior distribution. a, Cumulative χ-square distribution vs. d_{Mah}. The y axis denotes the probability mass which is counted by moving away from the maximum before an ellipsoid with distance d_{Mah} is reached. The different colours represent the changes of the cdf with an increasing number of rate parameters. The blue cdf at $d_{\mathrm{Mah}} = 1$ represents how much probability mass can be found from $\int_{-\sigma}^{\sigma} \mathrm{normal}(\theta, 0, \sigma) d\theta$, see inset. In one dimension we can expect to find the true value within 2σ around the mean with the usual probability of $P = 0.682$ for univariate normally-distributed random variables. The six parameters (brown) of the full rate matrix will almost certainly be beyond $d_{\mathrm{Mah}} = 1.0$. The higher the dimensions of the space the less important becomes the maximum of the probability density distribution for the typical set which is by definition the region where the probability mass resides. The mathematical reason for this is that the probability mass $P = \int_V \mathbb{P}(\boldsymbol{\theta}) dV$ is the integrated product of volume and probability density. **b,** The two methods to count volume in units of probability mass for the KF (green) and the RE (blue). For the optimistic noise assumptions $\sigma_{\mathrm{ex}} = 0.25 \cdot i$, $\sigma_{\mathrm{op}} = 0.025 \cdot i$ and a mean photon count per bound ligand per frame $\lambda_{\mathrm{b}} = 5$ the RE approach (blue) distributes the probability mass such that the HDCV never includes the true rate matrix. From $N_{\mathrm{ch}} > 100$ both HDCV estimates of the KF posterior (green curves) include the true value within a reasonable volume and show a similar behaviour. **c,** Binomial success statistics of HDCV to cover the true value vs. the expected probability constructed from the data of b). Calculated for $i = 0.25\sigma$ and $\sigma_{\mathrm{op}} = 0.025i$ and $\lambda_{\mathrm{b}} = 5$ and minimal background noise.

posterior distribution has its mode much closer to the true value for various parameter combinations and it seems that the posterior is approximately multivariate normal. Further, we recognize that the probability mass of the reasonably sized HDCV of the KF posterior includes the true values whereas the HDCV from the RE does not. In 6 dimensions we lack visual representations of the posterior. Since we showed that both algorithms are consistent for a given identifiable model, we are looking for a way to ask whether the posterior is accurate (has the posterior distribution the right shape). We can answer that question by asking, how much probability mass around the mode (or around multiple modes) needs to be counted to construct a HDCV Eq. 14 which includes the true values. Then we can ask for N_{set} data sets how often did we find the true values inside a volume $V(P)$ of a specific probability mass P of the posterior distribution

$$\mathrm{success} \sim \mathrm{binomial}(N_{\mathrm{set}}, P(V)). \tag{17}$$

An algorithm which estimates the parameters of the true process should fulfill this property simultaneously to being consistent. Otherwise credibility volumes or confidence volumes are meaningless. We explain (Appendix 8) in detail how to quantify the overall shape and n-dimensional posterior and comment on its geometrical meaning. One way is to use an analytical approximation via the cumulative Chi-squared distribution (Fig. 5**a,b**) The other way is to count the probability mass of n-dimensional histogram bins starting with the highest value until the first bin includes the true values (Fig. 5**b**).

Knowing how much volume/probability mass is needed to include the true rate matrix allows us to test whether all HDCVs constructed from the two probability distributions match the binomial distributions of the ideal model. For each data set and for each HDCV of a fixed probability mass there are two possible outcomes: The true rate matrix is inside or outside of that volume. For a chosen HDCV with a fixed probability volume, as indicated by a gray space in Fig. 5**b**, we count how many times the true matrix is included in the volume of that probability mass for each trail in a fixed amount of trials. Since the success is binomially distributed, we plot the expected mean of a perfect model $E[y] = N_{\mathrm{trials}} P_{\mathrm{true}}$ and binomial quantiles and compare them with the success rate found in our test runs (Fig. 5**c**) for both algorithms with both methods to determine the posterior

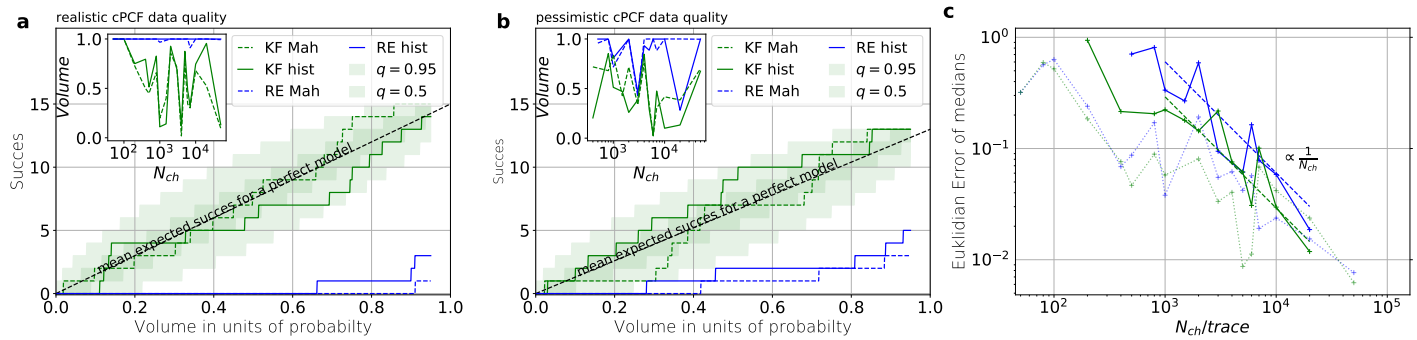


Figure 6. Overall shape and consistency for higher white noise levels. a, Binomial success statistics of HDCV to cover the true value vs. the expected probability. Calculated for $i = \sigma$ and $\sigma_{\text{op}} = 0.1i$ and $\lambda_{\text{b}} = 0.375$ and a strong background noise. **b,** Binomial success statistics of HDCV to cover the true value vs. the expected probability. For $10 \cdot i = \sigma$ and $\sigma_{\text{op}} = 1i$ and $\lambda_{\text{b}} = 0.375$ and a strong background noise. For both algorithms the adaptation of the sampler of the posterior was more fragile for small N_{ch}, leading to differences in the posterior if the posterior is constructed from different independent sampling chains. Those data sets were then excluded. We assume that these instabilities are induced in both algorithms by the shortcomings of the multivariate normal assumptions. **c,** Comparison of the Euclidean error vs. N_{ch} for the pessimistic noise case (solid lines) with Euclidean error for the optimistic noise case (dotted lines).

shape. The posterior of the KF distributes the probability mass in a consistent manner such that each volume includes the true rate matrix within the quantile range. In contrast, the RE approach fails for all data sets for all HDCVs (from 0 to 0.95 probability mass) and does not include the true values in one single case. Note, that all the binomial trials for each HDCV are made from the same set of data sets which explains the correlated deviation from the mean. For lower but realistic signal to noise ratios, where the fit quality decreases, e.g. by producing larger errors/wider posterior distributions (Fig. 6**a**), the statistics of the HDCV from the RE approach improve but are still outperformed by the KF. In particular, in our tested case of realistic experimental noise we never find the true values within a 0.65-HDCV if the data are analysed with a RE approach. Even for the highest noise level (Fig. 6**b**) the probability mass of the KF posterior needed to include the true rate matrix remains almost always smaller then the posterior mass of the RE approach. That means that the posterior mass of the KF is much closer to the true value distributed than the posterior mass of the RE. With the KF we find the true rate matrix for one data set in small volume $P < 0.05$ around the mode. To achieve the same with the RE approach we need at least a probability mass of 0.3. In the inset of Fig. 6**a,b** we do not observe a trend, thus no indication that the RE approach has a better performance for large values N_{ch} in this regard. This challenges the common argument that the RE approach should be equivalent to the KF for large N_{ch} because the ratio of mean signal vs. the intrinsic binding and gating noise is so large. Thus, including the autocorrelation into the analysis is important even for unrealistic large N_{ch}. One possible explanation is model a signal-to-noise ratio which scales $\propto N_{\text{ch}}$. From the multinomial distribution both algorithms inherit mean signals which scale $\propto N_{\text{ch}}$ and variances which scale in the terms dominating for large N_{ch} similarly with $\propto N_{\text{ch}}$. Thus, identical to the real signal, both algorithms model the scaling of the signal-to-noise ratio $\propto \sqrt{N}$. It is plausible, that both algorithms remain sensitive for the occurrence of autocorrelation of the noise even for largest signal-to-noise ratios. In Fig. 5**c** we compare the Euclidean error of the pessimistic high white noise case with an over-optimistic low noise case. We see, that when increasing N_{ch} there is a regime where the Euclidean error increases faster than $\sqrt{N_{\text{ch}}}^{-1}$ which we indicate with a coarse approximate fit $\propto N_{\text{ch}}^{-1}$. In that regime two effects happen simultaneously. First, the mean and the intrinsic fluctuations of the signal become more and more dominant over the experimental noise. Second, the standard deviation of intrinsic fluctuations becomes smaller relative to the mean signal. We speculate, that this produces together a learning rate which is faster than the usual asymptotic learning rate $\sqrt{N_{\text{ch}}}^{-1}$ of a regular model but relaxes asymptotically towards $\sqrt{N_{\text{ch}}}^{-1}$.

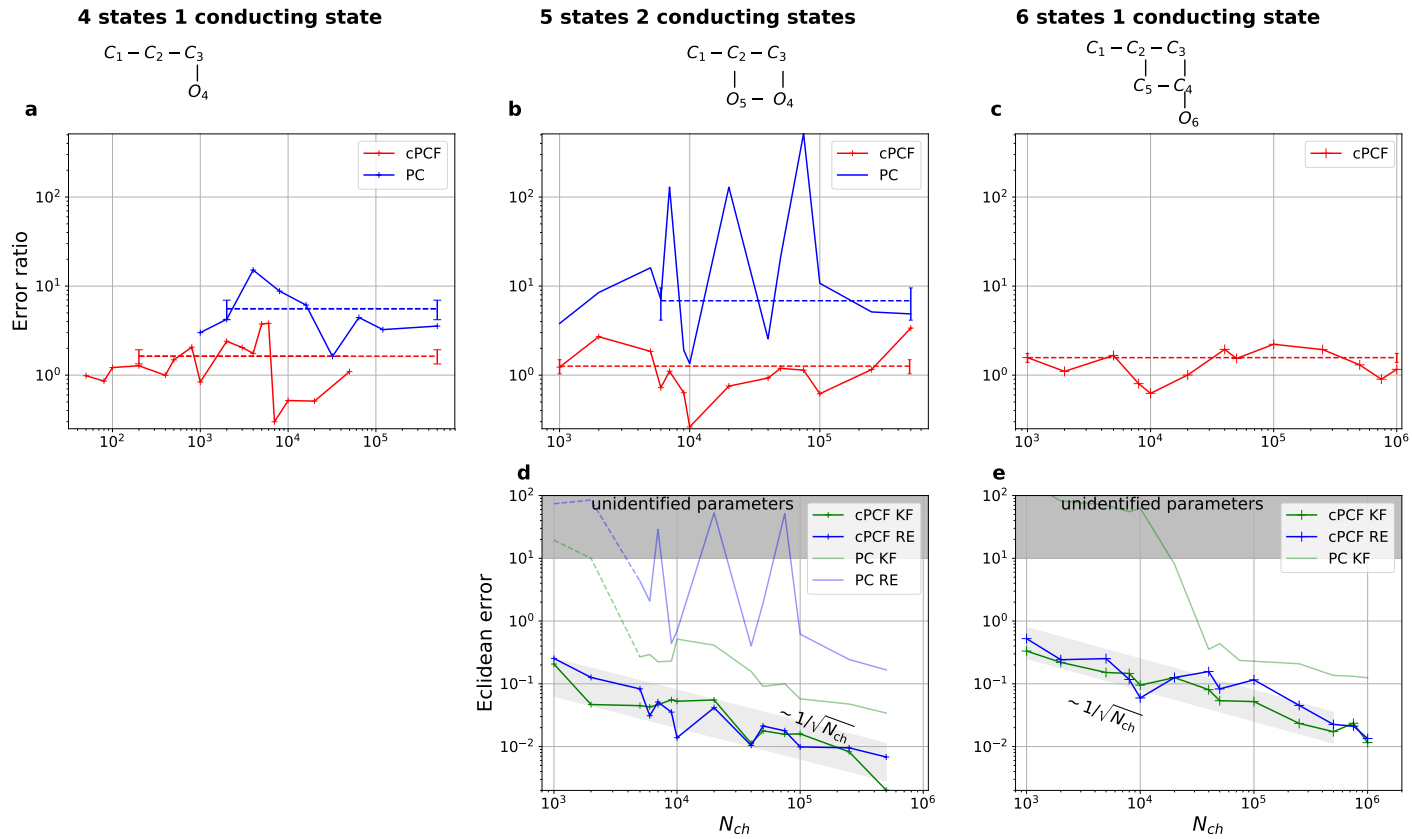


Figure 7. Scaling of the Euclidean error with data quantity and model complexity. Each panel column corresponds to a particular true process with increasing complexity from left to right, as indicated by the model schemes on top. Within all kinetic schemes each transition to the right adds one bound ligand. Each transition to left is an unbinding step. Vertical transitions are either conformational or opening transitions. Plots in each row share the same x- and y-axis respectively. All columns share the same abscissa. **a-c,** Error ratio for PC data (blue) and cPCF data (red). The dashed lines indicate the mean error ratio under the simplifying assumption that the error ratio does not depend on N_{ch}. The vertical bars are the standard deviations of the mean values. Theses values were calculated from the Euclidean errors shown in Fig. 3**c** and 4**a** for **a,** and panels **d-e,** for **b-c,** respectively. Results from the KF algorithm (green) and the RE algorithm (blue) are compared for PC (lighter shades) and cPCF (strong lines). The diagonal gray areas indicate a $\sim (N_{\mathrm{ch}})^{-0.5}$ proportionality. For simulating the underlying PC data, we used standard deviations of $\sigma_{\mathrm{op}} = 0.1$ and $\sigma = 1$ and for the cPCF data additionally a ligand of brightness $\lambda_{\mathrm{b}} = 5$. To facilitate the inference for the two more complex models, we assumed that the experimental noise and the single channel current are well characterized, meaning $i \sim \mathcal{N}(i|1, 0.01)$, $\sigma \sim \mathcal{N}(\sigma|1, 0.01)$ and $\sigma_{\mathrm{op}} \sim \mathrm{gamma}(\sigma_{\mathrm{op}}|1, 100)$. In the models containing loops (last 2 columns), a prior was used to enforce microscopic-reversibility and set to $k_{25}^{\star} \sim \mathrm{beta}(100, 100)$ multiplied by $k_1 = k_5 k_6 k_7 k_8 (k_2 k_3 k_4)^{-1} \cdot 0.995 + 0.01 \cdot k_1^{\star}$.

Statistical properties of both algorithms for more complex models.

We have seen in (Fig. 3**c**) and (Fig. 4**a**) that the RE and the KF algorithm are consistent estimators, while their error ratio (Fig. 7**a**) seems to have no trend to diminish with increasing N_{ch}. Adding a second observable increases parameter accuracy and adds identifiability for both algorithms since less aspects of the dynamics need to be statistically inferred (Fig. 4**a**). Furthermore, the second observable takes away much of the tendency (compare Fig. 4**b 1-6** with 8) of the RE approach to overinterpret (overfit) which leads to a shrinking of the error ratio 5.6 ± 1.4 for PC data to smaller values for cPCF data (Fig. 7**a**) (red) which are on average still bigger than one, while the Euclidean error is reduced (Fig. 4**a**). If we then keep the amount and quality of the PC/cPCF data but increase the complexity of the model which produced the data (Fig. 7**b,d**) from a 4-state to a 5-state model (see kinetic schemes above Fig. 7**a-c**) we see that for cPCF data the error ratio stays roughly the same (difference between Fig. 7**a** and **b**). For PC data instead both algorithms deliver an unidentified k_{21} for $N_{\mathrm{ch}} \leqq 2 \cdot 10^3$ (defined as an improper posterior). For larger N_{ch} the KF always identifies all parameters while the RE fails at $N_{\mathrm{ch}} \in \{7000, 2000, 75000\}$ to identify k_{54}. Thus, the KF reduces

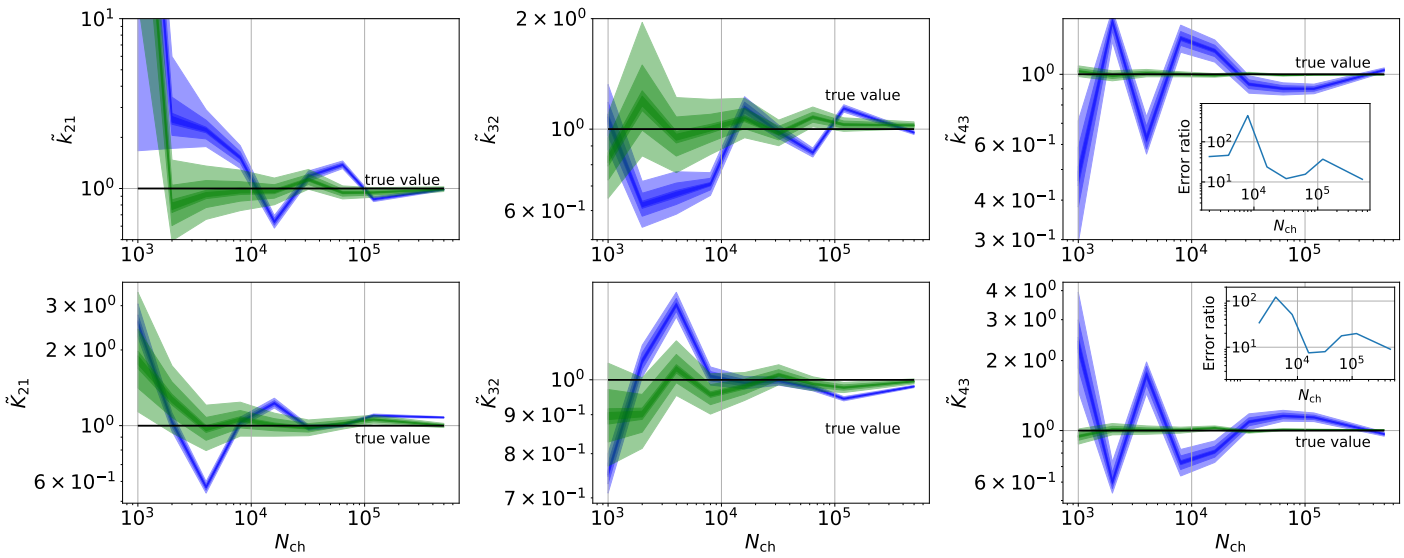


Figure 8. Convergence in the distribution of the HDCI for PC data obtained by the 4-state-1-open-state model. Comparison of a series of HDCIs shown as functions of N_{ch} for each parameter of the rate matrix obtained by the KF (green) and the RE algorithm (blue). The differing shades of green and blue indicate the set of $(0.9, 0.6, 0.2, 0.1)$-HDCIs. Only the interval $N_{\text{ch}} > 2 \cdot 10^3$ in which all parameters are identified is displayed. The data are taken from the KF vs. RE benchmark of Fig. 3**c** and 7**a**. The first row corresponds to three rates k_{ij} the second row to the equilibrium constants K_{ij}. All parameters are normalized by their true value. The insets show the error ratios of the respective single parameter estimates. Note that the error ratios on the single-parameter level can be much larger than the error ratios calculated from the Euclidean error if the errors of the respective parameters are small compared to other error terms in the Euclidean error Eq. 15.

the risk of unidentified parameters. To calculate the mean error ratio we exclude the values were some of the parameters are unidentified in total that still amounts to 6.8 ± 2.7 thus the advantage of the KF (given all parameters are identified) might increase with model complexity for PC data. The lowest Euclidean error for this kinetic scheme has cPCF data analysed with the KF. (Fig. 7**d**). A 6-state-1-open-states model with cPCF data has again an error ratio of the the usual scale (Fig. 7**c**). As expected, the Euclidean error continuously increases with model complexity (Fig. 7**d,e**). For PC data of the 6-state-1-open-states model even the likelihood of the KF is that weak (Fig. 7**e**) that it delivers unidentified parameters even for $N_{\text{ch}} = 10^4$ and we can detect heavy tailed distributions up until $N_{\text{ch}} = 10^5$. Using RE on PC data alone does not lead to parameter identification, thus no error ratio can be calculated.

Consistent with our findings, fluorescence data itself, should lower the advantage of the KF compared to PC data simply by signal-to-noise arguments. The stochastic aspect of the ligand binding is usually more dominated by the noise of Photon counting and background noise than the stochastic gating is dominated in current data by experimental noise. In terms of uncertainty quantification the advantage of the KF with cPCF varies with the model complexity (see, Appendix 9).

