Bayesian selection of Hidden Markov models for multi-dimensional ion channel data

Jan 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 the complex conformational dynamics of ion channels from ensemble currents is a daunting task due to limited information in the data leading to poorly determined model inference and selection. We address this problem with a parallelized Kalman filter for specifying Hidden Markov Models for current and fluorescence data. We demonstrate the flexibility of this Bayesian network by including different noises distributions. The accuracy of the parameter estimation is increased by tenfold compared to fitting Rate Equations. Furthermore, adding orthogonal fluorescence data increases the accuracy of the model parameters by up to two orders of magnitude. Additional prior information alleviates parameter unidenfiability for weakly informative data. We show that with Rate Equations a reliable detection of the true kinetic scheme requires cross validation. In contrast, our algorithm avoids overfitting by automatically switching of rates (continuous model expansion), by cross-validation, by applying the 'widely applicable information criterion' or variance-based model selection.

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 ***Sakmann*** (***2013***). The gating of ion channels is usually interpreted by kinetic schemes which are inferred from macroscopic currents with rate equations (REs) ***Sakmann*** (***2013***) 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***). 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.

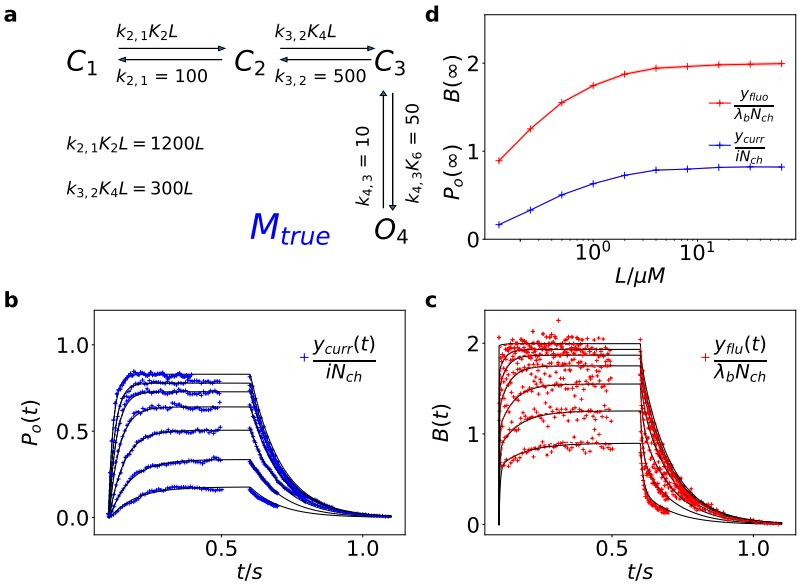
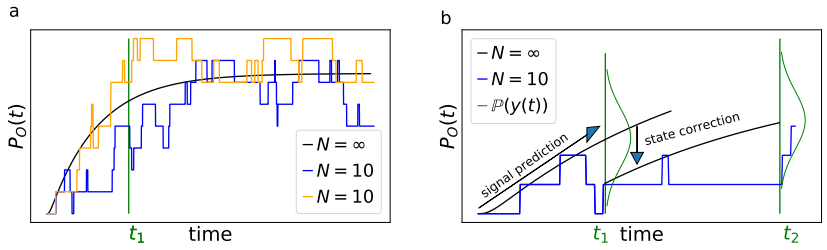


Figure 1. HMM used for the simulations. a, The kinetic scheme M_{true} used for simulating the data. The Markov state model (kinetic scheme) consists of two binding steps and one opening step. The rate matrix $\mathbf{K}$ is parametrized by the absolute rates $k_{i,j}$, the ratios K_i between on and 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 $i = 1$. **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 $\sigma^2_{\text{m}} = i^2$ and open-channel noise $\sigma^2_{\text{op}} = 0.1i^2$. 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. The ligand concentrations are $0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 64\,\mu\text{M}$. **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 rate equations (REs) which yield a point estimate of the rate matrix or its eigenvalues ***Colquhoun et al.*** (***1997***); ***Sakmann*** (***2013***); ***Alcantara et al.*** (***2002***); ***Wang et al.*** (***2012***). 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 (stochastic 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. Some rigorous attempts to incorporate the intrinsic noise of current data into the estimation ***Celentano and Hawkes*** (***2004***) suffer from cubic computational complexity in the amount of data points, rendering the algorithm impractical for real data. Stepanyuk suggested a faster algorithm ***Stepanyuk and Borisyuk*** (***2011***); ***Stepanyuk et al.*** (***2014***). Advanced approaches to analyze single-molecule data such as HMMs make use of solutions of the stochastic CME ***Jahnke and Huisinga*** (***2007b***) ***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 rates in addition to its signal observation probability distribution ***Rabiner*** (***1989***). HMM

Box 1. Illustration of two statistical problems in patch-clamp recordings addressed by a Bayesian network



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, the deviation at one time-point depends on the deviation on the previous time point **b,** Conceptual idea of the Kalman Filter (KF): the stochastic evolution of the ensemble signal is predicted and the prediction model updated recursively.

The 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 from the stochastic gating 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 (see blue and orange curves) from the mean time trace (black) that occur in real data measured from a *finite* ensemble. Assuming a white-noise process is never met in real data due to the Markovian nature of the system. **b,** In order to account for the autocorrelation in the signal, an optimal prediction of the signal distribution $\mathbb{P}(y)$ at the future time step t_2 should use the measurement y form the current time step t_1 to update the belief about the underlying hidden ensemble state $\mathbf{n}(t_1)$. Based on stochastic modelling of the time evolution of the channel ensemble, it then predicts $\mathbb{P}(y(t_2))$.

approaches are more accurate than fitting dwell time distributions for noisy recordings of rapidly gating channels ***Venkataramanan and Sigworth*** (***2002***) but the computational complexity 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 of the probability distribution of the hidden channel states. Notably, the KF is optimal in producing a minimal prediction error for the mean state. KFs have been used previously in several protein expression studies ***Komorowski et al.*** (***2009***); ***Finkenstädt et al.*** (***2013***); ***Fearnhead et al.*** (***2014***); ***Folia and Rattray*** (***2018***). 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.
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 10^4 channels per time trace, the KF outperforms the deterministic approach in estimating the model parameters and model selection. Moffatt ***Moffatt*** (***2007***) already demonstrated the advantage of the KF to learn absolute rates from time traces at equilibrium. Other benefits are the ability to infer the number of channels N_{ch} for each time trace, the single-channel current i and

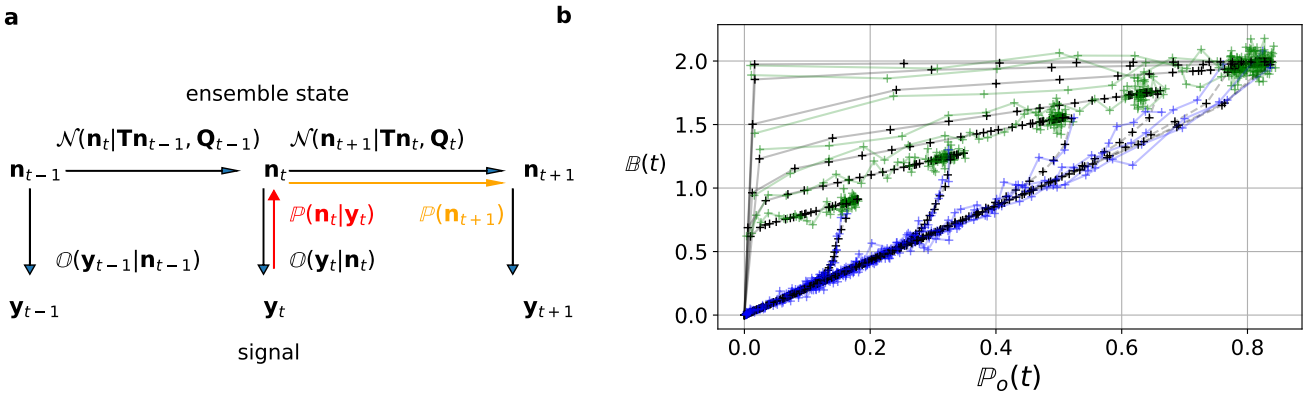


Figure 2. KF as Bayesian network. a, Graphical model of the conditional dependencies of the stochastic process. Horizontal black arrows represent the conditional multivariate normal transition probability $\mathcal{N}(\mathbf{n}_{t+1}|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t)$ of a continuous state Markov 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.

the mean number λ_b of photons from bound ligands per recorded frame. Thus no error-prone normalizations of the signal typical for deterministic (i.e. averaging) approaches is needed. The KF provides a likelihood which makes it possible to combine the time trace data during analysis with any other data that admits modelling with a likelihood.