Besides analyzing what causes the changes in the Euclidean error (Fig. 7**a, b**) at the single parameter, we now investigate whether the posterior is a proper representation of uncertainty. Thus, we look back at the HDCIs. The HDCIs of the 4-state-1-open-state (Fig. 8) of the PC data from Fig. 3 reveal an exacerbated over-confidence problem of the RE approach (blue) compared to cPCF-data (Fig. 4**b1-6**). This, underlines our conclusion of Fig. 5 and 6 that the Bayesian posterior sampled by the RE approach is misshaped. As a consequence a confidence volume derived from the curvature at the ML estimate of the RE algorithm understates parameter uncertainty. A possible way for ML methods to derive correct uncertainty quantification is by using bootstrapping data methods ***Joshi et al.*** (***2006***). Furthermore, the error ratios of each single parameter from its true value in the last column $\tilde{k}_{43}$ $\tilde{K}_{43}$ strongly increased their magnitudes (insets Fig. 8). Even error ratios of $5 \cdot 10^2$ are possible. Note, that the way we defined Eq. 15 suppresses the influence of the smaller parameter

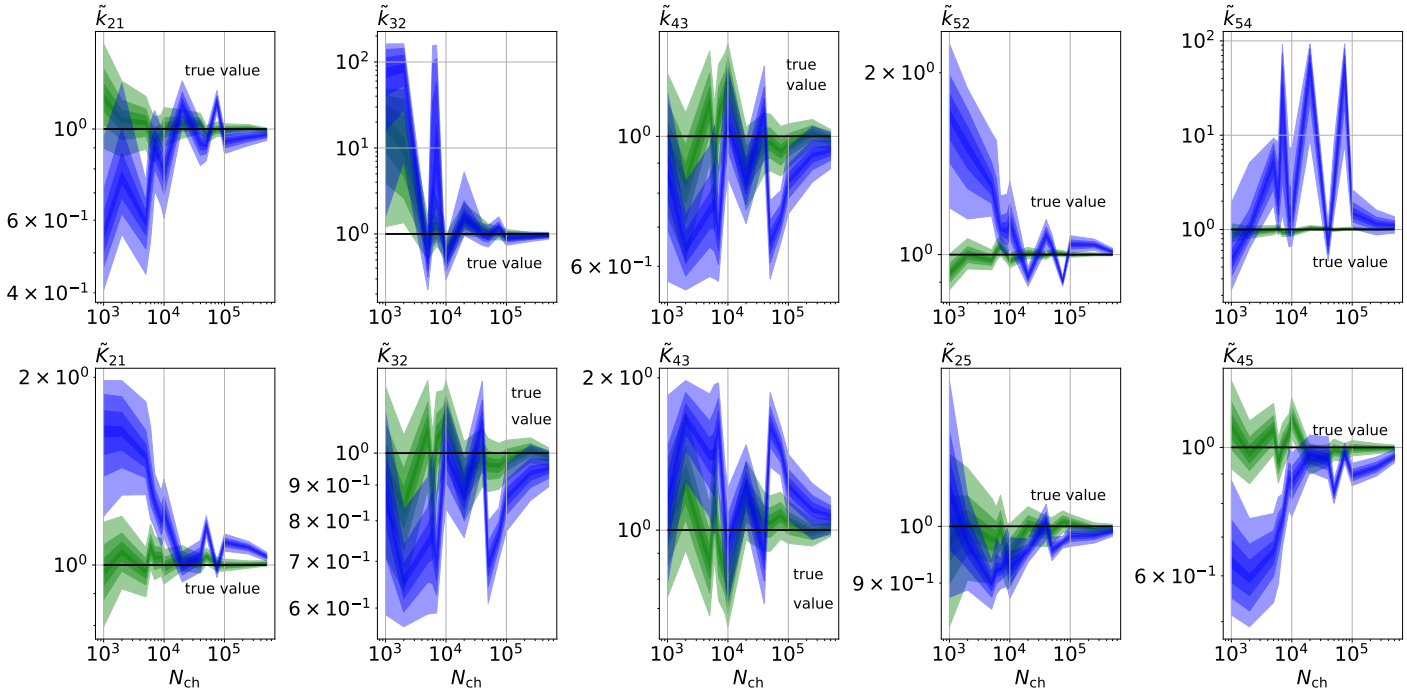


Figure 9. Convergence of the HDCIs for PC data for a 5-state-2-open-states model. Comparison of a series of HDCIs as functions of N_{ch} for each parameter of the rate matrix obtained by the KF (green) and the RE algorithm (blue). The HDCIs correspond to the data displayed in Fig. 7**b,d**. The first row corresponds to three rates k_{ij} the second row to the equilibrium constants K_{ij}. All parameters are normalized by their true value. $\tilde{K}_{25}$ is because of the microscopic-reversibility prior a parameter which is strongly dependend on the other rates and ratios. Refer to the caption of Fig. 7 for details about the prior that enforces microscopic-reversibility. Thus the deviations of $\tilde{K}_{25}$ are inherited from the other parameters.

errors in the overall error ratio. Thus the advantage (error ratio) of the KF over RE approach for a single parameter can be much larger or lower compared to the error ratio derived from the Euclidean error if the respective parameter is contributing less to the Euclidean error. The posterior of the KF (green) seems to be unbiased after the transition into the regime $N_{\mathrm{ch}} > 2 \cdot 10^3$ where all parameters are identified. Similarly, for the RE algorithm there is no obvious bias in the inference. If we use the RE algorithm and change from the 4-state to the 5-state model (PC data from Fig. 7**b**) , bias occurs (Fig. 9) in many inferred parameters, even for the highest N_{ch} investigated. ***Milescu et al.*** (***2005***) showed that one or the reason of the biased inference of the RE approach is its ignorance of autocorrelation of the intrinsic noise. We add here that the bias problem clearly aggravates with an increased model complexity. It is even present in unrealistically large patches which in principle could be generated by summing up 10^2 time traces with $N_{\mathrm{ch}} = 10^3$. In contrast, the KF algorithm reveals that its parameter inference is either unbiased or at least much less biased in the displayed N_{ch} regime. Furthermore, for both algorithms the position of the HDCI relative to the true value is for some parameters highly correlated, which corresponds to the correlation between optima of the ML method of ***Milescu et al.*** (***2005***); ***Moffatt*** (***2007***).

As a side note, unbiased parameter estimates are a highly desirable feature of an inference algorithm. For example, with a bias in the inference, repeated experiments do not lead to the true value if the arithmetic mean of the parameter inferences is taken. With bias even bootstrapping methods fail to produce reliable uncertainty quantification. Due to the variation of the data the k_{54} parameter is either identified in some neighbourhood of the true value or complete unidentified (Fig. 9), if the RE algorithm is used. The unidentified k_{54} occurs even at high quality data such as $N_{\mathrm{ch}} = 7.5 \cdot 10^4$. For the same data using the KF did not result in any unidentified parameters. Note, that comparable inference pathology's such as multimodal distributions of inferred parameter were also reported for the maximum likelihood RE algorithm for low quality PC data or too simple stimulation protocols ***Milescu et al.*** (***2005***).

In conclusion, the two different perspectives on parameter uncertainty: On the one hand distributions of ML estimates due to the random data ***Milescu et al.*** (***2005***); ***Moffatt*** (***2007***) and the Bayesian posterior distribution loose their tightly linked (and necessary) connection if the RE algorithm is used. Thus, KF robustifies also ML inferences of the rate matrix.

On the data side of the inference problem adding cPCF data eliminates the bias, reduces the variance of the position of the HDCI and eliminates unidentified parameters (Appendix 9 Fig. 1 and 2) for both investigated algorithms. This advantage increases with model-complexity.

For the 5-state and 6-state model, we applied microscopic-reversibility ***Colquhoun et al.*** (***2004***). We enforced it by hierarchical prior distribution (Methods Eq. 60) whose parameters can be chosen such that they allow only arbitrarily small violations of microscopic-reversibility. But the prior distribution can also be used to force some softer regularization around microscopic-reversibilty. Thus, transferring the unusually strictly applied algebraic constraint ***Salari et al.*** (***2018***) of microscopic-reversibilty to constraint with scalable softness we can model the lack information if microscopic-reversibilty is exactly fulfilled ***Colquhoun et al.*** (***2004***) by the given ion channel instead of forcing it upon the model.

Prior critique and model complexity

In the Bayesian framework the likelihood of the data and the prior generate the posterior. Thus, the performance of both algorithms can be influenced by appropriate prior distributions. We used a uniform prior over the rate matrix which is not optimal. Note, that uniform priors are widely used by several reasons. They appear to be unbiased, and are assumed to be a "no prior" option (which they are not). This is true for location parameters like mean values. In contrast, for other parameters, such as scaling parameters like rates or variances, a uniform prior adds bias to the inference towards faster rates ***Zwickl and Holder*** (***2004***). We suspect, that for the PC data even in the simplest model discussed here the lower data quality limit below which we detected unidentified parameters (improper posteriors) is caused by the uniform prior. This lower limit for the KF also

increases with the complexity of the model from $N_{\mathrm{ch}} < 2 \cdot 10^3$ for the 4-state model till $N_{\mathrm{ch}} \leqq 2 \cdot 10^4$ for 6-state-1-open-state model. Note, that it is hardly possible to fit the 6-state-1-open-state model with the RE approach for the same amount of PC data. We observe cPCF data eases this problem because the likelihood becomes more concentrated for all parameters. The likelihood dominates the uniform prior. Nevertheless, for most parts of the paper we used a uniform prior over the rates and equilibrium constants to be comparable with the usual default method: a plain ML which influences our results in data regimes in which the data is not strong enough to dominate the bias from the uniform prior. Thus both algorithms perform better with smarter informative or at least unbiased prior choices for the rate matrix.

In principle, to rule out an influence of the prior, unbiased priors should be used for the rates. The standard concept for unbiased least informative priors is to construct a Jeffreys prior ***Jeffreys*** (***1946***) for the rate matrix which is, however, beyond the scope of the paper.

The influence of the brightness of the ligands of cPCF data on the inference

To evaluate the advantage of cPCF data ***Biskup et al.*** (***2007***) with respect to PC data only (Fig. 10), we compare different types of ligands: Idealized ligands with brightness λ_{b}, emitting light only when bound to the channels, 'real' ligands which also produce background fluorescence when diffusing in the bath solution (Appendix 5) and current data alone. For datasets including fluorescence, the increased precision for the dissociation rate of the first ligand, $k_{2,1}$, is that strong that the variance of the posterior $\mathbb{P}(k_{2,1}, k_{3,2})$ nearly vanishes in the combined plot with the current data (nearly all probability mass is concentrated in a single point in Fig. 10**a**). The effect on the error of the equilibrium constants K_i is less strong. Additionally, the bias is reduced and even the estimation of N_{ch} is improved. The brighter the ligands are, the more the posterior of the rates decorrelates, in particular $\mathbb{P}(k_{2,1}, k_{3,2})$ (Fig. 10**a**). All median estimates of nine different cPCF data sets (Fig. 10**b**) differ by less than a factor 1.1 from the true parameter except $k_{3,2}$, which does not profit as much from the fluorescence data as $k_{2,1}$ (Fig. 10**c**). The 95th percentiles, l_{95} of $\mathbb{P}(k_{2,1})$ and $\mathbb{P}(K_1)$ follow $l_{95} \sim 1/\sqrt{\lambda_{\mathrm{b}}}$. Thus, with increasing magnitude of ligand brightness λ, the estimation of $k_{2,1}$ becomes increasingly better compared to that of $k_{3,2}$ (Fig. 10**c**). The posterior of the binding and unbinding rates of the first ligand contracts with increasing λ_{b}. The l_{95} percentiles of other parameters exhibit a weaker dependency on the brightness ($l_{95} \sim \lambda^{-0.1}$). For $\lambda_{\mathrm{b}} = 0.01$ photons per bound ligand and frame, which corresponds to a maximum mean signal of 20 photons per frame, the normal approximation to the Poisson noise hardly captures the asymmetry of photon counting noise included in the time traces. Nevertheless, l_{95} decreases about ten times when cPCF data are used (Fig. 10**c**). The estimated variance of

$$r(t_i) := \frac{y(t_i) - \left(\mathbf{H}\mathbb{E}[\mathbf{n}(t_i)]\right)}{\sqrt{\mathrm{var}[y(t_i)]}}, \tag{18}$$

with the mean predicted signal $\mathbf{H}\mathbb{E}[\mathbf{n}(t_i)]$, for PC or cPCF data is $\sigma^2(r_i) \approx 1$ (Fig. 10**d**) which means that the modeling predicts the stochastic process correctly up to the variance of the signal. Note that the mean value and covariance of the signal and the state form sufficient statistics of the process, since all involved distributions are approximately multivariate normal. The fat tails and skewness of $\mathbb{P}(k_{21})$ and $\mathbb{P}(k_{12})$ arises because the true model is too flexible for current data without further prior information. The KF allows to determine the variance (Fig. 10**e**) of the open-channel current noise for $\sigma_{\mathrm{op}} = 0.1i$. Data from cPCF experiments reduce even the variance of the posterior $\mathbb{P}(\sigma_{\mathrm{op}})$.

Sensitivity towards filtering before the analog-to-digital conversion of the signal

On the one side, every analog signal to be digitized needs analog filtering for antialiasing according to the Nyquist theorem. On the other side, every analog filter does not only suppress unwanted white noise but also distorts the dynamics (Fig. 11 a) of the signal of interest ***Silberberg and Magleby*** (***1993***). Therefore, ***Qin et al.*** (***2000***) recommend to avoid analog filtering as much as possible in single-channel analysis and let the HMM analyse the data in the rawest available form, even with simultaneous drift correction ***Sgouralis and Pressé*** (***2017***). One can also expect that analog

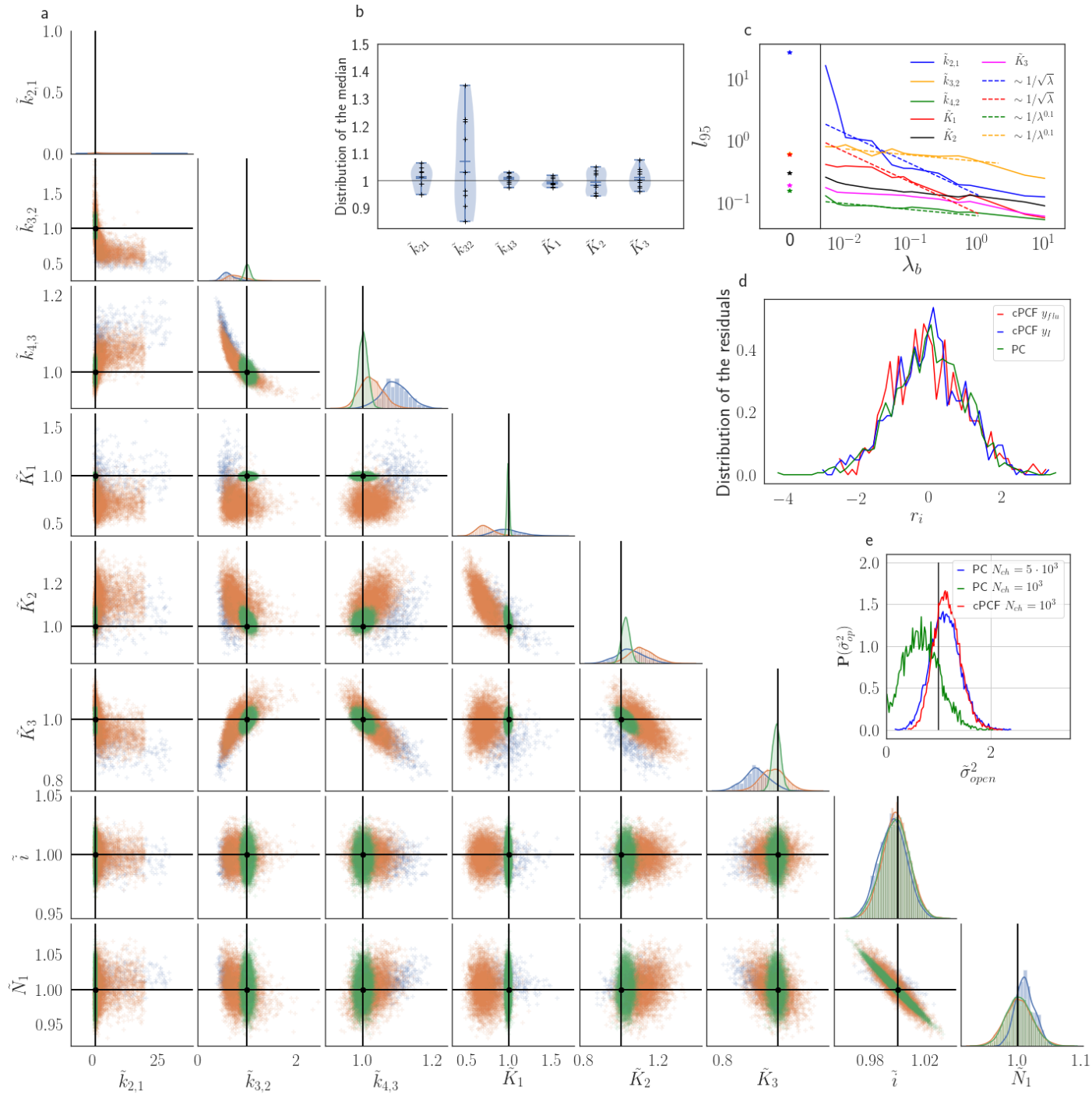


Figure 10. Benchmark of the KF for patch-clamp versus patch-clamp fluorometry data. a, Posteriors of PC data (blue), cPCF data with $\lambda_b = 0.00375$ (orange) and cPCF data with $\lambda_b = 0.375$ (green). For the data set with $\lambda_b = 0.375$ we additionally accounted for the superimposing fluorescence of unbound ligands in solution. In all cases $N_{ch} = 10^3$. The black lines represent the true values of the simulated data. The posteriors for cPCF $\mathbb{P}(k_{2,1}, k_{3,2})$ are centered around the true values that are hardly visible on the scale of the posterior for the PC data. **b,** Distribution of the median estimates visualized by a violin plot for the parameters of the rate matrix for 9 different data sets, with $\lambda_b = 0.375$ and superimposing bulk signal. The blue lines represent the median, mean and the maximal and minimal extreme value. **c,** The 95th percentile of the marginalized posteriors vs. λ_b normalized by the true value of each parameter. A regime with $l_{95} \sim 1/\sqrt{\lambda}$ is shown for $k_{2,1}$ and K_1, while other parameters show a weaker dependency on the ligand brightness. **d,** Histograms of the residuals r of cPCF with $\lambda_b = 2.5 \cdot 10^{-3}$ data and PC data. The randomness of the normalized residuals of the cPCF or PC data is well described by $r_i \sim \text{normal}(0, \sigma^2_{res} = 1)$. The estimated variance is $\sigma^2_{res} = 0.98 + 0.26$. Note that the fluorescence signal per frame is very low such that the normal approximation to Poisson counting statistics does not hold. **e,** Posterior of the open-channel noise $\mathbb{P}(\sigma^2_{op}/\sigma^2_{op,true})$ for PC data with $N_{ch} = 10^3$ (green) and $N_{ch} = 10^5$ (blue) as well as for cPCF data with $N_{ch} = 10^3$ (red) with $\lambda_b = 0.375$. Adding fluorescence data has roughly the same effect on the estimation of σ_{op} like using five times more ion channels to estimate σ^2_{op}. We assumed as prior for the instrumental variance $\mathbb{P}(\sigma^2) = \mathcal{N}(1, 0.01)$.

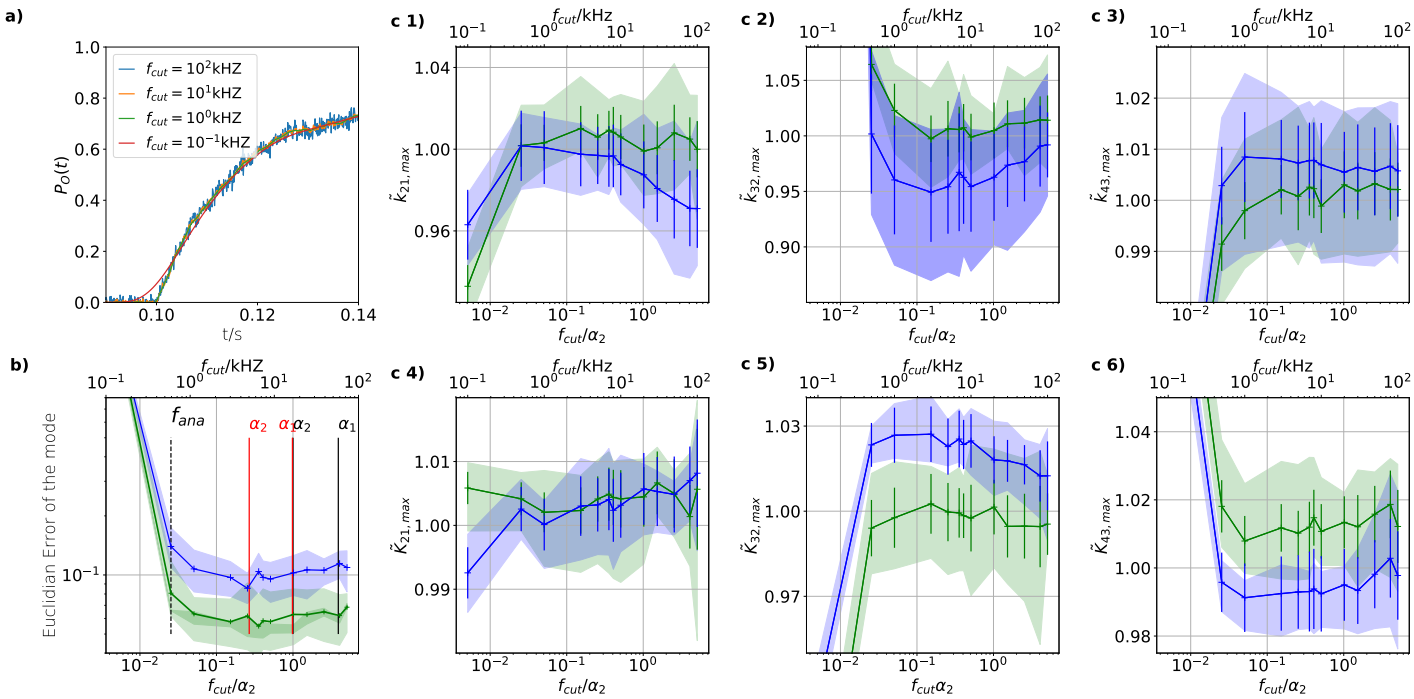


Figure 11. The influence of analog filtering on the accuracy and precision of the maxima of the posterior for cPCF data. In order to mimic an analog signal before the analog-to-digital conversion we simulated 7 different 100 kHz signals which were then filtered by a digital 4th-order (4 pole) Bessel filter. The activation curves were then analyzed with the Bayesian filter at 125 Hz and the deactivation curves at sampling rates between $166 - 500$ Hz. The slower frequency at which the Bayesian filter analyses the data is necessary because the applied Bessel filter has caused additional time correlations in the originally white noise of the signal. Thus, an all-data-points fit would immediately violate the white noise assumption of Eq. 4 which we restore by analyzing at a much lower frequency. We then let the time scales of the induced time correlations become larger and larger by decreasing f_{cut}. We chose for the analog signal $\sigma_{\mathrm{exp}}/i = 10$, $\sigma_{\mathrm{op}}/i = 0.1$, thus a stronger background noise, and we set the mean photon count per bound ligand as $\lambda_{\mathrm{b}} = 5$. For the ensemble size we choose $N_{\mathrm{ch}} = 10^3$. **a,** Time trace filtered with different f_{cut}. Except for 100 Hz (red) the signal distortion is visually undetectable. Nevertheless, the invisible signal distortions from analog filtering are problematic for both algorithms. **b,** Estimate of the distribution mean Euclidean error of the median of the posterior vs. the cut-off frequency of a 4 pole Bessel filter (upper scale is in units of kHz) or scaled to the channel time scale (lower scale, see text). The fastest two ratios $\alpha_{1,2}/\alpha_2$ for the highest ligand concentration are indicated by the black vertical lines. The fastest ratios $\alpha_{1,2}/\alpha_2$ for the next ligand concentration are indicated by the red vertical lines. The slowest eigenvalue ratio α_3/α_2 at the highest ligand concentration is beyond the left limit of the x-axis. The solid line is the mean median of 5 data sets of the respective RE posterior (blue) and KF posterior (green). The green shaded area indicates the 0.6 quantile (ranging from the 20th percentile till the 80th percentile), demonstrating the distribution of the error of the posterior median due to the randomness of the data.**c 1-3,** Accuracy (bias) and precision of the maxima of the posterior $\tilde{k}_{\mathrm{max},ij}$ of the posterior maxima of the rates vs. the cut-off frequency of a Bessel filter. The shaded areas indicate the 0.6 quantiles (ranging from the 20th percentile till the 80th percentile) due the variability among data sets while the error bars show the standard error of the mean. The deviation of the mean from the true value is an estimate of the accuracy of the algorithm while the quantile indicates the precision. **c 4-6,** Accuracy and precision of the maxima of the posterior $\tilde{K}_{\mathrm{max},ij}$ of the posterior maxima of the corresponding equilibria vs. the cut-off frequency of a Bessel filter.

filtering of a macroscopic signal is harmful for the inference of the KF and the RE approach. For the CCCO model considered herein we investigated the mean behavior of both algorithms with 7 data sets (simulated at 100kHz to mimic an analog signal). A digital 4th-order Bessel filter ***Virtanen et al.*** (***2020***) was then applied. The maximum analysing frequency f_{ana} of the KF used is $100-400$ Hz to be comparable to cPCF setups. Physically, the absolute cut-off frequency f_{cut} is irrelevant; what matters is the magnitude of f_{cut} relative to f_{ana} and to the eigenvalues α_i of the ensemble, since the eigenvalues determine the time evolution of the mean ensemble state, the autocorrelation, and Fourier spectrum of the fluctuations around the equilibrium distribution ***Colquhoun et al.*** (***1997***). The eigenvalues depend on the ligand concentration such that for a 4-state model for each ligand concentration there are 3 relevant time scales $-1/\alpha_i$ (where $i = 2, 3, 4$) plus the equilibrium solution which satisfies $\alpha_1 = 0$. For 10 different time series $3 \cdot 10 + 3$ the outcome is to have different values of α_i. For this reason we normalize in the following (Fig. 11) the cut-off frequencies by α_2 at the highest ligand concentration. We analyze the arithmetic mean from 7 different data sets of the median of the posterior of the rate matrix. The mean Euclidean error of the median (Fig. 11**b**) and a series of quantiles demonstrate that overall the error of the mean median of the posterior KF (green) is smaller than that obtained by the RE. For unfiltered data, the accuracy of the mean median of the KF is increased by ≈ 1.6. Based on the Euclidean error both algorithms benefit slightly from careful analog filtering for $f_{\mathrm{cut}}/\alpha_2 \geq 1$ while the offset remains rather constant. A strong negative effect of analog filtering starts for both algorithms around $f_{\mathrm{cut}} \approx 1\,\mathrm{kHZ}$. This is induced by $f_{\mathrm{cut}} \to f_{\mathrm{ana}}$ (see, Appendix 7) . In contrast, based on the level of each individual parameter of the rate matrix (Fig. 11**c 1-6**) the bias induced by analog filtering immediately starts with $f_{\mathrm{cut}} = 70\,\mathrm{kHz}$ (Fig. 11**c 1-3**). Note, that visual inspection of the signal (Fig. 11**a**) does not reveal signal distortions $f_{\mathrm{cut}} \geq 10\,\mathrm{kHz}$ though they are detected by both algorithms. For unfiltered data the maximum of the posterior for the RE approach is a biased estimate $E[\theta_{\mathrm{ME}}] \neq \theta_{\mathrm{true}}$ for at least the parameters $\tilde{k}_{21}, \tilde{K}_{21}, \tilde{K}_{32}$ of the true value θ_{true}, which is explained ***Milescu et al.*** (***2005***) by the fact that RE approaches ignore the autocorrelation of the intrinsic noise. Additionally, the data indicate that for $\tilde{K}_{43}$ the maximum of the posterior is even for the KF a biased estimate which we interpret as limitations induced by the fact that the mean vector and covariance-matrix do not constitute sufficient statistics as soon as Poisson distributed photon counting or open-channel noise blurs the signal. For the RE approach the additional bias induced by the analog filter on the mean maximum of all parameters of the posterior starts with $f_{\mathrm{cut}} \approx 70$ kHz or, in other words, at the fastest time scale in the whole data set. The total bias in the estimate is reduced for k_{21} with the additional bias from the analog filtering but increased for k_{32} which for the Euclidean error leads at first to a small increase in accuracy. The KF is more robust towards analog filtering, as the results alter less with f_{cut} (given a reasonable f_{cut}), and less biased for unfiltered data in the estimates of these parameters. On the one hand, the Euclidean error shrinks for $f_{\mathrm{cut}} > 10$kHz (Fig. 11**b**). On the other hand, on the single-parameter level (Fig. 11**c 1-6**), the parameter estimates pick up bias due the analog filtering even for high filter frequencies, in particular for the RE approach. Only for k_{43} the KF is more biased than the RE approach.

The KF is the unique minimal variance Bayesian filter for a linear Gaussian process ***Anderson and Moore*** (***2012***) which means, given that the assumptions of the KF are fulfilled by the true process of interest, the KF is mathematically proven the best model-based filter to apply. Consequently, analog filtering does not provide an advantage unless it removes specific high frequency external noise sources (colored noise). We demonstrate (Appendix 7) this for PC data and varied f_{cut} and f_{ana}. On the downside, increasing f_{ana} makes the results of both algorithms more fragile if $f_{\mathrm{cut}} \gg f_{\mathrm{ana}}$ does not hold. Thus, the critical edge in Fig. 11**b** is indeed induced by f_{cut} approaching f_{ana}. This suggests that the white noise assumption of both algorithms is violated. On the upside, if $f_{\mathrm{cut}} \gg f_{\mathrm{ana}}$ is given, the KF with an order of magnitude higher f_{ana} has a reduced bias of up to 20% for $f_{\mathrm{cut}} \to \infty$ for individual parameters compared to the KF with lower f_{ana}. Additionally, a higher f_{ana} reduces the variance. To reduce the bias of parameter estimates to a minimum, the experimental design offers two remedies, either doing cPCF experiments with additional discussed advantages or using

the KF at a high f_{ana} with even much higher f_{cut}.

By theoretical grounds a further argument for doing less analog filtering is that this benchmark analyses data of a finite state Markov process, which is a coarse proxy for the true process. In reality, relaxation of a protein is a high-dimensional continuous-state Markov process with infinitely many relaxation time scales (eigenvalues) ***Frauenfelder et al.*** (***1991***) which, however, might be grouped in slower experimentally accessible and non-accessible faster time scales ***Noé et al.*** (***2013a***). With larger data sets of higher quality from better experiments, the faster time scales might become accessible if they are not distorted by analog filtering. In conclusion, deciding on a specific kinetic scheme and inferring its parameters means finding a model which accommodates in the best way to the set of observed eigenvalues. Analog filtering hampers the RE, KF or HMM forward-backward algorithm ***Qin et al.*** (***2000***) to correctly describe the faster time scales.

Error due to finite integration time of fluorescence data

So far, we idealized the fluorescence data integration time as being instantaneously relative to the time scales of ensemble dynamics. In real experiments, the fluorescence signal of cPCF data has orders of magnitude longer minimal integration time T_{int} (time to record all voxels of a frame) or maximal integration frequency $f_{\text{int}} = 1/T_{\text{int}}$, than the possible sampling frequency of current recordings. We mimic the finite integration time

$$y_{\text{digital}}(t_i) = \int_{t_{\text{start}}}^{t_i = t_{\text{start}} + T_{\text{int}}} y_{\text{analog}}(t)dt \approx \sum_{j \in [t_{\text{start}}, t_{\text{start}} + T_{\text{int}}]} y(t_j)\Delta t \tag{19}$$

by summing with a sliding window the 100kHz signal including the white noise to obtain data at an effectively lower sampling frequency (Fig. 12**a**). Additionally we set the Bessel filter for the current data to $f_{\text{cut}}/\alpha_2 = 4.59$ or $f_{\text{cut}} = 90$ kHz. The fastest used analysing frequency is $f_{\text{ana}} = 500\,\text{Hz}$. We scale mean photo brightness λ_{b} and background noise down such that the signal-to-noise ratio of the lower integration frequency data is the same as of the high frequency data $\lambda_{\text{b}}/T_{\text{int}} = const$. We do that in order to separate the bias from the finite integration time from other effects such as a better signal to noise ratios for each integrated point. Note that we only analyzed the plot until $f_{\text{int}} = f_{\text{ana}}$. Both algorithms incur very similar bias due to the finite integration time (Fig. 12**b**). The KF (green) is more precise for high integration frequencies f_{cut}/α_2 until $f_{\text{cut}}/\alpha_2 \approx 0.08$ then the RE approach becomes more robust. Similar to Bessel-filtered current data (Fig. 11**b**) on the single parameter level the systematic deviations start early for example $f_{\text{int}} = 10$kHz for K_{21} (Fig. 12**c4**). Possibly the systematic deviations start (Fig. 12**c2**) already at $f_{\text{int}} = 50$kHz for k_{32}. The sudden increase of the Euclidean error (Fig. 12**b**) of the mean median at $f_{\text{cut}}/\alpha_2 \approx 0.2$ occurs in this case not due to f_{int} approaching f_{cut} but due to $f_{\text{int}} \lessapprox \alpha_{1,2}$ for many ligand concentrations. To show this we plot the results of the fitting of 5 different data sets without including the highest 4 ligand concentrations (red) which means the largest eigenvalues are much smaller (Fig. 12**b,c1-6**). Additionally we keep $f_{\text{int}} = const$. Though fluctuations of the posterior medians are higher, the KF becomes more robust. Note, that the fastest eigenvalues of these reduced data sets are indicated by the blue bars (Fig. 12**b, c4**). Based on the Euclidean error (Fig. 11**a** and Fig. 12**a**) the robustness of both algorithms against the cut-off frequency is compared with the robustness against the integration frequency found to be about an order of magnitude higher. That is related to a specific detail of the model used: the binding reaction, corresponds to the fastest time scales of the overall dynamic (difference between Fig. 1**b** and **c**), which is exposed by the fluorescence signal. Thus, kinetic analysis of any data should make sure that the corresponding frequency of the most dominant timescales of the time series are much slower than the respective f_{int} f_{cut} independently of the investigated algorithms.

Conclusions

We generalized the filter equations (Methods Eq. 37 , 38d, 57, 58 and 59) of the KF for analyzing the gating and binding dynamics of ligand-gated ion channels with a realistic signal-generating model for isolated patch-clamp (PC) and confocal patch-clamp fluorometry (cPCF) data including

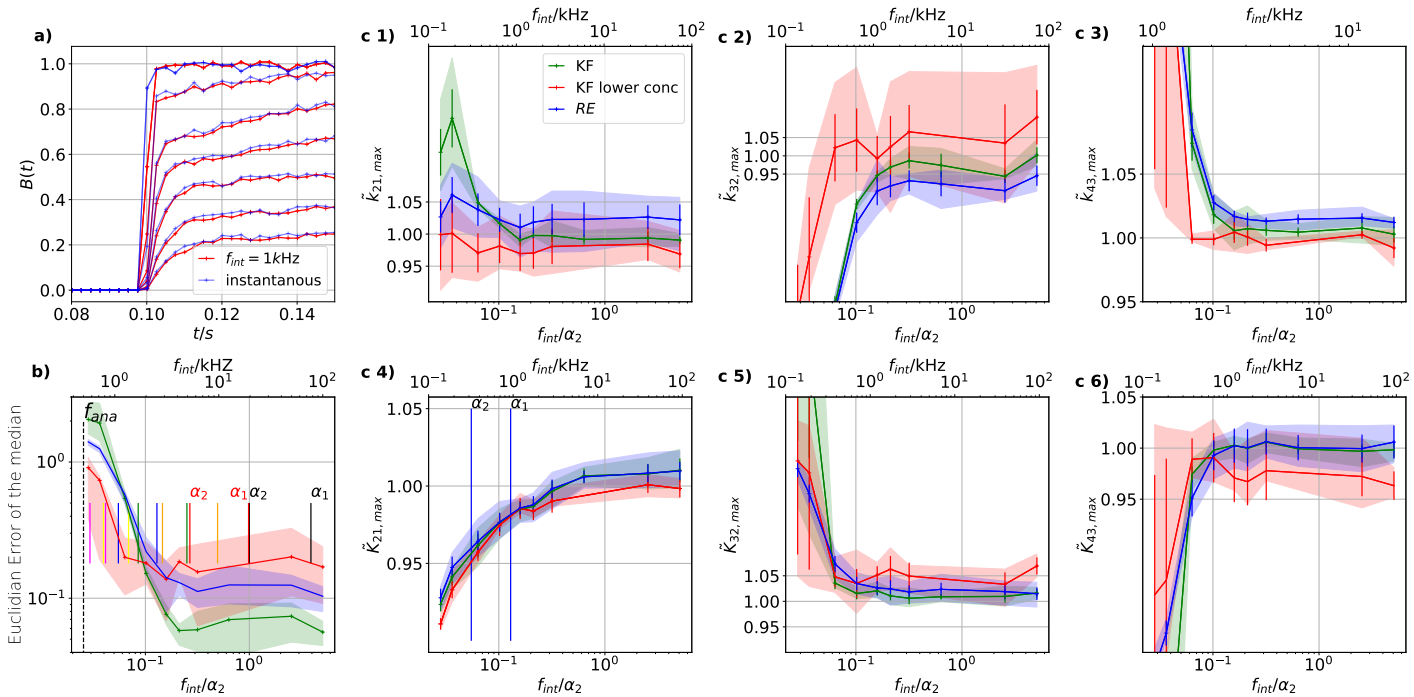


Figure 12. The influence of finite integration time of the fluorescence signal on the Accuracy and precision of the maxima of the posterior. We simulated 5 different 100kHz signals. All forms of noise were added and then the fluorescence signal was summed up using a sliding window to account for the integration time to produce one digital data point. The activation curves were then analyzed with the Bayesian filter at 125Hz and the deactivation curves at $166-500$kHz, see caption of fig. 8. We plot the $0.6-$quantile (interval between the 20th and the 80th percentile) to mimic $\pm$ one standard deviation from the mean as well as the mean of the distribution of the maxima of the posterior for different data sets. (Note, this is not equivalent to the mean and quantiles of the posterior of a single data set.) The quantiles represent the randomness of the data while the error bars indicate the standard error of the mean maximum of the posterior. Blue (RE) and green (KF) indicate the two algorithms with the standard data set while red (KF) shows examples that use only the six smallest ligand concentrations for the analysis in order to limit the highest eigenvalues. **a,** Instantaneous probing of the ligand binding (blue) compared with a probing signal which runs at $f_{\text{int}} = 1$kHz. The integrated brightness of the bound ligand is $\lambda_{\text{b}} = 5$ photons per frame. Although the red curves seem like decent measurements of the process except for the highest two shown ligand concentrations, the mean error is roughly an order of magnitude worse than for $f_{\text{int}} = 10$kHz. Note, that for visualization we plot at a higher frequency than the Kalman filter analyzed the data. **b,** Estimate of the distribution of the (Euclidean error of the mean median of the posterior) vs. the scaled integration frequency $f_{\text{int}}/\alpha_2 = 1/(\alpha_2 \cdot T_{\text{integration}})$. We use integration frequency instead of the integration time to make the plot comparable to the Bessel filter plot. The solid line is the mean median of 5 data sets of the respective KF posterior (green, red) and RE posterior (blue). The shaded areas indicate the $0.6-$quantile which visualizes the spread of the distribution of point estimates. The two fastest time scales (eigenvalues) at the highest ligand concentration are indicated by the vertical black lines, the time scales of the next lower ligand concentrations with the red vertical lines. **c 1-3,** Accuracy (bias) and precision of the maxima of the posterior $k_{ij,max}$ rates vs. the integration frequency. **c 4-6,** Accuracy and precision of the maxima of the posterior $K_{ij,max}$ of the corresponding equilibria vs. the cut-off frequency of a Bessel filter.

open-channel noise, photon-counting noise and background noise. Any other type of linear kinetic scheme (e.g. for voltage-dependent channels) and signal can be applied as long as the characteristics of the signal are sufficiently described by normal distributions. A Tutorial for Patch clamp data can be found on the git hub page https://github.com/JanMuench/Tutorial_Patch-clamp_data and for cPCF data https://github.com/JanMuench/Tutorial_Bayesian_Filter_cPCF_data.

Compared to previous attempts ***Moffatt*** (***2007***), this mathematical generalization is necessary (Fig. 3**b**) in order to use Bayesian filters on macroscopic PC or cPCF data. With our algorithm we demonstrate (Fig. 3**c** and 7) that the common assumption that for large ensembles of ion channels simpler deterministic modeling by RE approaches is on par with stochastic modeling, such as a KF, is wrong in terms of Euclidean error and uncertainty quantification (Fig. 5**a-c** and Fig. 6**a-b**). Enriching the data by fluorescence-based ligand binding reveals two regimes. In one regime the two-dimensional data increase the accuracy of the parameter estimates up to ≈ 10-fold (Fig. 4**a,c**). In the other regime of lower channel expression, enriching the data by the second observable, makes non-identified parameters to identified parameters. The second observable in cPCF data decreases the overfitting tendency (Fig. 4**a,b,d**) of the RE approach on the true process. Thus, in this regard the advantage of the KF becomes smaller. However, by exploiting Bayesian HDCV we gain a second perspective: We show for various signal-to-noise ratios (Fig. 5**a-c** and Fig. 6**a-b**) that the posterior sampled by a RE approach never covers the true values within a reasonable HDCV. Thus, the central feature of Bayesian statistics, exact uncertainty quantification by having the full posterior, is meaningless in combination with an RE approach (considering the type of data and set of signal-to-noise ratios that we tested). This even holds true for very pessimistic signal-to-noise assumptions Fig. 6**b**. If HDCVs based on an RE approach cannot be trusted, the same applies to confidence volumes based on the curvature of the likelihood. This is not the case for the KF which delivers properly shaped posteriors (Fig. 6**a-c** and Fig. 5**a-c**). Increasing the model complexity, at unchanged PC data quality (Fig. 7) shows that the RE approach displays unidentified rates even for large ion channel ensembles while our approach identified all parameters for the same data.

We also investigated the robustness of both algorithms against the cut-off frequency of a Bessel filter (Fig. 11) and showed the overall superior robustness of the KF against errors of analog filtering compared to the RE approach. Analog filtering has its limitations due to distorting the higher frequencies of the Fourier spectrum of the signal. Thus, one should let the KF sample as fast as possible, with a cut-off frequency of at least one order of magnitude higher than the sampling frequency of the KF.

Similar to the Bessel filter, the KF is more robust than the RE approach against errors due to the finite integration time. Nevertheless, it is crucial for both algorithms (Fig. 12), that the intrinsic time scales (1/eigenvalues) of the process to be analyzed are slower than the integration time of the data points. Otherwise the accuracy of the inference deteriorates.

Altogether, we demonstrated the performance of the generalized Kalman filter on ion channel data for inference of kinetic schemes. Nevertheless, our approach can approximate any other stochastic system and signal distributions of linear (pseudo-first-order) kinetics ***Sorenson and Alspach*** (***1971***). Perspectively, extensions of the Bayesian filter, e.g. by Bayesian Gaussian sum filters or similar numerically brute force concepts such as particle filters, can overcome modeling errors at low ion channel numbers or low photon fluxes.

Materials

We simulated state evolution $s(t)$ with either the software Cube ***Nicolai and Sachs*** (***2013***) for PC data or an inhouse Matlab routine (The code will be shared on request.) for cPCF data. The inhouse Matlab routine is an implementation of the Gillespie algorithm ***Gillespie Daniel T.*** (***1977***). Traces were summed up, defining the ensemble state vector $\mathbf{n}(t) := (n_1, n_2, n_3, n_4)^\top$, which counts the number of channels in each state. At first we used a 10 kHZ sampling frequency for the Gillespie algorithm but for investigating the errors induced by analog filtering the current signal and the finite integration time for each fluorescence data point the Gillespie algorithm sampled at 100

kHz. The KF,RE and Bayesian filter routines were implemented in Stan ***Carpenter et al.*** (***2017***) with the interface package PyStan and ran on a high performance computing cluster with O(100) Broadwell and SkyLake nodes. A Tutorial for Patch clamp data can be found on the git hub page https://github.com/JanMuench/Tutorial_Patch-clamp_data and for cPCF data https://github.com/JanMuench/Tutorial_Bayesian_Filter_cPCF_data.

Methods

Hereinafter, we derive the equations for our Bayesian filter for time series analysis of hidden linear chemical reaction networks (kinetic schemes). A detailed description of the experimental noise is provided in the Appendix 5.

The relation of Bayesian statistics to the Kalman filter

The following conventions are generally used: Bold symbols are used for multi-dimensional objects such as vectors or matrices. Calligraphic letters are used for (some) vectorial time series and double-strike letters are used for probabilities and probability densities. Within the Bayesian paradigm ***Hines*** (***2015***); ***Ball*** (***2016***), each unknown quantity, including model parameters $\boldsymbol{\theta}$ and time series of occupancies of hidden states $\mathfrak{N}_T = \{\mathbf{n}(t_i)\}_{i=1}^T$, are treated as random variables conditioned on observed time series data $\mathcal{Y}_T = \{\mathbf{y}(t_i)\}_{i=1}^T$. The prior $\mathbb{P}(\boldsymbol{\theta}) = \prod_j^{N_{\text{par}}} \mathbb{P}(\theta_j)$ or posterior distribution $\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T)$ encodes the available information about the parameter values before and after analysing the data, respectively. According to the Bayesian theorem, the posterior distribution

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) = \frac{1}{Z(\mathcal{Y}_T)} \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) \prod_j^{N_{\text{par}}} \mathbb{P}(\theta_j) \tag{20}$$

is a probability distribution of a parameter set $\boldsymbol{\theta}$ conditioned on $\mathcal{Y}_T$. The likelihood $\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})$ encodes the distribution of the data by modelling the intrinsic fluctuations of the protein as well as noise coming from the experimental devices. The prior provides either assumptions before measuring data or what has been learnt from previous experiments about $\boldsymbol{\theta}$. The normalization constant

$$Z(\mathcal{Y}_T) = \int \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{21}$$

ensures that the posterior is a normalized distribution. The KF is a special class of models in the family of Bayesian filters ***Ghahramani*** (***1997***), which is a generalisation of the classical KF. Due to its linear time evolution (Eq. 1), the KF is particularly useful for modeling time series data of ensembles dynamics of first order chemical networks. It delivers a set of recursive algebraic equations (Methods Eq. 28 and Eq. 32) for each time point, which allows to express the prior $\mathbb{P}(\mathbf{n}(t)|\mathcal{Y}_{t-1})$ and (after incorporating $\mathbf{y}(t)$) the posterior $\mathbb{P}(\mathbf{n}(t)|\mathcal{Y}_t)$ occupancies of hidden states $\mathbf{n}(t)$ for all t given a set of parameters $\boldsymbol{\theta}$. This means the KF solves the filtering problem (inference of $\mathfrak{N}_T$) by explicitly modeling the time evolution of $\mathbf{n}(t)$ by multivariate normal distributions. This allows us to replace $\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})$ of Eq. 20 by the expression of Eq. 9.

The Bayesian framework (as demonstrated in this article) has various properties which makes it superior to ML estimation (MLE) ***McElreath*** (***2018***). Those properties are in particular useful for the analysis of biophysical data since very often the dynamics of interest are hidden or latent in the data. Models with a hidden structure are called singular. For regular (non-singular) statistical models, maxima $\boldsymbol{\theta}_{\text{ML}}$ of the posterior or likelihood converge in distribution

$$\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\theta}_{\text{ML}} - \boldsymbol{\theta}_{\text{true}}) \sim \mathcal{N}(0, \mathbf{F}^{-1}(\boldsymbol{\theta}_{\text{true}})) \tag{22}$$

to the true value $\boldsymbol{\theta}_{\text{true}}$,where $\mathbf{F}^{-1}(\boldsymbol{\theta}_{\text{true}})$ is the inverse Fisher information matrix. Under those conditions it is justified to derive from the curvature of the likelihood at $\boldsymbol{\theta}_{\text{ML}}$ via the Cramer-Rao-bound theorem

$$\text{covar}[\boldsymbol{\theta}_{\text{ML}}] = \mathbf{F}^{-1}(\boldsymbol{\theta}_{\text{ML}}) \tag{23}$$

a confidence volume for the inferred parameters. In contrast, consider for example the type of data investigated in this study which probes the protein dynamics by current and light. Singularity means that the Fisher information matrix of a model is not invertible leading to the breakdown of the Cramer-Rao Bound theorem. Due to the breakdown, it cannot be guaranteed that even in the asymptotic limit the log-likelihood function can be approximated by a quadratic form ***Watanabe*** (***2007***). Thus, usually the MLE does not obey Eq. 22. Consequently, the posterior distribution is usually not a normal distribution either ***Watanabe*** (***2007***). Using the full posterior distribution without further approximations detects the resulting problems such as deviation from normality or non-identifiability of parameters, related to the singularity. In conclusion, the posterior is still a valid representation of parameter plausibility while ML fails.

Time evolution of a Markov Model for a single channel

In the following, we write the time t as function argument rather than a subscript. Following standard approaches, we attribute to each state of the Markov model an element of a vector space with dimension M. At a time, a channel can only be in a single state. This implies that the set of possible states is $\mathrm{S} := \{(1,0,0,\ldots),\ (0,1,0,\ldots),\ \ldots,(\ldots,0,1)\} \subset \{0,\ 1\}^M$. In the following, Greek subscripts refer to different states while Latin subscripts refer to different channels. By $\mathbf{s}(t) = \mathbf{e}_\alpha$ we specify that the channel is in state α at time t. Mathematically, $\mathbf{e}_\alpha$ stands for the α-th canonical unit Cartesian vector.

Assuming that the state transitions can be modeled by a first order Markov process, the path probability can be decomposed as the product of conditional probabilities as follows:

$$\mathbb{P}(\text{path}) = \mathbb{P}(\mathbf{s}(0), \mathbf{s}(1), \ldots, \mathbf{s}(T)) = \mathbb{P}(\mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(1) \mid \mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(2) \mid \mathbf{s}(1)) \cdots \mathbb{P}(\mathbf{s}(T) \mid \mathbf{s}(T-1)). \tag{24}$$

Markov models (MMs) and rate models are widely used for modeling molecular kinetics (Appendix 2). They provide an interpretation of the data in terms of a set of conformational states and the transition rates between these states. For exactness it remains indispensable to model the dynamics with a HMMs ***Noé et al.*** (***2013b***). The core of a hidden Markov model is a conventional Markov model, which is supplemented with a an additional observation model. We will therefore first focus on a conventional Markov model. State-to-state transitions can be equivalently described with a transition matrix $\mathbf{T}$ in discrete time or with a rate matrix $\mathbf{K}$ in continuous time, as follows:

$$\mathbf{T}_{\alpha,\beta} := \mathbb{P}(\mathbf{s}(t+1) = \mathbf{e}_\alpha \mid \mathbf{s}(t) = \mathbf{e}_\beta) = \exp(\mathbf{K} \cdot \Delta t)_{\alpha,\beta}, \tag{25}$$

where $\exp$ is the matrix exponential. We aim to infer the elements of the rate matrix $\mathbf{K}$, constituting a kinetic model or reaction network of the channel. Realizations of sequences of states can be produced by the Doob-Gillespie algorithm ***Gillespie Daniel T.*** (***1977***). To derive succinct equations for the stochastic dynamics of a system, it is beneficial to consider the time propagation of an ensemble of virtual system copies. This allows to ascribe a probability vector $\mathbf{p}(t)$ to the system, in which each element $p_\alpha(t)$ is the probability to find the system at t in state α. One can interpret the probability vector $\mathbf{p}$ as the instantaneous expectation value of the state vector $\mathbf{s}$.

$$\mathbf{p}(t) = \mathbb{E}(\mathbf{s}(t)) \tag{26}$$

The probability vector obeys the discrete-time Master equation

$$\mathbf{p}(t+1) = \mathbf{T}\mathbf{p}(t) \tag{27a}$$

$$\mathbb{E}(\mathbf{s}(t+1)) = \mathbf{T}\mathbb{E}(\mathbf{s}(t)) \tag{27b}$$

Time evolution of an ensemble of identical non-interacting channels

We model the experimentally observed system as a collection of non-interacting channels. A single channel can be modeled with a first-order MM. The same applies to the ensemble of non-interacting channels. We focus on modeling the time course of extensive macroscopic observables

Symbol	Meaning
$\boldsymbol{\theta}$	set of all unknown model parameters for which the posterior distribution is sampled
$\mathbf{n}(t)$	hidden ensemble occupancy vector of channel states in a specific patch at time t which is a continuous Markov state vector $\mathbf{n}(t) \in \mathbb{R}^M$
$\mathbf{P}(t)$	variance-covariance matrix of a hidden ensemble state $\mathbf{n}(t)$ in a specific patch at time t which contains the dispersion of the ensemble and the lacking knowledge of the algorithm about the true $\mathbf{n}(t)$
$\mathbf{T}$	transition matrix of a single channel
$\mathbf{K}$	rate matrix which is the logarithm of the transition matrix $\mathbf{T} = \exp(\mathbf{K}\Delta t)$
$\mathbf{H}$	observation matrix which projects the hidden ensemble state vector onto its mean signal.
$\mathbf{s}$	single-molecule Markov state vector
$k_{i,j}$	specific transition rate from state j to state i, $[\mathbf{K}]_{i,j} = k_{i,j}$
K_i	ratio of two transition rates i.e. an equilibrium constant
$\mathbf{y}(t)$	data point at time t
T	number of observations in a time series
$\mathcal{Y}_T$	time series of T data points, $\mathcal{Y}_T = \{\mathbf{y}(t_i)\}_{i=1}^T$
$\mathfrak{N}_T$	time series of T hidden ensemble states, $\mathfrak{N}_T = \{\mathbf{n}(t_i)\}_{i=1}^T$
$N_{\mathrm{ch},j}$	number of channels in patch number j
i	mean electrical current through a single-channel
σ_{m}^2	variance of the current including all noise from the patch and the recording system
σ_{op}^2	variance of the current noise generated by a single open-channel
λ_{b}	mean brightness of a bound ligand
λ_{Fl}	mean brightness of the fluorescence signal from bulk and bound ligands
σ_{bulk}^2	variance of the fluorescence generated by unbound ligands after subtraction of the image obtained for the reference dye
M	number of single-channel states which is the dimension of $\mathbf{n}(t) \in \mathbb{N}^M$ in the KF algorithm
N_{obs}	dimensions of the observational space
$\mathbb{F}(\mathcal{Y})$	true probability density of $\mathcal{Y}$, i.e. the true data-generating process
$\mathbb{L}(\mathcal{Y}\|\boldsymbol{\theta})$	likelihood function of the model parameters
$\mathbb{P}(\boldsymbol{\theta}\|\mathcal{Y})$	posterior distribution of the model parameters
$\mathbb{P}_{\mathrm{pred}}(\widetilde{\mathcal{Y}}\|\mathcal{Y})$	predictive distribution of the new data points $\widetilde{\mathcal{Y}}$
$\mathbb{O}(\mathbf{y}\|\mathbf{n})$	distribution of observables for a single time step
$\mathcal{N}(\cdot\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$
$\mathbb{E}[\cdot]$	mean value

Table 1. Important symbols

such as the mean current and fluorescence signals as well as their fluctuations. A central quantity is the vector $\mathbf{n}(t)$ which is the occupancy of the channel states at time t:

$$\mathbf{n}(t) = \sum_{i=1}^{N_{\text{ch}}} \mathbf{s}_i(t) \tag{28}$$

This quantity, like $\mathbf{s}(t)$, is a random variate. Unlike $\mathbf{s}(t)$, its domain is not confined to canonical unit vectors but to $\mathbf{n} \in \mathbb{N}^M$. From the linearity of Eq. 28 in the channel dimension and from the single-channel CME Eq. 27b one can immediately derive the equation for the time evolution of the mean occupancy $\overline{\mathbf{n}}(t) = \mathbb{E}[\mathbf{n}(t)]$:

$$\overline{n}_\alpha(t+1) = \sum_\beta T_{\alpha,\beta} \overline{n}_\beta(t) \tag{29}$$

with the transition matrix $\mathbf{T}$. The full distribution $\mathbb{P}(\mathbf{n}(t+1)|\mathbf{n}(t))$ is a generalized multinomial distribution. To understand the generalized multinomial distribution and how it can be constructed from the (conventional) multinomial distribution, consider the simplified case where all channels are assumed to be in the same state α. Already after one time step, the channels will have spread out over the state space. The channel distribution after one time step is parametrized by the transition probabilities in row number α of the single-channel transition matrix $\mathbf{T}$. According to the theory of Markov models, the final distribution of channels originating from state α is the multinomial distribution

$$\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha) = \mathbb{P}(n_1, \dots, n_M \mid \mathbf{n}(t) = n_\alpha \mathbf{e}_\alpha) = \frac{n_\alpha!}{n_1! \cdots n_M!} T_{1,\alpha}^{n_1} \cdots T_{M,\alpha}^{n_M} \tag{30}$$

In general, the initial ensemble will not have only one but multiple occupied channel states. Because of the independence of the channels, one can imagine each initial sub-population spreading out over the state space independently. Each sub-population with initial state α gives rise to its own final multinomial distribution that contributes $n_\beta^{(\alpha)}$ transitions into state β to the total final distribution. The total number of channels at $t+1$ in each state can then be simply found by adding the number of channels transitioning out of the different states α.

$$\mathbf{n}(t+1) = \sum_\alpha \mathbf{n}^{(\alpha)}(t+1) \tag{31}$$

Evidently, the total number of channels is conserved during propagation. The distribution of $\mathbf{n}(t+1)$, defined by Eqs. 30 and 31, is called the generalized multinomial distribution:

$$\mathbf{n}(t+1) \sim \text{general}-\text{multinomial}(\mathbf{n}(t), \mathbf{T}) \tag{32}$$

While no simple expression exists for the generalized multinomial distribution, closed form expressions for its moments can be readily derived. For large N_{ch} each $\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha)$ can be approximated by a multivariate-normal distribution such that also $\text{general}-\text{multinomial}(\mathbf{n}(t), \mathbf{T})$ has a multivariate-normal approximation. In the next section we combine the kinetics of channel ensembles with the KF by a moment expansion of the governing equations for the ensemble probability evolution.

Moment expansion of ensemble probability evolution

The multinomial distribution (30) has the following mean and covariance matrix

$$\overline{\mathbf{n}}^{(\alpha)}(t+1) = n_\alpha \mathbf{T}_{:,\alpha} \tag{33}$$

$$\mathbf{\Sigma}^{(\alpha)}(t+1) = n_\alpha \text{diag}(\mathbf{T}_{:,\alpha}) - n_\alpha \mathbf{T}_{:,\alpha,:} \mathbf{T}_{:,\alpha}^\top \tag{34}$$

where $\mathbf{T}_{:,\alpha}$ denotes the column number α of the transition matrix and $\text{diag}(\mathbf{T}_{:,\alpha})$ describes the diagonal matrix with $\mathbf{T}_{:,\alpha}$ on its diagonal. Combining Eq. 31 with Eqs. 33 and 34 we deduce the mean and variance of the generalized multinomial distribution:

$$\mathbb{E}\left[\mathbf{n}(t+1) \mid \mathbf{n}(t)\right] = \sum_\alpha n_\alpha(t) \mathbf{T}_{:,\alpha} = \mathbf{T}\mathbf{n}(t) \tag{35}$$

$$\mathrm{cov}\left[\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right] = \sum_{\alpha} n_{\alpha}(t) \left(\mathrm{diag}(\mathbf{T}_{:,\alpha}) - \mathbf{T}_{:,\alpha}\mathbf{T}_{:,\alpha}^{\top}\right) = \mathrm{diag}\left(\mathbf{T}\mathbf{n}(t)\right) - \mathbf{T}\mathrm{diag}\left(\mathbf{n}(t)\right)\mathbf{T}^{\top} \tag{36}$$

Note that Eqs. 35 and 36 are conditional expectations that depend on the random state $\mathbf{n}$ at the previous time t and not only on the previous mean $\overline{\mathbf{n}}$. To find the absolute mean, the law of total expectation is applied to Eq. 35, giving

$$\overline{\mathbf{n}}(t+1) = \mathbb{E}\left[\mathbb{E}\left[\mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] = \mathbf{T}\overline{\mathbf{n}}(t), \tag{37}$$

in agreement with the simple derivation of Eq. 29. We introduce a shorthand $\mathbf{P}(t) := \mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))$ for the absolute covariance matrix of $\mathbf{n}(t+1)$. Similarly, $\mathbf{P}(t)$ can be found by applying the law of total variance decomposition ***Weiss*** (***2005***) to Eqs. 36 and 35, giving

$$\begin{aligned}
\mathbf{P}(t+1) &= \mathbb{E}\left[\mathrm{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] + \mathrm{cov}\left[\mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t)), \mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t))\right] && \text{(38a)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathrm{cov}(\mathbf{T}\mathbf{n}(t), \mathbf{T}\mathbf{n}(t)) && \text{(38b)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathbf{T}\mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))\mathbf{T}^{\top} && \text{(38c)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathbf{T}\mathbf{P}(t)\mathbf{T}^{\top} && \text{(38d)}
\end{aligned}$$

where we have introduced the shorthand $\mathbf{P}(t) = \mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))$ in the last line. Eqs. 37, 38d are compact analytical expressions for the mean and the covariance matrix of the occupancy vector $\mathbf{n}$ at $t+1$ that depend on the mean $\overline{\mathbf{n}}$ and covariance matrix $\mathbf{P}$ at the previous time step t. Chaining these equations for different time steps $t = 0, \ldots, T$ allows to model the whole evolution of a channel ensemble. Moreover, these two equations together with the output statistics of $\mathbb{O}(\mathbf{y}|\mathbf{n}(t))$ are sufficient to formulate correction equations Eq. 59 of the KF ***Moffatt*** (***2007***); ***Anderson and Moore*** (***2012***). These equations will be used in a Bayesian context to sample the posterior distribution of the model parameters. The sampling entails repeated numerical evaluation of the model likelihood. Therefore, analytical equations for the ensemble evolution that can be quickly evaluated on a computer millions of times are indispensable. This was achieved by deriving Eqs. 37, 38d. Comparing Eq. 38d with the KF prediction equation ***Anderson and Moore*** (***2012***) for $\mathbf{P}(t)$ we obtain the state-dependent covariance matrix of Eq. 3 as

$$\mathbf{Q}(\mathbf{T}, \overline{\mathbf{n}}(t)) = \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} \tag{39}$$

In the following section on properties of measured data and the KF, we no longer need to refer to the random variate $\mathbf{n}(t)$. All subsequent equations can be formulated by only using the mean hidden state $\overline{\mathbf{n}}(t)$ and the variance-covariance matrix of the hidden state $\mathbf{P}(t)$. We therefore drop the overbar in $\overline{\mathbf{n}}(t)$ so that the symbol $\mathbf{n}(t)$ refers from now on to the mean hidden state.

Modeling simultaneous measurement of current and fluorescence

In the following, we develop a model for the conditional observation distribution $\mathbb{O}(\mathbf{y}|\mathbf{n}(t))$ (Appendix 5 for experimental details). Together with the hidden ensemble dynamics this will enable us to derive the output statistics of the KF (see, below). Let $\mathbf{y}(t)$ be the vector of all observations at t. Components of the vector are the ion current and fluorescence intensity.

$$\mathbf{y}(t) = \begin{pmatrix} \text{fluorescence intensity}(t) \\ \text{ion current}(t) \end{pmatrix} = \begin{pmatrix} y_{\mathrm{flu}}(t) \\ y_{\mathrm{curr}}(t) \end{pmatrix} \tag{40}$$

As outlined in the introduction part, in Eq. 4 we model the observation by using a conditional probability distribution $\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t))$ that only depends on the mean hidden state $\mathbf{n}(t)$, as well as on fixed channel and other measurement parameters. $\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t))$ is modeled as a multivariate normal distribution with mean $\mathbf{H}\mathbf{n}(t)$ and variance-covariance matrix $\mathbf{\Sigma}(t)$, that can in general depend on the mean state vector $\mathbf{n}(t)$ (much like the covariance matrix of the kinetics in Eq. 38d). The observation

matrix $\mathbf{H} \in \mathbb{R}^{N_{\text{obs}} \times M}$ projects the hidden state vector $\mathbf{n}(t)$ onto $\mathbf{Hn}(t) \in \mathbb{R}^{N_{\text{obs}}}$, the observation space. The observation distribution is

$$\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t)) = \mathcal{N}\left(\mathbf{y}(t)|\mathbf{Hn}(t), \mathbf{\Sigma}(\mathbf{n}(t))\right) \Leftrightarrow \mathbf{y}(t) = \mathbf{Hn}(t) + \boldsymbol{\nu}(t). \tag{41}$$

This measurement model is very flexible and allows to include different types of signals and error sources arising from both the molecules and the instruments. A summary of the signals and sources of measurement error and their contributions to the parameters of $\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t))$ is provided by Tab. 2. Below we address the two types of signals and four noise sources one by one. For

	ion current		fluorescence	
	current signal	measurement noise	fluorescence signal	background fluorescence
signaling states	open state	-	ligand-bound states	-
error term	open-channel noise	measurement noise	photon counts	bulk noise
affected signal	current	current	fluorescence	fluorescence
distribution	normal$(in_4, \sigma_{\text{op}}^2 n_4)$	normal$(0, \sigma_{\text{m}}^2)$	Poisson$(\lambda_b n_i(t))$	scaled Skellam
contribution to $\mathbf{H}$	$H_{2,4} = i$	-	$\mathbf{H}_{1,:} = (0, \lambda_b, 2\lambda_b, 2\lambda_b)$	-
contribution to $\mathbf{\Sigma}$	$\Sigma_{2,2} = \sigma_{\text{op}}^2 n_4(t)$	$\Sigma_{2,2} = \sigma_{\text{m}}^2$	$\Sigma_{1,1} = (0, \lambda_b, 2\lambda_b, 2\lambda_b)\mathbf{n}(t)$	$\Sigma_{1,1} = \sigma_{\text{back}}^2$

Table 2. Summary of signals and noise sources for the exemplary CCCO model with the closed states $\alpha = 1, 2, 3$ and the open state $\alpha = 4$. The observed space is two-dimensional with y_{Fl} = fluorescence and y_I = ion current. The fluorescence signal is assumed to be derived from the difference of two spectrally different Poisson distributed fluorescent signals. That procedure results in a scaled Skellam distribution of the noise.

this we decompose the observation matrix and the observation noise covariance matrix into the individual terms:

$$\mathbf{H} = \mathbf{H}_{\text{I}} + \mathbf{H}_{\text{binding}} \tag{42}$$

$$\mathbf{\Sigma}(t) = \mathbf{\Sigma}_{\text{open}}(t) + \mathbf{\Sigma}_{\text{meas.}} + \mathbf{\Sigma}_{\text{binding}}(t) + \mathbf{\Sigma}_{\text{back}} \tag{43}$$

In the following, we report the individual matrices for the exemplary CCCO model with one open state $\alpha = 4$ and three closed states $\alpha = 1, 2, 3$. Matrices can be constructed analogously for the other models. For the definition of $\mathbf{\Sigma}_{\text{back}}$ refer to (Appendix 5).

Macroscopic current and open-channel noise

We model the current and the intrinsic fluctuations of the open-channel state $\mathbf{s} = \mathbf{e}_4$ (the open channel noise) by a state-dependent normal distribution with mean $in_4(t)$ where $n_4(t)$ is the number of channels in the open state at t and i is the single-channel current. The additional variance of the single-channel current is described by σ_{open}^2. The sum of the instrumental noise of the experimental setup and the open channel noise is modeled as uncorrelated (white) normally distributed noise with the mean $\mathbb{E}[\nu_{\text{I}}(t)] = 0$ and variance $\mathbb{E}[\nu_{\text{I}}^2(t)] = \sigma_{\text{op}}^2 n_4(t) + \sigma_{\text{m}}^2$. By making the open-channel noise dependent on the hidden state population $n_4(t)$, we fully take advantage of the flexibility of Bayesian filters which admits an (explicitly or implicitly) time-dependent observation model. By tabulating the parameters of the two normal distributions into $\mathbf{H}$ and $\mathbf{\Sigma}$, we obtain

$$\mathbf{H}_{\text{I}} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \end{pmatrix} \tag{44}$$

$$\mathbf{\Sigma}_{\text{open}}(t) + \mathbf{\Sigma}_{\text{meas.}} := \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{\text{op}}^2 n_4(t) + \sigma_{\text{m}}^2 \end{pmatrix} \tag{45}$$

One can now ask for the variance of a data point $y(t)$ given the epistemic and aleatory uncertainty of $\mathbf{n}(t)$ encoded by $\mathbf{P}(t)$ in Eq. 38d. By using the law of total variance the signal variance follows as:

$$\mathrm{var}(\mathbf{y}(t)) = \mathbb{E}[\mathrm{var}[\mathbf{y}(t)|\mathbf{n}(t)]] + \mathrm{var}[\mathbb{E}[\mathbf{y}(t)|\mathbf{n}(t)]] \tag{46a}$$

$$= \mathbb{E}[\sigma_{\mathrm{op}}^2 n_4(t) + \sigma_{\mathrm{m}}^2] + \mathrm{var}[\mathbf{H}_{\mathrm{l}}\mathbf{n}(t)] \tag{46b}$$

$$= \sigma_{\mathrm{op}}^2 \mathbb{E}[n_4(t)] + \sigma_{\mathrm{m}}^2 + (\mathbf{H}_{\mathrm{l}}\mathbf{P}(t)\mathbf{H}_{\mathrm{l}}^{\top})_{2,2} \tag{46c}$$

See, Appendix 6 for further details.

Fluorescence and photon-counting noise

The statistics of photon counts in the fluorescence signal are described by a Poisson distribution with emission rate λ_{Fl}

$$y_{\mathrm{Fl}}(t) \sim \mathrm{Pois}(\lambda_{\mathrm{Fl}}(t)). \tag{47}$$

The total emission rate λ_{Fl} can be modeled as a weighted sum of the specific emission rates λ_b of each ligand class $\{0, 1, 2\}$. The weights are given by the stoichiometric factors which reflect the number of bound ligands. In order to cast the Poisson distribution into the functional form of the observation model (Eq. 41), we invoke the central limit theorem to approximate

$$y_{\mathrm{Fl}} \sim \mathrm{Pois}(\lambda_{\mathrm{Fl}}) \approx \mathcal{N}(\lambda_{\mathrm{Fl}}(t), \lambda_{\mathrm{Fl}}(t)) \tag{48}$$

The larger λ_{Fl} the better is the approximation. We assume, that the confocal volume is equally illuminated. For our model of ligand fluorescence, we assume for a moment that there is no signal coming from ligands in the bulk. We will drop this assumption in the next section. With these assumptions, we arrive at the following observation matrix

$$\mathbf{H}_{\mathrm{binding}} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{49}$$

The matrix $\mathbf{H}$ aggregates the states into two conductivity classes: non-conducting and conducting and three different fluorescence classes. The first element $(\mathbf{Hn})_1$ is the mean fluorescence $\lambda_{\mathrm{Fl}}(t) = \lambda_b[n_2(t) + 2(n_3(t) + n_4(t))]$. The variance-covariance matrix $\mathbf{\Sigma}_{\mathrm{binding}}$ can be derived along the same lines using Eq. 48. We find

$$\mathbf{\Sigma}_{\mathrm{binding}}(t) := \begin{pmatrix} (\mathbf{Hn}(t))_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{50}$$

Under these assumptions the observation matrix can be written as follows

$$\mathbf{H} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & i \end{pmatrix} \tag{51}$$

Output statistics of a Kalman Filter with two-dimensional state-dependent noise

Now simultaneously measured current and fluorescence data $\mathbf{y} \in \mathbb{R}^2$, obtained by cPCF, are modeled. Thus, the observation matrix fulfills $\mathbf{H} \in \mathbb{R}^{2\times M}$. One can formulate the observation distribution as

$$\mathbf{y}(t) = \mathbf{Hn}(t) + \boldsymbol{\nu}_{\mathrm{m}}(t) + \begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix} \Leftrightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{Hn}(t), \mathbf{\Sigma}(t)). \tag{52}$$

The vector $\boldsymbol{\nu}_{\mathrm{m}}$ denotes the experimental noise, with $\mathbb{E}[\boldsymbol{\nu}_{\mathrm{m}}] = 0$ and variance given by the diagonal matrix $\mathbf{\Sigma}_{\mathrm{meas}} + \mathbf{\Sigma}_{\mathrm{back}}$. The second noise term arises from Poisson-distributed photon counting statistics and the open-channel noise. It has the properties

$$\mathbb{E}\left[\begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}\right] = 0 \tag{53}$$

and

$$\mathrm{cov}\left(\begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}, \begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}\right) = \mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{binding}}(t). \tag{54}$$

The matrix $\mathbf{\Sigma}$ is a diagonal matrix. To derive the covariance matrix $\mathrm{cov}(\mathbf{y}(t))$ we need to additionally calculate $\mathrm{var}(y_{\mathrm{fluo}}(t))$ and $\mathrm{cov}(y_{\mathrm{fluo}}(t), y_{\mathrm{patch}}(t))$. By the same arguments as above we get

$$\mathrm{var}[y_{\mathrm{fluo}}(t)] = \mathbb{E}[\mathrm{var}(y(t)|\mathbf{n}(t))] + \mathrm{var}[\mathbb{E}(y(t)|\mathbf{n}(t)] \tag{55a}$$

$$= \mathbb{E}[\sigma^2_{\mathrm{back}} + (\mathbf{Hn}(t))_1] + \mathrm{var}(\mathbf{Hn}(t)) \tag{55b}$$

$$= \sigma^2_{\mathrm{back}} + (\mathbf{Hn}(t))_1 + (\mathbf{HP}(t))\mathbf{H}^{\top})_{1,1} \tag{55c}$$

The cross terms can be calculated by using the law of total covariance

$$\mathrm{cov}(y_{\mathrm{patch}}, y_{\mathrm{fluo}}) = \mathbb{E}[\mathrm{cov}(y_{\mathrm{patch}}, y_{\mathrm{fluo}}|\mathbf{n})] + \mathrm{cov}(\mathbb{E}(y_{\mathrm{patch}}|\mathbf{n}), \mathbb{E}(y_{\mathrm{fluo}}|\mathbf{n})) \tag{56a}$$

$$= 0 + \mathrm{cov}(\mathbf{H}_{2,:}\mathbf{n}, \mathbf{H}_{1,:}\mathbf{n}) \tag{56b}$$

$$= \mathbf{H}_{2,:}\,\mathrm{cov}(\mathbf{n}, \mathbf{n})\mathbf{H}^{\top}_{1,:} = \mathbf{H}_{2,:}\mathbf{P}(t)\mathbf{H}^{\top}_{1,:} \tag{56c}$$

yielding the matrix

$$\mathrm{cov}(\mathbf{y}, \mathbf{y}) = \mathbf{HP}(t)\mathbf{H}^{\top} + \mathbf{\Sigma}(t) \tag{57}$$

We assumed that the Poisson distribution is well captured by the normal approximation. In cPCF data the ligand binding to only a sub-ensemble of the channels is monitored, which we assume to represent the conducting ensemble such that $N_{\mathrm{ch,FL}} = N_{\mathrm{ch,I}}$. For real data further refinement might be necessary to model the randomness of the sub-ensemble in the summed voxels. With the time evolution equations for the mean (Eq. 35) and for the covariance matrix (Eq. 38d) as well as with the expressions for the signal variance we possess all parameters that are needed in the correction equation of the KF ***Kalman*** (***1960***); ***Anderson and Moore*** (***2012***).

The correction step

For completeness we write down the correction step of the KF though its derivation can be found in ***Chen et al.*** (***2003***); ***Anderson and Moore*** (***2012***); ***Moffatt*** (***2007***). The mean ensemble state $\mathbf{n}(t)$ is corrected by the current data point

$$\mathbf{n}_{\mathrm{posterior}}(t) = +\mathbf{n}_{\mathrm{prior}}(t) + \mathbf{K}_{\mathrm{Kal}}\left(\mathbf{y}(t) - \mathbf{Hn}_{\mathrm{prior}}(t)\right) \tag{58}$$

Where Kalman gain matrix $\mathbf{K}_{\mathrm{Kal}} := \mathbf{P}(t)_{\mathrm{prior}}\mathbf{H}^{\top}\mathbf{\Sigma}^{-1}$ evaluates the intrinsic noise against the experimental noise. How precise are my model predictions about $\mathbf{n}(t)$ compared with the information gained about $\mathbf{n}(t)$ by measuring $\mathbf{y}(t)$. The covariance $\mathbf{P}(t)$ of the ensemble state $\mathbf{n}(t)$ is corrected by

$$\mathbf{P}_{\mathrm{posterior}}(t) = \mathbf{P}_{\mathrm{prior}}(t) - \mathbf{K}_{\mathrm{Kal}}\left(\mathbf{HP}_{\mathrm{prior}}(t)\mathbf{H} + \mathbf{\Sigma}(t)\right)\mathbf{K}^{\top} \tag{59}$$

Eq. 58,59, 37 and 38d form the filtering equations which summarize the algorithm. One initialises the first $\mathbf{n}(0)$ and $\mathbf{P}(0)$ and with an equilibrium assumption.

Microscopic-reversibility as a hierarchical prior

We applied microscopic-reversibility ***Colquhoun et al.*** (***2004***) by a hierarchical prior distribution. Usually, micro-reversibility is strictly enforced by setting the product of the rates of the clockwise loop $k_1, k_2, k_3 k_4$ equal to the anti-cockwise loop k_5, k_6, k_7, k_8 and then solving for the desired rate parameter to be replaced. This means that the classical approach can be described by drawing the resulting rate from a dirac delta distribution prior with

$$k_1 \sim \delta(k_1 - \frac{k_5 k_6 k_7 k_8}{k_2 k_3 k_4}) \tag{60}$$

Following Eq. 60 we can model microscopic-reversibility with any hierarchical prior distribution whose limit for a vanishing variance is Eq. 60. For mathematical convenience we defined the hierarchical prior by a beta distribution close to uniform distribution

$$k_1^\star \sim \mathrm{beta}(100.01, 100.01) \tag{61}$$

and by rescaling and adding an offset

$$k_1 = \frac{k_5 k_6 k_7 k_8}{k_2 k_3 k_4} \cdot 0.995 + 0.01 \cdot k_1^\star \tag{62}$$

we derived a conditional prior which allows at maximum a ± 0.005 relative deviation from the strict microscopic-reversibility. The ± 0.005 micro-reversibility constraint is applied in (Fig. 7**b-d**). In this way, one could model or even test possible small violation of microscopic-reversibility if smaller beta parameters such as $\mathrm{beta}(1, 1)$ would be chosen.

Acknowledgments

The authors are grateful to Dr. E. Schulz for designing a software to simulate channel activity in ensemble patches and to Dr. Thomas Eick for performing the simulations. F.P. acknowledges funding from the Yen Post-Doctoral Fellowship in Interdisciplinary Research and from the National Cancer Institute of the National Institutes of Health (NIH) through Grant CAO93577. The authors are also indebted to M. Habeck and I. Schroeder for comments on the manuscript, to M. Bücker for help with the computer cluster at the Friedrich Schiller University Jena, and to F. Noé, R. Blunck, G. Mirams and S. Presse for helpful discussions. This work was supported by the Research Unit 2518 DynIon (Project P2) and the TRR 166 ReceptorLight (Project A5) of the Deutsche Forschungsgemeinschaft to K.B.

References

Alcantara P, Cardenas LM, Swillens S, Scroggs RS. Reduced Transition between Open and Inactivated Channel States Underlies 5HT Increased INa+ in Rat Nociceptors. Biophys J. 2002; 83(1):5–21. doi: 10.1016/S0006-3495(02)75146-1.

Anderson BD, Moore JB. Optimal filtering. Courier Corporation; 2012.

Auger-Methe Marie, Field Chris, Albertsen Christoffer M , Derocher Andrew E , Lewis Mark A , Jonsen Ian D , Mills Flemming Joanna. State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Scientific Reports. 2016 may; 6:26677. doi: 10.1038/srep26677; 10.1038/srep26677.

Ball F. MCMC for Ion-Channel Sojourn-Time Data: A Good Proposal. Biophys J. 2016; 111(2):267–268. doi: 10.1016/j.bpj.2016.02.042.

Ball F G CYKJB, A O. Bayesian inference for ion–channel gating mechanisms directly from single–channel recordings, using Markov chain Monte Carlo. Proc R Soc Lond A. 1999; .

Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:170102434. 2017; .

Biskup C, Kusch J, Schulz E, Nache V, Schwede F, Lehmann F, Hagen V, Benndorf K. Relating ligand binding to activation gating in CNGA2 channels. Nature. 2007 feb; 446:440. doi: 10.1038/nature05596; 10.1038/nature05596.

Brown C, Dalal R, Hebert B, Digman M, Horwitz A, Gratton E. Raster image correlation spectroscopy (RICS) for measuring fast protein dynamics and concentrations with a commercial laser scanning confocal microscope. Journal of microscopy. 2008; 229(1):78–91.

Bruening-Wright A, Elinder F, Larsson HP. Kinetic Relationship between the Voltage Sensor and the Activation Gate in spHCN Channels. J of Gen Physiol. 2007; 130(1):71–81. doi: 10.1085/jgp.200709769.

Calderhead B, Epstein M, Sivilotti L, Girolami M. 13. In: Bayesian Approaches for Mechanistic Ion Channel Modeling Totowa, NJ: Humana Press; 2013. p. 247–272.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. Journal of statistical software. 2017; 76(1):1–32.

Celentano JJ, Hawkes AG. Use of the Covariance Matrix in Directly Fitting Kinetic Parameters: Application to GABAA Receptors. Biophys J. 2004; 87(1):276–294. doi: 10.1529/biophysj.103.036632.

Chang G. On kalman filter for linear system with colored measurement noise. Journal of Geodesy. 2014; 88(12):1163–1170.

Chen Z, et al. Bayesian filtering: From Kalman filters to particle filters, and beyond. Statistics. 2003; 182(1):1–69.

Chung SH, Moore JB, Xia L, Premkumar L, Gage PW. Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. Philos T of the Roy Soc of Lond Series B Bio Sci. 1990; 329(1254):265–285.

Clancy CE, Rudy Y. Cellular consequences of HERG mutations in the long QT syndrome: precursors to sudden cardiac death. Cardiovascular research. 2001; 50(2):301–313.

Colquhoun D, Dowsland KA, Beato M, Plested AJ. How to impose microscopic reversibility in complex reaction mechanisms. Biophysical journal. 2004; 86(6):3510–3518.

Colquhoun D, Hawkes GA. A Q-Matrix Cookbook. In: Sakmann B., Neher E. (eds) Single-Channel Recording. Springer, Boston, MA; (1995).

Colquhoun D, Hawkes GA. The Principles of the Stochastic Interpretation of Ion-Channel Mechanisms. In: Sakmann B., Neher E. (eds) Single-Channel Recording. Springer, Boston, MA; (1995).

Colquhoun D, Hawkes GA, Bernard K. On the stochastic properties of single ion channels. P of the Roy Soc of London Series B Biological Sciences. 1981; 211(1183):205–235. doi: 10.1098/rspb.1981.0003.

Colquhoun D, Hawkes GA, Bernard K. Relaxation and fluctuations of membrane currents that flow through drug-operated channels. Proc R Soc Lond B. 1997; .

Deuflhard P, Weber M. Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl. 2005; 398(Supplement C):161–184. doi: 10.1016/j.laa.2004.10.026, special Issue on Matrices and Mathematical Biology.

Epstein M, Calderhead B, Girolami MA, Sivilotti LG. Bayesian Statistical Inference in Ion-Channel Models with Exact Missed Event Correction. Biophys J. 2016; 111(2):333–348. doi: 10.1016/j.bpj.2016.04.053.

Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science. 1991; 254(5038):1598–1603. doi: 10.1126/science.1749933.

Fredkin DR, Rice JA. Maximum likelihood estimation and identification directly from single-channel recordings. P of the Roy Soc of London Series B: Biological Sciences. 1992; 249(1325):125–132.

Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. J of the Roy Statistical Soc: Series A (Statistics in Soc). 2019; 182(2):389–402.

Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2014 Nov; 24(6):997–1016. doi: 10.1007/s11222-013-9416-2.

Gelman A, Lee D, Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization. J of Educational and Behavioral Statistics. 2015; 40(5):530–543.

Gelman A, Rubin DB. A single series from the Gibbs sampler provides a false sense of security. Bayesian statistics. 1992; 4:625–631.

Gelman A, Rubin DB, et al. Inference from iterative simulation using multiple sequences. Statistical science. 1992; 7(4):457–472.

Ghahramani Z. Learning dynamic Bayesian networks. In: International School on Neural Networks, Initiated by IIASS and EMFCSC Springer; 1997. p. 168–197.

Gillespie CS, Golightly A. Bayesian inference for the chemical master equation using approximate models. In: Ninth International Workshop on Computational Systems Biology, WCSB 2012, June 4-6, Ulm, Germany, vol. 4; 2012. p. 23.

Gillespie Daniel T . Exact stochastic simulation of coupled chemical reactions. The J of Phys Chem. 1977; 81(25):2340–2361.

Gin E, Falcke M, Wagner LE, Yule DI, Sneyd J. Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors. J of Theoretical Biology. 2009; 257(3):460–474.

Goldschen-Ohm MP, Wagner DA, Petrou S, Jones MV. An epilepsy-related region in the GABAA receptor mediates long-distance effects on GABA and benzodiazepine binding sites. Mol Pharmacol. 2010; 77(1):35–45.

de Gunst MM, Künsch H, Schouten J. Statistical analysis of ion channel data using hidden Markov models with correlated state-dependent noise and filtering. Journal of the American Statistical Association. 2001; 96(455):805–815.

Hines KE. A Primer on Bayesian Inference for Biophysical Systems. Biophys J. 2015; 108(9):2103–2113.

Hines KE, Bankston JR, Aldrich RW. Analyzing Single-Molecule Time Series via Nonparametric Bayesian Inference. Biophys J. 2015; 108(3):540–556. doi: 10.1016/j.bpj.2014.12.016.

Hines KE, Middendorf TR, Aldrich RW. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. The J of General Physiology. 2014; 143(3):401–416. doi: 10.1085/jgp.201311116.

Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J of Machine Learning Research. 2014; 15(1):1593–1623.

Horn R, Lange K. Estimating kinetic constants from single channel data. Biophys J. 1983; 43(2):207–223. doi: 10.1016/S0006-3495(83)84341-0.

Hwang Y, Kim JS, Kweon IS. Sensor noise modeling using the Skellam distribution: Application to the color edge detection. In: 2007 IEEE conference on computer vision and pattern recognition IEEE; 2007. p. 1–8.

Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. J Math Biol. 2007 Jan; 54(1):1–26. doi: 10.1007/s00285-006-0034-x.

Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. J of mathematical biology. 2007; 54(1):1–26.

Jaynes ET, Kempthorne O. Confidence intervals vs Bayesian intervals. In: Foundations of probability theory, statistical inference, and statistical theories of science Springer; 1976.p. 175–257.

Jeffreys H. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences. 1946; 186(1007):453–461.

Joshi M, Seidel-Morgenstern A, Kremling A. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. Metabolic engineering. 2006; 8(5):447–455.

Kalman RE. A new approach to linear filtering and prediction problems. J of basic Engineering. 1960; 82(1):35–45.

Kalstrup T, Blunck R. Dynamics of internal pore opening in KV channels probed by a fluorescent unnatural amino acid. Proc of the Nat Academy of Sci. 2013; 110(20):8272–8277.

Kalstrup T, Blunck R. S4–S5 linker movement during activation and inactivation in voltage-gated K+ channels. Proc of the Nat Academy of Sci. 2018; 115(29):E6751–E6759.

Kurtz TG. The relationship between stochastic and deterministic models for chemical reactions. The J of Chem Phys. 1972; 57(7):2976–2978.

Kusch J, Biskup C, Thon S, Schulz E, Nache V, Zimmer T, Schwede F, Benndorf K. Interdependence of Receptor Activation and Ligand Binding in HCN2 Pacemaker Channels. Neuron. 2010; 67(1):75–85. doi: 10.1016/j.neuron.2010.05.022.

Kusch J, Thon S, Schulz E, Biskup C, Nache V, Zimmer T, Seifert R, Schwede F, Benndorf K. How subunits cooperate in cAMP-induced activation of homotetrameric HCN2 channels. Nature Chemical Biology. 2011 dec; 8:162. doi: 10.1038/nchembio.747; 10.1038/nchembio.747.

Lei CL, Clerx M, Whittaker DG, Gavaghan DJ, de Boer TP, Mirams GR. Accounting for variability in ion current recordings using a mathematical model of artefacts in voltage-clamp experiments. Philosophical Transactions of the Royal Society A. 2020; 378(2173):20190348.

McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC; 2018.

Middendorf TR, Aldrich RW. The structure of binding curves and practical identifiability of equilibrium ligand-binding parameters. J of General Physiology. 2017; 149(1):121–147.

Milescu LS, Akk G, Sachs F. Maximum Likelihood Estimation of Ion Channel Kinetics from Macroscopic Currents. Biophys J. 2005; 88(4):2494–2515. doi: 10.1529/biophysj.104.053256.

Moffatt L. Estimation of Ion Channel Kinetics from Fluctuations of Macroscopic Currents. Biophys J. 2007; 93(1):74–91. doi: 10.1529/biophysj.106.101212.

Munsky B, Trinh B, Khammash M. Listening to the noise: random fluctuations reveal gene network parameters. Mol Syst Biol. 2009; 5(1):318. doi: 10.1038/msb.2009.75.

Navarro MA, Salari A, Milescu M, Milescu LS. Estimating kinetic mechanisms with prior knowledge II: Behavioral constraints and numerical tests. Journal of General Physiology. 2018; 150(2):339–354.

Neal RM, et al. MCMC using Hamiltonian dynamics. Handbook of markov chain monte carlo. 2011; 2(11):2.

Neher E, Sakmann B. Single-channel currents recorded from membrane of denervated frog muscle fibres. Nature. 1976; 260(5554):799–802. doi: 10.1038/260799a0.

Nicolai C, Sachs F. Solving ion channel kinetics with the QuB software. Biophysical Reviews and Letters. 2013; 8(03n04):191–211.

Noé F, Doose S, Daidone I, Löllmann M, Sauer M, Chodera JD, Smith JC. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. P Natl Acad Sci USA. 2011; 108(12):4822–4827. doi: 10.1073/pnas.1004646108.

Noé F, Wu H, Prinz JH, Plattner N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. J Chem Phys. 2013; 139(18):184114. doi: 10.1063/1.4828816.

Noé F, Wu H, Prinz JH, Plattner N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. The J of Chem Phys. 2013; 139(18):184114. doi: 10.1063/1.4828816.

Oyrer J, Maljevic S, Scheffer IE, Berkovic SF, Petrou S, Reid CA. Ion channels in genetic epilepsy: from genes and mechanisms to disease-targeted therapies. Pharmacol Rev. 2018; 70(1):142–173.

Qin F, Auerbach A, Sachs F. Hidden Markov Modeling for Single Channel Kinetics with Filtering and Correlated Noise. Biophys J. 2000; 79(4):1928–1944. doi: 10.1016/S0006-3495(00)76442-3.

Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc of the IEEE. 1989 Feb; 77(2):257–286. doi: 10.1109/5.18626.

Rosales R, Stark JA, Fitzgerald WJ, Hladky SB. Bayesian Restoration of Ion Channel Records using Hidden Markov Models. Biophys J. 2001; 80(3):1088–1103. doi: 10.1016/S0006-3495(01)76087-0.

Rosales RA. MCMC for hidden Markov models incorporating aggregation of states and filtering. Bull Math Biol. 2004 Sep; 66(5):1173–1199. doi: 10.1016/j.bulm.2003.12.001.

Sakmann B. Single-channel recording. Springer Science & Business Media; 2013.

Salari A, Navarro MA, Milescu M, Milescu LS. Estimating kinetic mechanisms with prior knowledge I: Linear parameter constraints. Journal of General Physiology. 2018; 150(2):323–338.

Sgouralis I, Pressé S. Icon: an adaptation of infinite hmms for time traces with drift. Biophysical journal. 2017; 112(10):2117–2126.

Siekmann I, Fackrell M, Crampin EJ, Taylor P. Modelling modal gating of ion channels with hierarchical Markov models. Proc of the Roy Soc A- Math Phys. 2016; 472(2192):20160122.

Siekmann I, Sneyd J, Crampin EJ. MCMC Can Detect Nonidentifiable Models. Biophys J. 2012; 103(11):2275–2286. doi: 10.1016/j.bpj.2012.10.024.

Siekmann I, Wagner LE, Yule D, Fox C, Bryant D, Crampin EJ, Sneyd J. MCMC Estimation of Markov Models for Ion Channels. Biophys J. 2011; 100(8):1919–1929. doi: 10.1016/j.bpj.2011.02.059.

Silberberg SD, Magleby KL. Preventing errors when estimating single channel properties from the analysis of current fluctuations. Biophysical journal. 1993; 65(4):1570–1584.

Sorenson HW, Alspach DL. Recursive Bayesian estimation using Gaussian sums. Automatica. 1971; 7(4):465–479.

Stepanyuk A, Borisyuk A, Belan P. Maximum likelihood estimation of biophysical parameters of synaptic receptors from macroscopic currents. Frontiers in Cellular Neuroscience. 2014; 8:303.

Stepanyuk AR, Borisyuk AL, Belan PV. Efficient Maximum Likelihood Estimation of Kinetic Rate Constants from Macroscopic Currents. PloS One. 2011 12; 6(12):1–18. doi: 10.1371/journal.pone.0029731.

Taraska JW, Puljung MC, Olivier NB, Flynn GE, Zagotta WN. Mapping the structure and conformational movements of proteins with transition metal ion FRET. Nature methods. 2009; 6(7):532.

Taraska JW, Zagotta WN. Structural dynamics in the gating ring of cyclic nucleotide–gated ion channels. Nature structural & molecular biology. 2007; 14(9):854.

Van Kampen NG. Stochastic processes in physics and chemistry, vol. 1. Elsevier; 1992.

Vats D, Knudson C. Revisiting the gelman-rubin diagnostic. arXiv preprint arXiv:181209384. 2018; .

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC, et al. Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. Bayesian Analysis. 2021; .

Venkataramanan L, Sigworth F. Applying hidden Markov models to the analysis of single ion channel activity. Biophys J. 2002; 82(4):1930–1942.

Verkerk AO, Wilders R. Pacemaker activity of the human sinoatrial node: effects of HCN4 mutations on the hyperpolarization-activated current. Europace. 2014; 16(3):384–395.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020; 17:261–272. doi: 10.1038/s41592-019-0686-2.

Wang W, Feng X, Xuhui Z, Jing Y, Ming Y, Jiuping D. Optimal Estimation of Ion-Channel Kinetics from Macroscopic Currents. PLoS One. 2012 04; 7(4):1–12. doi: 10.1371/journal.pone.0035208.

Watanabe S. Almost All Learning Machines are Singular. In: 2007 IEEE Symposium on Foundations of Computational Intelligence; 2007. p. 383–388. doi: 10.1109/FOCI.2007.371500.

Weiss NA. A Course in Probability. Boston, U.S.A.: Addison-Wesley; 2005.

Wu S, Vysotskaya ZV, Xu X, Xie C, Liu Q, Zhou L. State-Dependent cAMP Binding to Functioning HCN Channels Studied by Patch-Clamp Fluorometry. Biophys J. 2011; 100(5):1226–1232. doi: 10.1016/j.bpj.2011.01.034.

Wulf M, Pless SA. High-Sensitivity fluorometry to resolve ion channel conformational dynamics. Cell reports. 2018; 22(6):1615–1626.

Zheng J, Zagotta WN. Gating rearrangements in cyclic nucleotide-gated channels revealed by patch-clamp fluorometry. Neuron. 2000; 28(2):369–374.

Zwickl DJ, Holder MT. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. Systematic Biology. 2004; 53(6):877–888.

Appendix 1

Efficiency of the Hamiltonian Monte Carlo sampler

Bayesian posterior sampling requires a fast and efficient sampler which requires in turn a fast evaluation of the likelihood and prior distribution. Optimally, the calculation time should increase as little as possible with the amount of data. This is why we chose the KF framework and parallelized the processing of a time trace across many CPUs. Compared to the maximization of the likelihood, sampling from the posterior requires approximately one to multiple orders of magnitude times more evaluations of the likelihood function to have a decent representation of the posterior. For example, in ***Moffatt*** (***2007***) both algorithms need roughly 10^2 to 10^3 evaluations of the likelihood function to converge. This is only a small fraction of the number of evaluations needed here to obtain a converged estimate of the posterior (in fact the "warm-up" phase of the sampler is barely half completed with such a small number of evaluations).

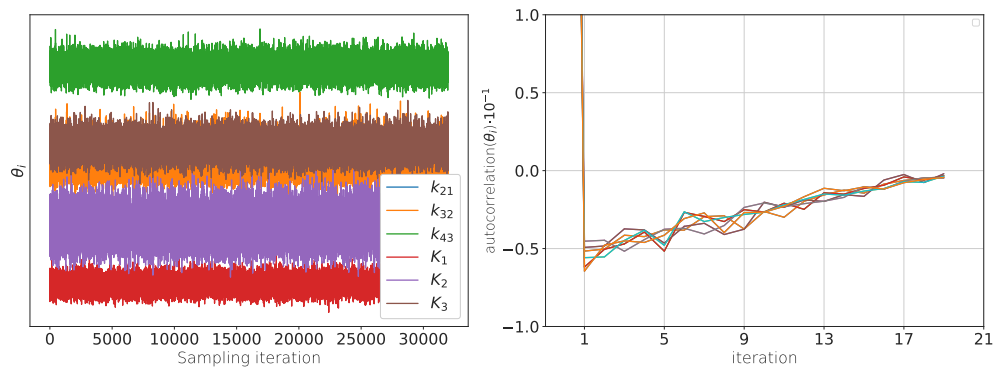


Figure 1. a, Sample traces from the 6-dimensional posterior of the rate matrix recorded after the warm-up of the sampler. **b,** These traces are anti-correlated (in time) such that the effective sample size with which all quantities of relevance are estimated is smaller than the actual number of samples. Note that even at one iteration the absolute value of the auto-correlation is never larger than 0.05. The standard error of the bootstrapped mean auto-correlation is in the order of 10^{-4}.

The ratio of the number of used samples to the number of likelihood evaluations can serve as a simple efficiency measure. However, this ratio is a too optimistic measure as revealed by a look into the internal dynamics of Hamiltonian Monte Carlo sampler (HMC). HMC treats the current parameter sample as part of a phase space point of a (dynamic) Hamiltonian system. The force acting on the system is the derivative of the negative logarithm of the posterior density ***Betancourt*** (***2017***). Each suggested sample is derived from a ballistic movement with random kinetic energy in the augmented parameter space. Calculating this movement is done by integration of the corresponding Hamiltonian equations. Therefore, one suggested sample requires many evaluations of the gradient of the posterior distribution. How many evaluations were needed in turn depends on the shape of the posterior and the length of the ballistic movement.

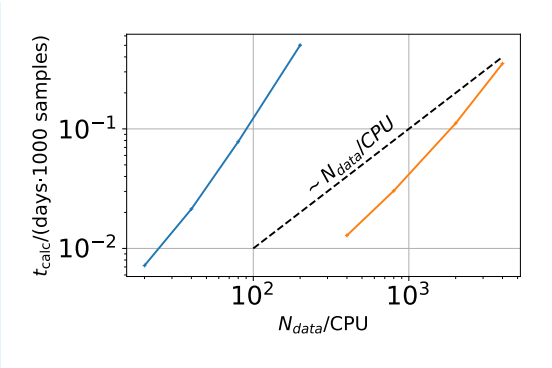


Figure 2. Computational time in units of 1000 samples per day vs. $N_{\text{data}}/\text{CPU}\ 10^{-4}$. The two colors represent two different implementations. The difference between the two implementations is that the derived quantities which are the predicted mean signal and the predicted signal variance for each data point in the less efficient sampling algorithm are saved. In the more efficient algorithm these derived quantities are discarded.

Strong changes of the gradient of the posterior (high curvature of the posterior) require more integration steps. Setting all those sampler parameters requires either expert knowledge about the statistical model or a sampler which automatically learns optimal parameter settings (warm-up) before it starts the actual sampling. Nevertheless, HMC generates samples (Fig. 1**a**) from high-dimensional, correlated posteriors more efficiently than many other classical Markov chain Monte Carlo methods ***Betancourt*** (***2017***). By more efficient we mean the product of the speed of drawing one correlated sample with the amount of iterations needed for the autocorrelation function of those samples to vanish sufficiently. By construction, the samples from any Markov chain Monte Carlo method are correlated (Fig. 1**b**). The quality of the adaptive NUTS HMC sampler provided by Stan can by seen in Fig. 1**b**. The sample traces show a tiny anticorrelation whose absolute value is hardly larger than 0.05. This is achieved only by letting the sampler adapt to the geometry of the posterior in the warm-up phase. The samples collected during the warm-up phase are discarded.

Hereinafter, we show that likelihood and sampler considerations are one aspect of computational efficiency. The other is the implementation of a parallelized algorithm. The blue curve (Fig. 2) represents the scaling of the KF with an implementation of the algorithm which saves besides the sampled parameters the predicted mean and the covariance of the signal. This means that for each parameter sample $5 \cdot N_{\text{traces}} \cdot N_{\text{data}}/\text{CPU}$ more memory operations are needed on the computer node. Note, since we assigned each ligand concentration jump and the following data points to its own CPU the following relation holds: $N_{\text{data}}/\text{CPU} = N_{\text{data}}/\text{trace}$. Skipping saving those derived quantities (orange curve) creates a speedup of roughly 2 orders of magnitude. The derived quantities are redundant quantities which we suggest to calculate by feeding a small subsample of the posterior samples to the KF filter which then reanalyzes the traces by the given draws but does not redraw from the posterior. See Git-Hub: https://github.com/JanMuench/Tutorial_Patch-clamp_data

We compare the computation time for the KF (green) and the RE approach (blue) for the 4-state model (Fig 3**a**) versus the the data quality by N_{ch}. The calculation time t_{calc} rises for both algorithms roughly like $\sim \sqrt{N_{\text{ch}}}$. For each curve the likelihood evaluations stay constant for the whole plot. Thus this scaling relates to the integration time of the HMC sampler. Taking the average of the last 5 calculation times shows that the KF is about 2.7 ± 0.2 times slower than the RE approach. The computation time for cPCF data of the 5-state-2-open states model (Fig 3**b**) shows a similar $\sim \sqrt{N_{\text{ch}}}$-regime until it seems to become independent of N_{ch}. Note, the different N_{ch} interval which is displayed.

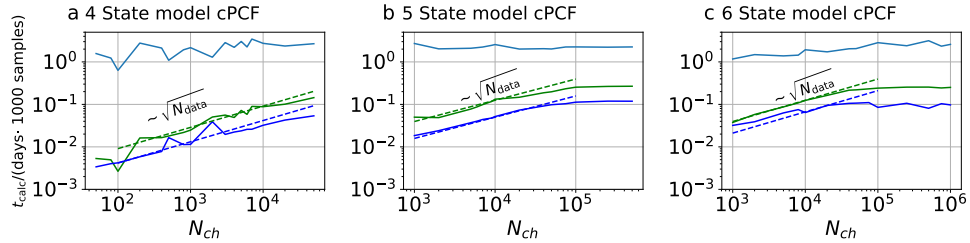


Figure 3. Computational time for cPCF data. Computational time in units of 1,000 samples per day vs. the amount of ion channels per trace N_{ch}. The green curve belongs to the results from the KF, the blue curve to the results from the RE approach. **a,** 4-state model with cPCF data. The posterior of the RE approach was sampled from 3 independent chains. We drew 18,000 samples per chain which included a large fraction of 6,000 warm-up samples per chain. The posterior of the KF was sampled from 4 independent chains. We drew 26,000 samples per chain which included a large fraction of 4,000 warm-up samples per chain. **b** Computational time of the 5-state-2-open-states model with cPCF data. The KF algorithm drew 9,000 samples which included a large fraction of 6,000 warm-up samples per chain. The RE algorithm drew 10,000 samples which included a large fraction of 5,000 warm-up samples per chain. In total we used 4 independent chains. **c,** Computational time of the 6-state-1-open-state model with cPCF data. The KF algorithm drew 10,000 samples which included a large fraction of 6,000 warm-up samples per chain. The RE algorithm drew 12,000 samples which included a large fraction of 7,000 warm-up samples per chain. In total we used four independent sampling chains.

Taking the average of the last 4 computation times shows that the KF is about 2.24 ± 0.01 times slower than the RE approach. Surprisingly, with increasing model complexity the ratio of the computation time becomes not necessarily larger. The computation time for cPCF data of the 6-state-1-open state model (Fig 3**c**) shows a similar $\sim \sqrt{N_{\mathrm{ch}}}$-regime until it becomes independent of N_{ch}. Taking the average of the last 4 calculation times reveals that the KF is about 3.3 ± 0.3 times slower than the RE approach. With increasing model complexity, the ratio of the computation times becomes larger. Note, that there is little variation of the absolute calculation time of the algorithm across model complexity which is a good prospect for more complex kinetic schemes. Nevertheless, to achieve this computational speed we used in total 72 CPUs on one node, 20 CPUs per chain for 20 time traces. Thus, nodes with more then 80 CPUs would enable a better performance. Overall, the more complex likelihood calculations of the KF require 2-3 times more calculation for the tested models and cPCF data.

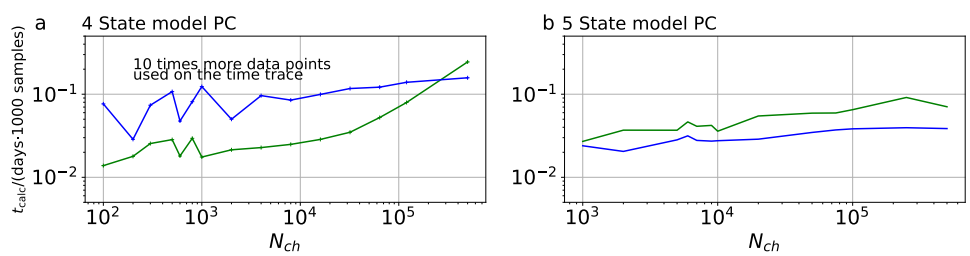


Figure 4. Computational time for PC data The computational time in units of 1,000 samples per day are plotted vs. the amount of ion channels per trace N_{ch}. The green curve belongs to the results from the KF, the blue curve to the results from the RE approach. **a,** 4-state model with PC data. In contrast to the cPCF data (Fig. 3), f_{ana} is increased by an order of magnitude. Thus, 10 times more data points per time trace and CPU are evaluated. These results were obtained by 4 independent chains which drew 9,000 samples per chain, including a large fraction of 7,000 warm-up samples per chain.**b,** 5-states model with PC data. Note that here f_{ana} equals the analyzing frequency of cPCF data (Fig. 3).

Surprisingly, for most part of the displayed interval the KF is faster than the RE approach (Fig 4**a**) for PC data. Considering the more expensive likelihood evaluation (prediction and correction) of the KF, this result reminds that this computational time benchmark is a downstream test that integrates all aspects of the sampling. One possible explanation could be the following: We describe in the main text (Fig. 4**b1-6**) for the RE approach, that due to the treatment of every data point as an individual draw from a multinomial distribution, HDCVs (Fig. 5-6) are too narrow. This problem gets exacerbated (Fig 3**c**) as we used 10 times more data points. The narrower the posterior is the more integration steps from the sampler are needed to calculate the proposed trajectory with sufficient precision. Another example displayed (Fig 4**b**), has an increased kinetic scheme complexity. Note that we reduced the analysing frequency 10 times to be comparable with the analyzing frequency of cPCF data. Under this condition, the intuition from likelihood calculation complexity is confirmed that the KF is slower. Taking the average of the last 4 calculation times shows that the KF is about 1.8 ± 0.2 slower than the RE algorithm for PC data.

Convergence diagnostics

In the limit of infinitely many samples drawn from a posterior by any MCMC method, given that posterior space is simply connected, the sampling is guaranteed to converge to the typical set ***Betancourt*** (***2017***). For a finite amount of samples from the posterior it is not guaranteed that the samples are representative for the typical set of the posterior. To ensure that the trace has converged to the typical set is an active field of statistical research. A discussion of these important consistency checks ***Gelman and Rubin*** (***1992***); ***Gelman et al.*** (***1992***); ***Vats and Knudson*** (***2018***); ***Gabry et al.*** (***2019***); ***Vehtari et al.*** (***2021***) ensuring that the sampled traces represent the typical set, is beyond the scope of this paper. In short, we usually worked with 4 independent sampling chains of the posterior ***Gelman and Rubin*** (***1992***). On the one hand, we visually checked that all 4 chains do report the same typical set to verify that the sampler is fine. This convergence can be monitored by the $\hat{R}$ statistics ***Gelman et al.*** (***1992***); ***Vehtari et al.*** (***2021***). $\hat{R}$ evaluates differences between the within- and between-chain parameter estimates. A bad mixing is indicated by large values of $\hat{R}$. Perfectly converged chains would be reported with $\hat{R} = 1$. Additionally, the effective sample size N_{eff} of each object of interest needs to be monitored. Both of them are reported by the "summary" method from the "fit" object in PyStan. We show (Table 1) an edited version of the output from this method. The last two columns would indicate signs of missing convergence. The second column reports the Markov sampling error telling if we drew enough samples in total. Note that any quantity far out in the tails would have a much larger error. There were cases when the sampler was not fine for all chains, meaning that 3 chains showed the same posterior but one showed something different. In such cases we simply discarded that chain. This occurred in particular more often with the RE approach. As we are benchmarking methods and know the true parameter values, it was rather obvious which chain did not work well. In particular, the successful binomial test of the HDCV is a downstream test that confirms besides the assumptions of the algorithm also whether the sampler worked sufficiently. For the Bayesian filter, sampler problems occurred at very low ion channel counts below $N_{\mathrm{ch}} = 200$.

	mean	se_{mean}	sd	2.5%	50%	97.5%	n_{eff}	Rhat
k_{21}	507.65	0.06	11.91	484.74	507.46	531.62	41262	1.0
k_{32}	481.15	0.16	28.63	428.34	480.19	540.32	33610	1.0
k_{43}	509.95	0.04	6.99	496.39	509.91	523.77	36006	1.0
K_1	0.5	2.6e-5	5.2e-3	0.49	0.5	0.51	40548	1.0
K_2	0.51	6.7e-5	0.01	0.49	0.51	0.54	33228	1.0
K_3	0.49	4.2e-5	7.5e-3	0.47	0.49	0.5	31445	1.0
$N_{ch}[1]$	1007.1	0.14	20.05	968.84	1006.8	1047.2	19838	1.0
$N_{ch}[2]$	1006.1	0.14	20.04	967.83	1005.8	1046.0	19874	1.0
$N_{ch}[3]$	1003.9	0.14	20.0	965.66	1003.6	1043.8	19877	1.0
$N_{ch}[4]$	1007.0	0.14	20.09	968.67	1006.6	1047.1	19843	1.0
$N_{ch}[5]$	1005.9	0.14	20.09	967.61	1005.6	1045.9	19874	1.0
i	1.0	7.0e-5	9.8e-3	0.98	1.0	1.02	19971	1.0
σ^2_{exp}	0.25	1.2e-5	2.5e-3	0.25	0.25	0.26	43142	1.0
σ^2_{op}	2.5e-3	1.2e-7	2.5e-5	2.5e-3	2.5e-3	2.5e-3	42707	1.0
σ^2_{fluo}	1.3e-5	2.3e-8	4.6e-6	1.0e-5	1.2e-5	2.5e-5	41306	1.0
λ	9.98	1.1e-4	0.02	9.94	9.98	10.02	35109	1.0

Table 1. Typical summary statistics of the posterior as printed by the "summary" method from the "fit" object in Pystan. We deleted the 25% and 75% percentiles to decrease the size of the table and changed some of the names with respect to the used name in the algorithm to the symbol used in this article. Note that we assumed to measure 2 ligand concentrations (which includes 4 jumps) from one patch. Thus, we assumed for 10 ligand concentrations that they are coming from 5 patches which leads to a 5 dimensional channels per patch vector $\mathbf{N}_{ch}$. The bias of N_{ch} i discussed in the main article, is also observed here.

HMC parameter settings

The sampling efficiency of an HMC algorithm is sensitive to three parameters ***Neal et al.*** (***2011***), ***Hoffman and Gelman*** (***2014***). The three parameters, discretization time ϵ, the metric $\mathbf{M}$ and the number of steps taken L can be predefined or/and are adapted in the warm-up phase of the sampler. A typical number of iterations N_{warm} in the warm-up phase for each chain is between $N_{warm} = 3 \cdot 10^3 \ldots 7 \cdot 10^3$. Since a sufficient N_{warm} depends on the correlation structure of the posterior, which depends on the quality and quantity of the data, we did not engage to optimize the number of warm-up steps. We rather checked whether for a set of different data (for example with different N_{ch}) all chains are fine.

Stan

Stan is a probabilistic programming language that provides the research community with easy access to HMC sampling, variational Bayes, and optimization procedures like Maximum posterior/ML approaches. Stan uses automatic analytical differentiation of the statistical model such that no analytical treatment by hand has to be done for any tested model. The only requirement from the user is to define a statistical model of the data. The Stan compiler calculates the analytical derivatives of the model and then translates the statistical model to C++ code and after that into an executable program. Furthermore, Stan uses adaptive HMC such that the length of the ballistic trajectory as well as the variance of the random kinetic energy to create the ballistic trajectory and the number of integration steps are optimized automatically. Then the program can be started from any high-level data processing language of the user's choice such as Python, Matlab, Julia, R or from the command line.

Appendix 2

Markov Models for a single ion channel

Markov models and rate models are widely used for modeling molecular kinetics. They provide an interpretation of the data in terms of a set of functional states and the transition rates between these states. Markov models can be estimated from experimentally recorded data as well as from computer simulation data. The use of Markov models with one-step memory is supported by the concept of the molecular free energy landscape. Molecular energy landscapes are typically characterized by conformationally well defined free-energy minima that are separated by free-energy barriers. State transitions in molecules are thermally activated barrier-crossing events on this landscape ***Frauenfelder et al.*** (***1991***) leading to a rapid equilibration of the system in the vicinity of this new minimum. Memory of other minima that have been visited in the past is not required. Regarding the wide spectrum of time scales at which processes in a protein take place, one has to be aware that there is typically a small number of relaxation modes with excessively long autocorrelation times and many relaxation modes with much faster autocorrelation times. To model the slow, experimentally accessible processes, it is sufficient to retain the small number of slow modes ***Noé et al.*** (***2011***). It has been shown rigorously that working with the set of slow modes is equivalent to model the state dynamics with a small number of fuzzily defined metastable states in the full conformational space ***Deuflhard and Weber*** (***2005***). Later it has been shown that the set of slow modes can be well approximated with a hidden Markov model ***Noé et al.*** (***2013a***).

Appendix 3

Eigenvalues and fitting of sums exponential decays: Mean time evolution of the spectral components of the time evolution

From the experimental perspective we are used to fit sums of exponential decays to time series data of any signal such as currents or fluorescence.

$$I(t) = I_\infty + \sum_{i=1}^{M} I_i \exp(-t/\tau_i) \tag{63}$$

The mathematical background for this procedure is the spectral decomposition of the solution of the corresponding RE equation

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}\mathbf{p}(t), \tag{64}$$

which is a vector differential equation whose evolution is governed by the rate matrix $\mathbf{K}$. The general solution to the mean time evolution of the system is

$$\mathbf{p}(t) = \exp(\mathbf{K} \cdot t)\mathbf{p}(t = 0),$$

A Markov state model has as many real valued eigenvalues as it has states M. With those eigenvalues one can decompose the solution into its spectral components

$$\mathbf{p}(t) = C_1 \cdot \exp(\alpha_0 t)\tilde{\mathbf{p}}_0 + \sum_{i=1}^{M-1} C_i \cdot \exp(\alpha_i t)\tilde{\mathbf{p}}_i$$

All of the eigenvalues obey $\alpha \leq 0$. One of them always fulfills $\alpha_0 = 0$ which belongs to the equilibirum probability distribution of the ion channel. With $exp(0) = 1$ we can write the sum as

$$\mathbf{p}(t) = C_1 \cdot \tilde{p}_0 + \sum_{i=1}^{M-1} C_i \cdot \exp(\alpha_i t)\tilde{p}_i \tag{65}$$

The eigenvalue $\alpha_0 = 0$ determines the equilibrium solution. Due to $\alpha_i \leq 0$ the solution is stable and converges to

$$\mathbf{p}(t = \infty) = C_1 \cdot \tilde{p}_0$$

The time evolution of the mean signal in its spectral components can then be derived

$$I(t) = \mathbf{H}\left(C_1 \cdot \tilde{p}_0 + \sum_{i=1}^{M-1} C_i \cdot \exp(\alpha_i t)\tilde{p}_i\right) \tag{66}$$

by multiplying with the observation matrix $\mathbf{H}$ and setting $\alpha_i = -\frac{1}{\tau_i}$ we derived Eq. 63 ***Colquhoun and Hawkes*** (***1995***). To derive the likelihood based on this deterministic approximation ***Milescu et al.*** (***2005***) one assumes for each data point that the ensemble state $\mathbf{n}(t)$ is drawn from

$$\mathbf{n}(t) \sim \text{multinomial}(N_{\text{ch}}, \mathbf{p}(t)) \tag{67}$$

with the additional approximation of the multinomial distribution by a multivariate normal distribution. Subsequently, the equation

$$\mathbf{y}(t) \sim \text{normal}(\mathbf{H}n(t), \Sigma(t)) \tag{68}$$

provides the contribution of the recording system to the signal of the ensemble state. In contrast, the KF approximates a

$$\mathbf{n}(t+\Delta t) \sim \text{generalized}-\text{multinomial}(\mathbf{n}(t), \mathbf{T}) \tag{69}$$

distribution with multivariate normal distributions (see Methods) and then the details of the experimental setup are adjoined

Appendix 4

A note about Kalman filtering within other frameworks such as variational Bayes, maximum a posteriori and ML

We like to emphasize that the key property of the Kalman filter likelihood for an autocorrelated time series (see Eq. 9) to factorize over time enables us to treat each data point as if it were independent. The alternative is to treat the whole time series as one high-dimensional data point as done in ***Celentano and Hawkes*** (***2004***). A ML Kalman filter ***Moffatt*** (***2007***) would try to optimize Eq. 9, which is usually done in the log space to improve numerical stability. Taking the logarithm of Eq. 9 gives also an insight for the reader, who might be more experienced in using residual sums of squares for fitting.

$$\log \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \log \prod_{t=2}^{N_T} \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}[\mathbf{n}_t], \mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t) = \sum_{t=2}^{N_T} \log \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}[\mathbf{n}_t], \mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t), \tag{70}$$

where we used the logarithmic algebra rule $\log(\prod_i a_i) = \sum_i \log(a_i)$ Using the functional form of the multi-variate normal distribution, we arrive at

$$\log \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \sum_{t=2}^{N_T} \log(\frac{1}{\sqrt{(2\pi)^2 \det(\mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t)}})(-0.5) \cdot (\mathbf{y}_t - \mathbf{H}\mathbb{E}[\mathbf{n}_t])[\mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t]^{-1}(\mathbf{y}_t - \mathbf{H}\mathbb{E}[\mathbf{n}_t])^\top, \tag{71}$$

which is a vectorial residual sum of squares whose summands are weighted by $\mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t$. The past observations of time series are encoded in $\mathbf{P}_t$ and $\mathbb{E}[\mathbf{n}_t]$, which are calculated with the Kalman filtering equations before the likelihood can be evaluated.

Generally, the user of our algorithm is not restricted to the full Bayesian inference. He/she can also stick to optimizing Eq. 71. This is because in the probabilistic programming language Stan one only formulates the statistical model. The method to evaluate the statistical model and the data is provided by the Stan compiler and the high level programming language interface. Together they allow to switch quickly between HMC sampling, variational Bayesian posterior estimation, maximum a posteriori and ML. In particular, the variational Bayesian posterior might be a way to ease the computational burden due HMC sampling while keeping many of the benefits of knowing the full posterior distribution.

In the previous versions before PyStan 3.0 all three approaches were supported. Currently, with PyStan 3.0 only HMC sampling of the posterior is supported but, if desired, any other interface can be used (Matlab, Julia, R, CmdStan, Pystan 2.9, CmdStanPy,...) to perform optimization or variational Bayesian approaches. No actual changes in the Stan code are required, eventually with the exception that the CPU based parallelization needs to be eliminated. Given that one chooses optimization, the differences between maximum a posteriori/penalized maximum likelihood or only ML are just a matter of prior choice which need to be coded into the statistical model. Apart from those minor points, there is no need of extensive re-implementation of the code.

Appendix 5

The fluorescence signal of cPCF experiments

First four moments of a photomultiplier signal

In this work, the KF analysis assumes Poisson statistics for the fluorescence signal in cPCF. Many commercial microscopes are not equipped with photon counting detectors or detectors are not operated in photon-counting mode, often to due to ease of use or limitation in dynamic range. Therefore, it is important to verify that the fluorescence signal follows, at least approximately, Poisson counting statistics. In particular, for the KF it is assumed that higher order statistics, such as skewness and excess kurtosis, vanish. The central assumption of the derivation of our Bayesian filter is $\mathrm{var}[y_{\mathrm{fl}}] = \mathbb{E}[y_{\mathrm{fl}}]$

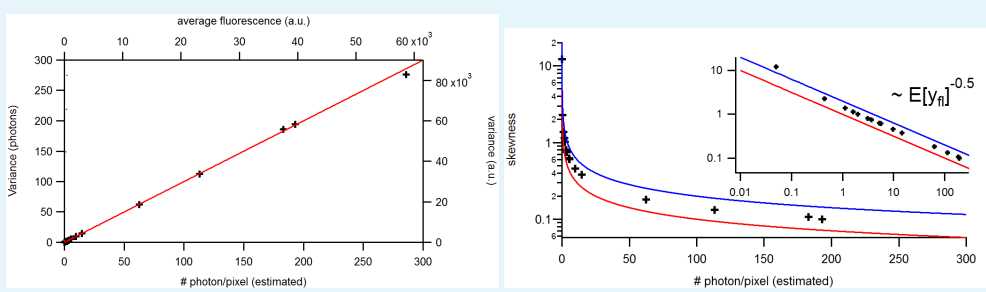
Here we show that this assumption for the detectors used in our previously used LSM 710 (Carl Zeiss, Jena) system ***Biskup et al.*** (***2007***); ***Kusch et al.*** (***2010***) is fulfilled by re-scaling to photon-numbers. The measured variance obeys $\mathrm{var}[y_{\mathrm{fl}}] = a \cdot \mathbb{E}[y_{\mathrm{fl}}]$. It depends linearly on the mean signals (Appendix Fig. 1a).

$$\mathrm{var}[y] = a\mathbb{E}[y] \tag{72}$$

$$\mathrm{var}[\mu x] = a\mathbb{E}[\mu x] \tag{73}$$

$$\mu^2\,\mathrm{var}[x] = a\mu\,\mathrm{E}[x] \tag{74}$$

For the scaled signal x being Poisson distributed follows $\mu = a$. Re-scaling of the signal by $1/a$ provides approximately Poisson distributed values. A linear fit yields $a = 205$a.u.$(16\,\mathrm{bit})$/photon (for 680 V PMT voltage, 3.26 μs pixel dwell time). Appendix Fig. 1 b,c) shows that excess kurtosis and skewness remain small at all levels of photons/pixel but are somewhat higher than theoretically predicted for Poisson-distributed data. The proportionalities are correctly described by the Poisson distribution assumption but the skewness and the kurtosis are too small by a constant factor of $\sqrt{2}$ and 4, respectively. This finding has to be verified for different experimental conditions, because at lower concentration/particle densities and higher count rates, particle number fluctuations can dominate statistics ***Brown et al.*** (***2008***). For comparison another option would be a Gamma distribution which has the mean and the variance of $\mathbb{E}[y] = k\theta$ and $var[y] = k\theta^2$, respectively. Thus, the applied scaling requires that $\theta = 1$. The Gamma distribution has a higher skewness by factor two (independently of θ) than a Poisson distribution and overscores the skewness and excess kurtosis of the detector. For simplicity only the Poisson distribution is considered in this work. In conclusion: Typical cPCF fluorescence signal detection rates are well approximated by a Gamma or Poisson distribution which in turn have the desired property that can be approximated by a normal distribution.



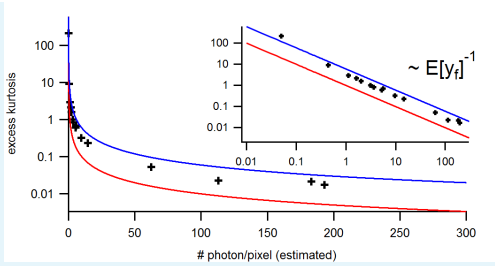


Figure 1. Benchmark of the signal statistics for experimental solution data recorded under cPCF conditions: The concentrations of the fluorescent ligand were 0.25, 3 and 15 μM and a reference dye was present. The laser intensities covered 1.6 orders of magnitude at constant detection settings. The data points were obtained from $1.4 \cdot 10^6$ pixel. The red and blue lines indicate the theoretical prediction for a Poisson and Gamma distribution, respectively, assuming $\theta = 1$. **a**, Variance vs. average. The linear relation allows to relate the measured a.u. (top, right axis) to photons (bottom, left axis). **b**, Skewness. **c**, Excess kurtosis. The higher moments are small but the values are slightly larger than theoretically predicted. The insets provide a corresponding log-log plot. Important for the KF algorithm is that skewness and excess kurtosis is small.

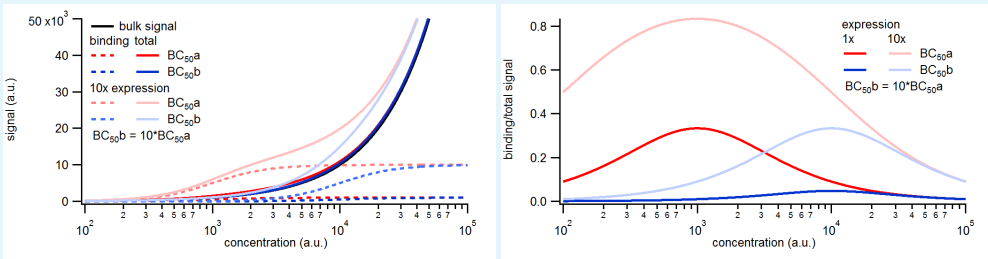


Figure 2. Simulated binding signals. **a,** Comparison of binding of a labeled ligand at two concentrations. A simple two-ligand binding process was simulated with the Hill equation for the two expression levels of 1,000 or 10,000 binding sites per patch and a BC_{50} of 1,000 (BC_{50}a) or 10,000(BC_{50}b), respectively, given in molecules per observation unit. The observed signal is the sum of the signal from ligands free in solution and bound to the receptors. The solution signal scales linearly with the concentration, while the binding signal saturates. **b,** Relative contribution of the binding signal to the total signal. Note that the contribution of the binding signal scales linearly with the expression level and inversely with the BC_{50}.

Background noise statistics

In cPCF measurements with fluorescence-labeled ligands, the signals of the ligands bound to the receptors overlap with the signals from freely diffusing fluorescence-labelled ligands in the bulk. This bulk signal is subtracted from the total signal***Biskup et al.*** (***2007***). While the mean difference signal $y_{\mathrm{f1,k}}(t)$ of the confocal voxel k represents the bound ligands in that voxel, its noise $y_{\zeta,k}(t)$ originates from both bound and bulk ligands. The additional bulk signal, e.g. the fraction of bulk solution inside that voxel, varies from voxel to voxel and can hardly be described theoretically. Nevertheless, it can be determined experimentally ***Biskup et al.*** (***2007***). At low expression levels or at ligand concentrations above low nano-molar levels, this background signal is not negligible. It scales linearly with the ligand concentration, while the signal from bound receptors depends on the affinity, as estimated by the concentration of half maximum binding BC_{50}, and the number of ion channels in the membrane of the observed volume. The binding signal saturates at high concentrations (Appendix Fig. 2). Thus, both high affinity (low BC_{50}) and high expression reduce the relative contribution of the background to the overall signal, improving the signal to noise ratio.

Practically, the bulk signal is estimated by counter-staining the solution with a spectrally distinct reference dye ***Biskup et al.*** (***2007***). The spatial distribution of this dye mimics the spatial distribution of the freely diffusing ligands. The bulk absolute concentration as well

as the molecular brightness of the reference dye and the labeled ligand differ. Hence, the binding signal is calculated as the average pixel intensity of the scaled difference image between the signal of labeled ligand and reference dye according to

$$y_{\mathrm{fl,k}} = y_{\mathrm{lig,total}} - \hat{\lambda}_{\mathrm{lig,back}} - (y_{\mathrm{fl,ref}} - \hat{\lambda}_{\mathrm{ref,back}}) \frac{\hat{\lambda}_{\mathrm{lig}} - \hat{\lambda}_{\mathrm{lig,back}}}{\hat{\lambda}_{\mathrm{ref}} - \hat{\lambda}_{\mathrm{ref,back}}}, \tag{75}$$

where $\hat{\lambda}_{\mathrm{lig,back}}$ and $\hat{\lambda}_{\mathrm{ref,back}}$ are the arithmetic mean background signals of the ligand and reference dye recorded beyond the membrane where no signal should be recorded. They represent a signal offset which needs to be subtracted. The mean intensities in the bulk, $\hat{\lambda}_{\mathrm{bulk}}$ and $\hat{\lambda}_{\mathrm{ref}}$, are estimated outside the pipette. In order to get the correct scaling, the mean intensities need to be corrected by the respective mean background signals. If $\frac{\hat{\lambda}_{\mathrm{lig}} - \hat{\lambda}_{\mathrm{lig,back}}}{\hat{\lambda}_{\mathrm{ref}} - \hat{\lambda}_{\mathrm{ref,back}}} = 1$ holds then $y_{\mathrm{fl,bin}}$ would be Skellam distributed ***Hwang et al.*** (***2007***). The total signal is then $y_{\mathrm{fl}} = \sum_k y_{\mathrm{fl,k}}$. This procedure creates $\mathbb{E}[y_\zeta] = 0$ but adds an additional noise term $\boldsymbol{\zeta}(t_j)$. For the general case of different intensities, we name the distribution 'scaled Skellam distributed'. The scaling variance of the background noise in each voxel of the difference image

$$\sigma_\zeta^2 = \lambda_{\mathrm{lig}} + \frac{\lambda_{\mathrm{lig}}^2}{\lambda_{\mathrm{ref}}} \tag{76}$$

is derived from simulated data in the Appendix 5. λ_{lig} and λ_{ref} are the fluorescence intensity from the freely diffusing ligands and reference dye molecules per voxel, respectively. λ_{lig} and λ_{ref} are proportional to the volume fraction of the voxel, which is occupied by the bulk, and to the respective concentrations. To achieve a symmetric $\mathbb{P}(\zeta)$, one can set $\lambda_{\mathrm{lig}} = \lambda_{\mathrm{ref}}$. The summed variance of all selected voxels can be tabulated according to

$$\Sigma_{back} = \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{77}$$

To mimic an experiment which creates time series data $\zeta(t)$, we draw Poisson numbers for the signal from the membrane $Poisson(\mathbf{Hn}(t))$ and for the signal from the bulk we draw numbers from the two respective Poisson distributions. Then subtraction of the two background signals is performed according to

$$y_{\mathrm{bulk}} = y_{\mathrm{lig,bulk}} - y_{\mathrm{ref,bulk}} \frac{\lambda_{\mathrm{lig,bulk}}}{\lambda_{\mathrm{ref,bulk}}} \tag{78}$$

assuming that the dark count signal has been correctly subtracted. Then we add the bulk signal to the bound ligand signal. In this way we produce a time trace with colored noise by the Gillespie algorithm and add white noise to time traces as it is observed in real experiments.

Deriving the moments of the background noise for the difference signal

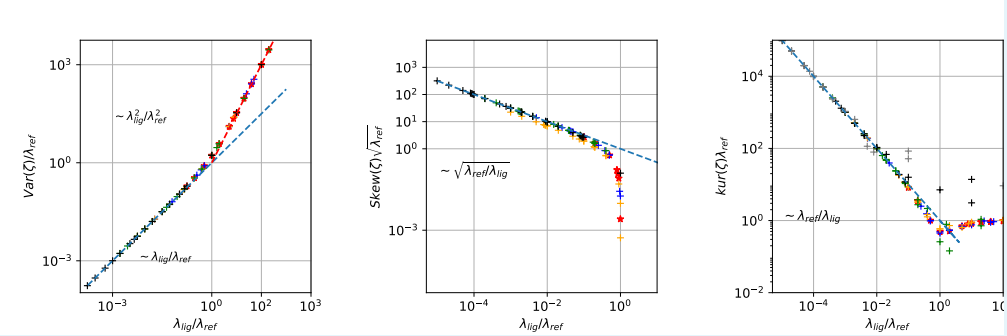


Figure 3. Master curves of 2nd till 4th centralized moment of photon counting noise ζ arising from the difference signal of fluorescent ligands and the dye in the bulk. The curves are created from $4 \cdot 10^5$ draws from Poisson distributions with different combinations of intensities for the reference dye λ_{ref} and of the intensity of the confocal voxel fraction λ_{lig}.

For the KF the variance, skewness and kurtosis arising from the background noise has to be calculated. Skewness and excess kurtosis of the distribution have to be small compared to the total variance of the signal including all noise sources because only in this case the KF algorithm can be considered as the optimal solution for the filtering and inference problem ***Anderson and Moore*** (***2012***). In the following the 2nd to 4th moment of ζ are derived. The noise intensity parameter of the reference dye λ_{ref} is proportional to $\rho_{\mathrm{ref}}\lambda_{\mathrm{bulk}}$ with V being the confocal volume fraction containing fluorophores and ρ_{ref} the density of the fluorophores in this volume. In Appendix Fig. 3 we deduce master curves for the variance skewness and excess kurtosis of the white noise by drawing $4 \cdot 10^5$ Poisson numbers from the respective Poisson distribution and subtract them from each other according to Appendix Eq. 78. The variance is derived empirically to be

$$\frac{\sigma_\zeta^2}{\lambda_{\mathrm{ref}}} = \frac{\lambda_{\mathrm{lig}}}{\lambda_{\mathrm{ref}}} + \frac{\lambda_{\mathrm{lig}}^2}{\lambda_{\mathrm{ref}}^2}. \tag{79}$$

In Appendix Fig. S. 3 **a**, we confirm the intuition $\lambda_{\mathrm{ref}} \to \infty \Rightarrow \mathrm{var}(\zeta) = \lambda_{\mathrm{lig}}$. Optimally, the skewness should be zero to avoid a biased estimate of θ when the data are analyzed by the KF. Empirically, for $\lambda_{lig} \ll \lambda_{ref}$ the skewness holds

$$\mathrm{skew}(\zeta)\sqrt{\lambda_{\mathrm{ref}}} = \sqrt{\frac{\lambda_{\mathrm{ref}}}{\lambda_{\mathrm{lig}}}} \tag{80}$$

Additionally for $\lambda_{lig} < \lambda_{ref}$ the skewness holds

$$\mathrm{skew}(\zeta)\sqrt{\lambda_{\mathrm{ref}}} \leq \sqrt{\frac{\lambda_{\mathrm{ref}}}{\lambda_{\mathrm{lig}}}} \tag{81}$$

It is zero when $\lambda_{ref} = \lambda_{lig}$. The KF is optimal if the kurtosis excess approaches zero, in other words if ζ is distributed normally. Empirically the kurtosis holds this

$$\mathrm{kur}(\zeta)\lambda_{\mathrm{ref}} \leq \frac{\lambda_{\mathrm{ref}}}{\lambda_{\mathrm{lig}}} \tag{82}$$

for $\lambda_{\mathrm{ref}} \leq \lambda_{\mathrm{lig}}$. The relative intensity λ_{lig} of the voxel fraction compared to the intensity λ_b depends on the affinity of the ligand to the receptor, the number of receptors in the patch, and the density of the fluorphores ρ_{lig} at the patch. For larger concentrations the ratio should be $\lambda_{\mathrm{lig}}/\lambda_{\mathrm{ref}}$.

Appendix 6

Output statistics of Bayesian filters

Classical Kalman Filter without open-channel noise

Assuming that current measurements are only compromised by additive technical white noise ν but do not contain open-channel noise ν_{op}, then our noise model reduces to

$$y(t) = \mathbf{H}\mathbf{n}(t) + \nu(t) \Leftrightarrow y \sim \mathbb{O}(y|\mathbf{n}) = \mathrm{normal}(\mathbf{H}\mathbf{n}(t), \sigma_{\mathrm{m}}^2) \tag{83}$$

The noise term ν_{m} has a mean of $\mathbb{E}[\nu_{\mathrm{m}}] = 0$ and variance $\mathbb{E}[\nu_{\mathrm{m}}^2] = \sigma_{\mathrm{m}}^2 = const$. One has to keep in mind that we have to add an extra variance term originating from the dispersion of channels over the state space, as encoded by $\mathbf{P}(t)$ and $\mathbf{n}(t)$. The uncertainty $\mathbf{P}(t)$ is calculated using Methods Eq. 30. The variance of the total output is

$$\begin{aligned}
\mathrm{var}(y(t), y(t)) &= \mathbb{E}[(y(t) - \mathbb{E}[y(t)])(y(t) - \mathbb{E}[y(t)])^{\top}] && \text{(84a)}\\
&= \mathbb{E}[(y(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])(y(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])^{\top}] && \text{(84b)}\\
&= \mathbb{E}[(\mathbf{H}\mathbf{n}(t) + \nu(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])(\mathbf{H}\mathbf{n}(t) + \nu(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])^{\top}] && \text{(84c)}\\
&= \mathbf{H}\mathbb{E}[(\mathbf{n}(t) - \mathbb{E}[\mathbf{n}(t)])(\mathbf{n}(t) - \mathbb{E}[\mathbf{n}(t)]^{T}]\mathbf{H}^{\top} + \mathbb{E}[\nu(t)^2] && \text{(84d)}\\
&= \mathbf{H}\mathbf{P}(t)\mathbf{H}^{\top} + \sigma_{\mathrm{m}} && \text{(84e)}\\
& && \text{(84f)}
\end{aligned}$$

The two cross terms $\mathbb{E}[\nu(t_1)(\mathbf{n} - \mathbb{E}[\mathbf{n}])^T\mathbf{H}^T]$ and $\mathbb{E}[\mathbf{H}(\mathbf{n} - \mathbb{E}[\mathbf{n}])\nu(t_1)^T]$ are zero since ν is independent of $\mathbf{n}$ and $\mathbb{E}[\nu_{\mathrm{m}}] = 0$. Our derivation is equivalent to marginalization over the predicted normal prior of the ensemble state $\mathbb{P}(\mathbf{n}(t)|\mathcal{Y}_{t-1})$ at the time of the measurement except that the prior distribution could be any probability distribution with some mean and variance. Eq. 84 is the classical KF variance prediction of a signal. The first term in Eq. 84, describes the variance from stochastic gating and that the ensemble state is hidden. Notably, by Methods Eq. 30 we realize that $\mathrm{var}(y(t))$ contains information about $\mathbf{T}$ and $\mathbf{n}(t-1)$, which we can exploit with the KF framework.

A generalized Kalman filter with state-dependent open-channel noise

In addition to the standard KF with only additive noise ***Moffatt*** (***2007***); ***Anderson and Moore*** (***2012***); ***Chen et al.*** (***2003***), fluctuations arising from the single-channel gating lead to a second white-noise term $\nu_{\mathrm{op}} n_4(t)$, causing state-dependency of our noise model. The output model is then

$$y(t) = \mathbf{H}\mathbf{n}(t) + \nu_{\mathrm{m}}(t) + \nu_{\mathrm{op}}(t) \Leftrightarrow y \sim p(y|\mathbf{n}) = \mathrm{normal}(y|\mathbf{H}\mathbf{n}(t), \sigma_{\mathrm{m}}^2 + n_4(t)\sigma_{\mathrm{op}}^2) \tag{85}$$

The second noise term ν_{op} is defined in terms of the first two moments $\mathbb{E}(\nu_{\mathrm{op}}) = 0$ and therefore $\mathrm{var}(\nu_{\mathrm{op}}) = \mathbb{E}(\nu_{\mathrm{op}}^2) = \sigma_{\mathrm{op}}^2 \mathbf{n}_4(t)$. To the best of our knowledge such a state-dependent noise makes the following integration intractable

$$\begin{aligned}
\mathbb{P}(y(t)) &= \int \mathrm{normal}(y|\mathbf{H}\mathbf{n}, \sigma_{\mathrm{m}}^2 + n_4\sigma_{\mathrm{op}}^2)\ \mathrm{normal}(\mathbf{n}|\overline{\mathbf{n}}(t), \mathbf{P}(t))\,\mathrm{d}n && \text{(86a)}\\
&= \frac{1}{const}\int \exp\left(\frac{(y - \mathbf{H}\mathbf{n})^2}{2(\sigma_{\mathrm{m}}^2 + n_4\sigma_{\mathrm{op}}^2)}\right)\exp\left(\frac{1}{2}(\mathbf{n} - \overline{\mathbf{n}}(t))\mathbf{P}^{-1}(\mathbf{n} - \overline{\mathbf{n}}(t))^{\top}\right)\,\mathrm{d}n && \text{(86b)}
\end{aligned}$$

When assuming that the relative fluctuations of $\mathbf{n}(t)$ are on average small then n_4 in the denominator is close to $\mathbb{E}(n_4)$ of the state. Thus, the incremental likelihood can be written

as in the standard KF, with the only difference that the measurement noise is the sum of two components.

$$y(t) \sim \mathrm{normal}(\mathbf{H}\overline{\mathbf{n}}(t), \sigma_{\mathrm{m}}^2 + \sigma_{\mathrm{op}}^2 \overline{n}_4(t) + \mathbf{HPH}^\top) \tag{87}$$

To see that this approximation of the variance is correct, we apply the law of total variance decomposition ***Weiss*** (***2005***).

$$\mathrm{var}(y(t)) = \mathbb{E}[\mathrm{var}[y(t)|\mathbf{n}(t)]] + \mathrm{var}[\mathbb{E}[y(t)|\mathbf{n}(t)]] \tag{88a}$$

$$= \mathbb{E}[\Sigma + \sigma_{\mathrm{op}}^2 n_4(t)] + \mathrm{var}[\mathbf{Hn}(t)] \tag{88b}$$

$$= \sigma_m^2 + \sigma_{\mathrm{op}}^2 \mathbb{E}[n_4(t)] + \mathbf{HP}(t)\mathbf{H}^\top \tag{88c}$$

The terms $\mathbf{HP}(t)\mathbf{H}^\top + \sigma_m^2$ are the standard output covariance matrix. Again $\mathbf{P}(t)$ contains information about $\mathbf{T}$, $\mathbf{n}(t-1)$ while the additional variance term includes information about the current $\mathbf{n}(t)$. The information contained in the noise influences likelihood in two ways. By the variance or covariance of the current $\mathbf{y}(t)$ but also for $\mathbf{y}(t+1)$ in correction step by the Kalman gain $\mathbf{K}_{\mathrm{Kal}}$ matrix defined in the next section.

Appendix 7

Error induced for the RE or KF approach by analog filtering of PC data

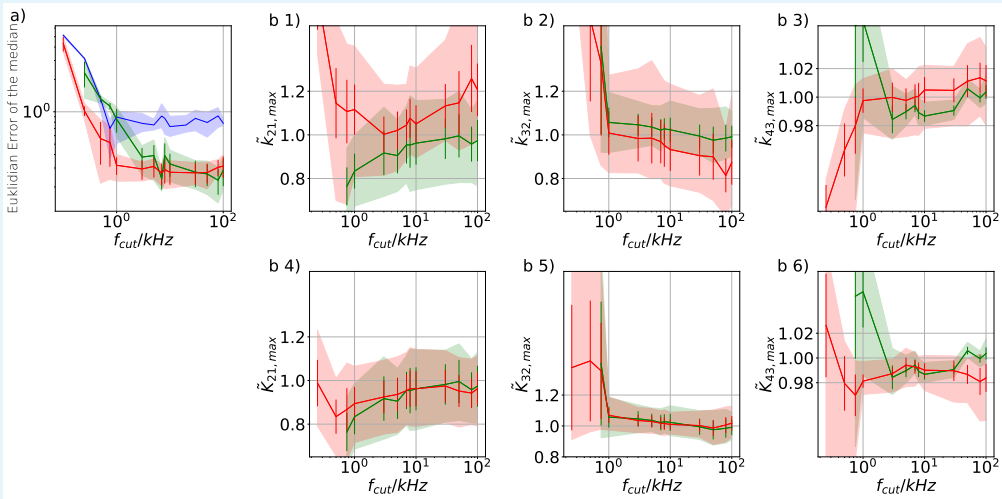


Figure 1. Influence of analog filtering on the accuracy and precision of the maximum of the posterior from PC data We chose for the analog signal $\sigma_{\text{exp}}/i = 10$, $\sigma_{\text{op}}/i = 0.1$, thus a strong background noise. For the ensemble size we chose $N_{\text{ch}} = 4 \cdot 10^3$ such that the likelihood dominates the uniform prior. **a)** Estimate of the distribution of the accuracy (mean Euclidean error of the median of the posterior) vs. the cut-off frequency of a 4th-order Bessel filter scaled to the channel time scale. The solid line is the mean median of 5 data sets of the respective RE posterior (blue) and KF posterior (green). The green shaded area indicates the 0.6 quantile (ranging from 0.2 till 0.8) demonstrating the distribution of the error of the median of the posterior due to the randomness of data.**b 1-3)** Accuracy and precision of the maxima of the posterior $\tilde{k}_{ij,max}$ of the maxima of the posterior rates vs. the cut-off frequency of a Bessel filter. The shaded areas indicate the 0.6 quantiles (ranging from 0.2 till 0.8) due to the variability from data set do data set while the error bars show the standard error of the mean. **b 4-6)** Accuracy and precision of the maxima of the posterior $\tilde{K}_{ij,max}$ of the maxima of the posterior of the corresponding equilibria vs. the cut-off frequency of a Bessel filter.

In order to mimic an analog signal before the analog-to-digital conversion we simulated 5 different data sets of 100 kHz signals which were then filtered by a digital fourth-order Bessel filter. The activation curves were then analyzed with the Bayesian filter at 125 Hz and the deactivation curves at sampling rates between $166 - 500$ Hz. Operating the Bayesian filter at a lower frequency is necessary because due to applying the Bessel filter the former white noise of the signal obtained additional time correlations. Thus, an all data points fit would immediately violate the white noise assumption of Eq. 4 which we restore by analyzing at a much lower frequency. We then let the time scales of the induced time correlations become larger and larger by decreasing f_{cut}. We show the results of the Euclidean Error for 3 different cases for PC data. The RE approach (blue) and the Bayesian filter (red) share the same analysing frequency. In contrast, an increased $f_{\text{ana}} = 1660 - 5000$ Hz (green) is used for the Bayesian filter. Both algorithms (blue, red) with the same f_{ana} show a similar rather constant region separated by an offset. Similar, to cPCF data the KF is more robust. The Bayesian filter with $f_{\text{ana}} = 1660 - 5000$ Hz (green) does not show this constant region but outperforms the Bayesian filter with the lower f_{ana} if the recording system is operated with minimal analog filtering. This becomes even more apparent when we consider the single parameter deviation vs. f_{cut} (Appendix 7 Fig 1 **b1-6**). Note, the strong dependence of the critical f_{cut} at which the performance of all algorithms becomes suddenly worse scales with f_{ana}. Additionally, it is crucial, independently of the algorithm, that $f_{\text{cut}} \gg f_{\text{ana}}$. In Fig. **b1-6)** we demonstrate the distribution of the errors on the single parameter level and compare

only the KF for different f_{ana}. Similar to the findings in the main text, on the one hand, the Euclidean error (red) shows some robustness against f_{cut}. On the other hand, the influence of f_{cut} on the individual parameter level sets in immediately and is complex. The best results are achieved with minimal analog filtering and a high f_{ana}.

Appendix 8

Details of the specified two methods to count the probability mass needed to include the true value

In Fig. 4**c** we compare the true against the expected success probabilities of finding the complete true rate matrix within an n-dimensional HDCV for different expected success probabilities of a perfect model. By "expected" and "perfect" we refer to a fictitious ideal algorithm which exactly models exhaustively all details of the true process, including all measurement details. The first way assumes a multivariate normal posterior distribution with $\mathbb{P}(\boldsymbol{\theta}|y)_{\text{post}} \approx \text{normal}(\boldsymbol{\theta}|E[\boldsymbol{\theta}], \boldsymbol{\Sigma})$. Then the ellipsoid of a constant probability density is exactly the surface of a HDCV of given probability mass P. For a two-dimensional representation see Fig. 4**d** below the diagonal. In one dimension, the ellipsoid consists of the two points with a distance $d_{\text{Mah}} = \sqrt{(\theta - E[\theta])^2}/\sigma$ from the mean (inset Fig. 5**a**). In general, the n-dimensional ellipsoids around the mean of a multivariate normal distribution can be described by points which have the same Mahalanobis distance d_{Mah} from the mean.

$$d_{\text{Mah}}(\boldsymbol{\theta}) = \sqrt{(\boldsymbol{\theta} - E[\boldsymbol{\theta}])\Sigma^{-1}(\boldsymbol{\theta} - E[\boldsymbol{\theta}])} \tag{89}$$

For a multivariate standard normal distribution without correlation the Mahalanobis distance becomes the Euclidean distance $\sqrt{(\boldsymbol{\theta} - E[\boldsymbol{\theta}])^T(\boldsymbol{\theta} - E[\boldsymbol{\theta}])}$. Rewriting the multivariate normal distribution in terms of the Mahalanobis distance

$$\text{normal}(\boldsymbol{\theta}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp[-0.5 d(\boldsymbol{\theta})^2_{\text{Mah}}] \tag{90}$$

reveals that all points $\boldsymbol{\theta}$ with $d_{\text{Mah}} = const$ have the same probability density. The random variable d^2_{Mah} is χ-square distributed. Thus the χ-square distribution (Fig. 5**a**) is the probability density of drawing an $\boldsymbol{\theta}$ in n dimensions and finding it on the ellipsoid's $n-1$-dimensional surface, at a distance d_{Mah} from the mean. This allows us to use the cumulative χ^2-square distribution function to calculate the probability mass inside the ellipsoid which is the desired HDCV. By evaluating $d_{\text{Mah}} = \sqrt{(\boldsymbol{\theta}_{\text{true}} - E[\boldsymbol{\theta}])\Sigma^{-1}(\boldsymbol{\theta}_{\text{true}} - E[\boldsymbol{\theta}])}$ and $\chi^2_{\text{cdf}}(d^2_{\text{Mah}})$ we specify how much volume, in units of probability mass, has to be counted until the volume includes the true value. Note, that with increasing dimensionality (Fig. 5**a**) the probability mass is shifted away from the mode of the probability density. For general probability densities, this is also true in higher dimensions where one only rarely finds the true value within a small distance to the maximum of the probability density. Note, that there is no mathematical theorem for singular models ***Watanabe*** (***2007***) saying that the posterior approaches asymptotically a multivariate normal with increasing data quality/quantity. Consequently, the underlying assumption that the posterior distribution is multivariate normal, is situation dependent valid or highly questionable. Thus the displayed method should be validated additionally in an independent way. To determine the probability mass needed to include the true value into the HDCV for such a posterior, we can also use a histogram-based method. One starts by constructing an n-dimensional histogram from the samples of the posterior and by initializing a global variable with zero. Then we start counting from the bin with most counts and check whether the true value falls inside this bin. If not, we add the probability mass inside that bin to the global variable. Repeating this for next lower bins eventually leads to the detection of the bin with the true rate matrix inside. On detection, the global variable contains the sought-after value P. While this procedure does not depend on a multivariate normal assumption, it is prone to errors due to the discrete bins and the finite samples from the posterior.

Nevertheless, both procedures show (Fig. 5**b**) good agreement when plotting the volume/

probability mass needed to include the true parameter value vs. N_{ch}. Again $N_{\mathrm{ch}} \approx 200$ seems to be a reasonable data quality to trust the statistics of the posterior using the KF. In contrast, the posterior of the RE approach never includes the true values with a reasonable HDCV. Note, that a probability mass of ≈ 1 to reach the true value means that qualitatively speaking the estimate of the RE approach is infinitely far away from the true value in terms of the Mahalanobis distance. Further, due to the finite sampling the method of histogram counting is not qualified for the largest HDCVs approaching $P \approx 1$. The two developed methods contrast the Euclidean distance from the true values to the maximum mode, median or even the mean of the distribution against all higher moments of the distribution. Thus it tests the overall shape of the posterior.

Appendix 9

Uncertainty quantification for the 5-state and 6-state model with cPCF data

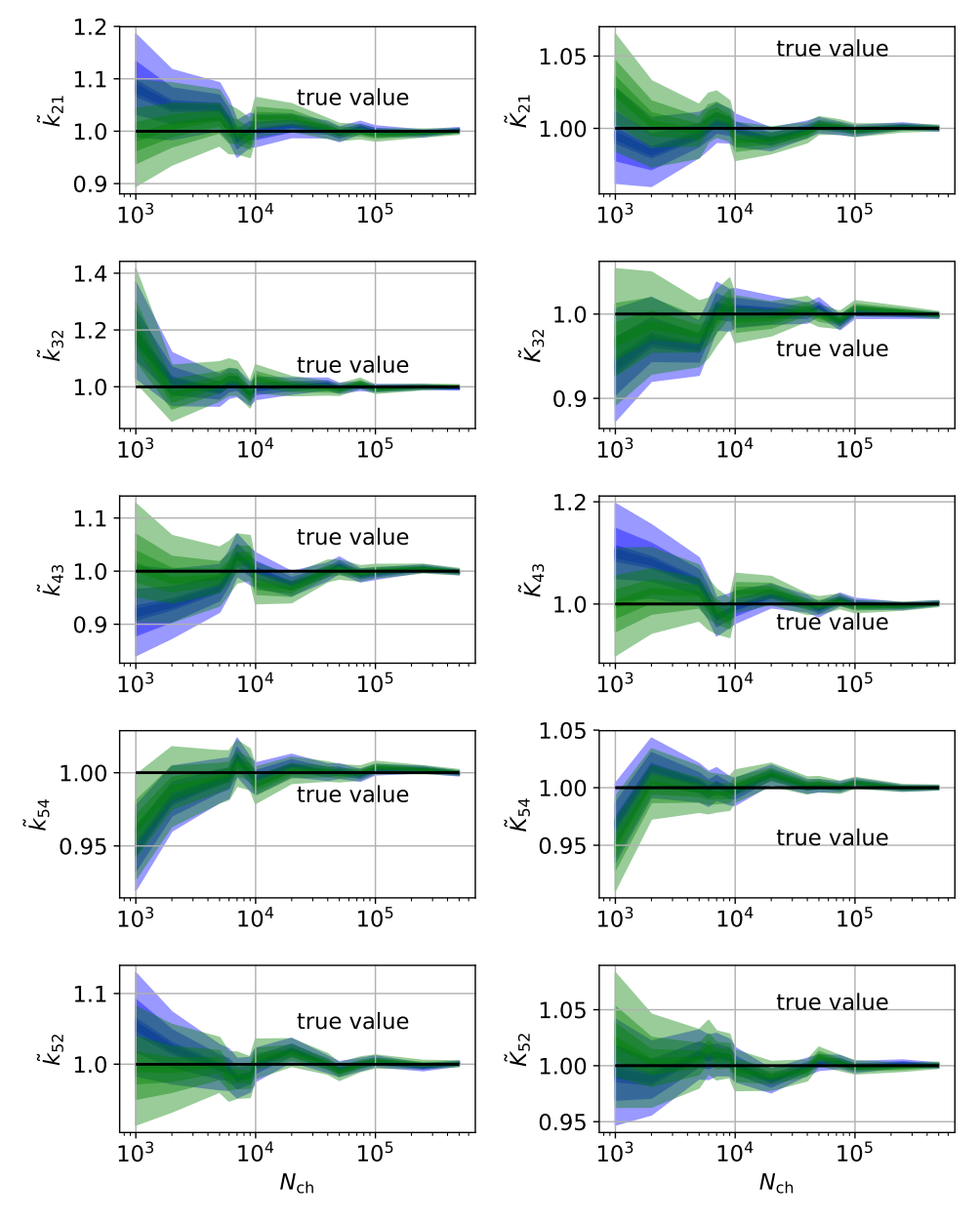


Figure 1. HDCIs vs. N_{ch} for the 5-state-2-open-states model for cPCF data. The single parameter level corresponds to the Euclidean error of Fig. 7 **d**. The prior that enforces microscopic-reversibility is identical to the prior mentioned in the Fig. 7 of the main text.

The effect of the second observable on the single parameter level for the 5-state model can be seen in App. Fig. 1. The biased estimates and the unidentified parameters of the RE approach for PC data (Fig. 9 of the main text) are eliminated by the fluorescence data.

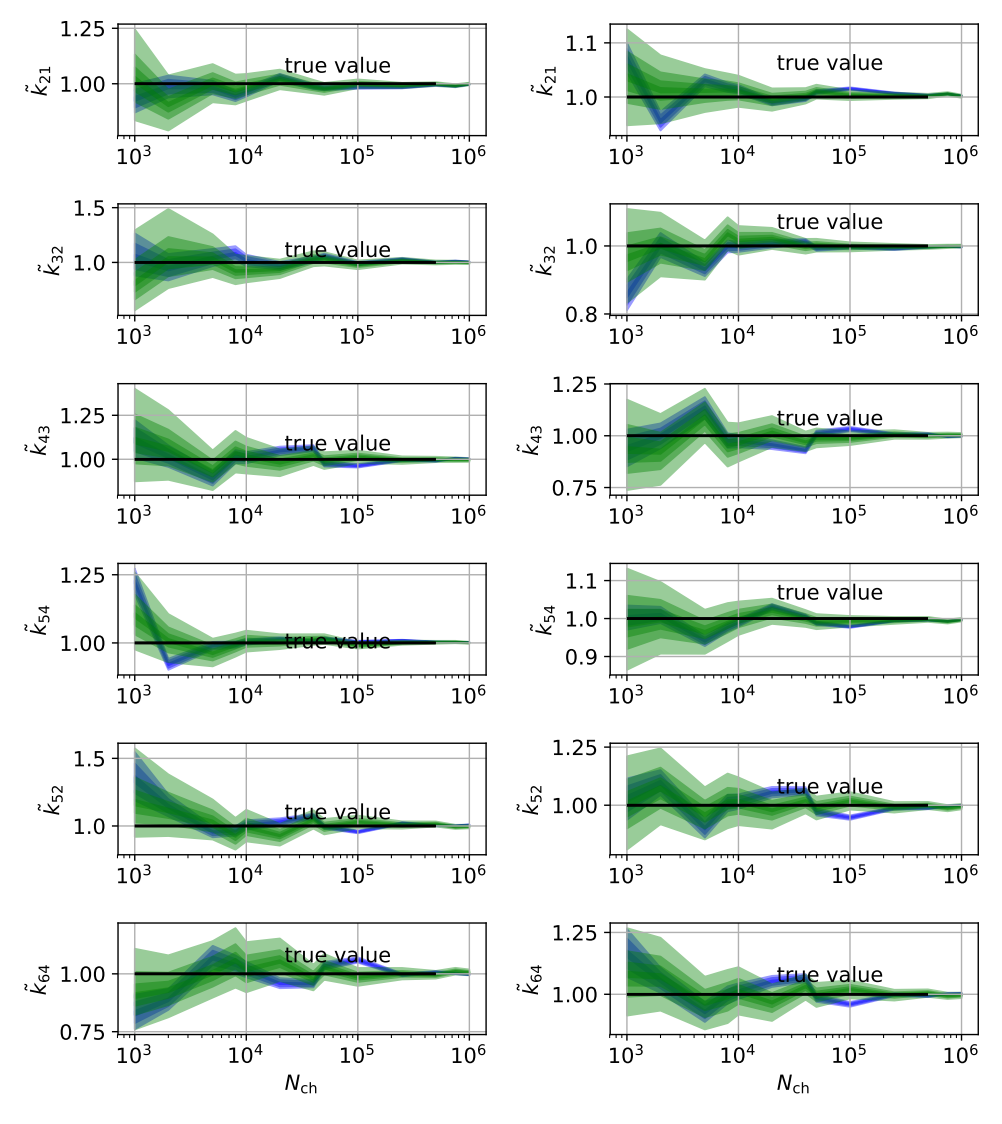


Figure 2. HDCIs vs. N_{ch} for the 6-states-1-open-state model for cPCF data. The single parameter level corresponds to the Euclidean error of Fig. 7 **e**. Compared to the main text we switched rows with columns of the panels, the first column represents now the rates and the second column the equilibrium constants. The prior that enforces microscopic-reversibility is identical to the prior mentioned in the Fig. 7 of the main text.

Even the over-confidence problem seems to be decreased (compare with Fig. 4**b1-6**). In contrast, the HDCIs of the 6-states-1-open-state model for cPCF data (Appendix 9 Fig. 2) display again a much increased over-confidence problem of the rate equation approach. This indicates that not simply counting of states but the actual topology of the kinetic scheme and the structure of the observation matrix $\mathbf{H}$ determines the scale of the over-confidence problem. Apparently, the more information comes from the fluorescence data relative to the information from the current data the less grave is the over-confidence problem and the less different are the Euclidean errors.