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 maximum likelihood (ML) estimation ***Moffatt*** (***2007***) and did not derive model selection strategies. 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 inference provides outmatching tools for modeling: First, information from other experiments, simulations or from theory can be integrated through prior probabilities. Hence, uncertainties in the model parameters prior to the experiment are correctly accounted for in the analyses of the new data. For weakly informative data we demonstrate the beneficial effect of incorporating theoretical knowledge such as diffusion limited binding by prior distributions onto the posterior. Second, 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***). Third, a Bayesian approach provides superior model selection tools for singular models such as HMMs ***Kienker*** (***1989***).

The best fitting model will be defined as that one with the highest predictive accuracy, estimated either by cross-validation against held-out test data or by information criteria ***Gelman et al.*** (***2014***). Information criteria allow for model testing on training data instead of hold-out data by performing a bias-corrected computation of the predictive accuracy ***Gelman et al.*** (***2014***). We use the recently developed Widely Applicable Information Criterion (WAIC) ***Watanabe*** (***2010***) relying on the Bayesian paradigm. In contrast to its predecessor, the Akaike Information Criterion (AIC), WAIC asymptotically approximates the predictive accuracy of the model correctly, even for singular

models ***Watanabe*** (***2010***) such as HMMs or KFs. Moreover, we show that fitting current data with REs, both AIC and WAIC fail to detect overfitting, which demonstrates the importance of correctly modeling the intrinsic noise. Additionally, we propose a second-moment based model selection criterion which is enabled by the KF and improved by simultaneous measurement of fluorescence and current signals using cPCF.

Results and Discussion

Simulation of relaxing cPCF data

As an exemplary HMM we assume a ligand-gated channel with two ligand binding steps and one open-closed isomerization (see Fig. 1a). We define the ensemble state vector

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

which counts the number of channels in each state $\mathbf{s}$ (see Methods). A qualitative description of two statistical problems inherent in a stochastic a time series with an RE approach and the Basic idea of the KF is outlined in Box. 1. At first we assume that the fluorescence signal originates only from bound ligands (Fig. 3). Later also the signal of unbound ligands and the correction using a reference dye will be included (see Figs. 4-7, Appendix, and Methods section). Example data are shown in Figs. 1b-d.

Kalman filter derived from a Bayesian network

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 network (see Methods), i.e. a continuous state HMM with a multivariate normal transition probability ***Ghahramani*** (***1997***) (Fig. 2**a**). To make use of the KF, we assume the following general form of the dynamic model: The evolution of the hidden state vector $\mathbf{n}(t)$ is determined by a linear model that is parametrized by the state evolution matrix $\mathbf{T}$

$$\mathbf{n}_{t+1} \sim \mathcal{N}(\cdot|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t) = \mathbf{T}\mathbf{n}_t + \boldsymbol{\omega}_t, \tag{2}$$

where $\sim$ means sampled from and $\mathcal{N}$ is a shorthand for the multivariate normal distribution. 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 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 determines the noise $\mathrm{cov}[\boldsymbol{\omega}_t, \boldsymbol{\omega}_t] = \mathbf{Q}(\mathbf{T}, \mathbf{n}_{t-1})$ (Methods Eq. 34d). In short, Eq. 1a defines a continuous state 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 $\mathbf{H}$.

$$\mathbf{y}_t \sim \mathbb{O}(\cdot|\mathbf{H}\mathbf{n}_t) := \mathcal{N}(\cdot|\mathbf{H}\mathbf{n}_t, \boldsymbol{\Sigma}_t) = \mathbf{H}\mathbf{n}_t + \boldsymbol{\nu}_t \tag{3}$$

The noise of the measurement setup (Appendix 3 and Eq. 39) 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. 3 defines the state-conditioned observation distribution $\mathbb{O}$ (Fig. 2**a**).

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 network (Fig. 2**a**), exploits the conditional dependencies of the assumed stochastic process. A Bayesian network recursively predicts (prior) distributions for the next $\mathbf{n}_t$

$$\mathbb{P}(\mathbf{n}_t) = \int \mathbb{P}(\mathbf{n}_t|\mathbf{n}_{t-1})\mathbb{P}(\mathbf{n}_{t-1}|\mathbf{y}_{t-1})\mathrm{d}\mathbf{n}_{t-1}, \tag{4}$$

given what is known at time $t-1$. The KF as a special Bayesian network assumes that the transition probability is multivariate normal according to Eq. 2a

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

Each prediction of $\mathbf{n}_t$ (Eq. 5) is followed by a correction step,

$$\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t) = \frac{\mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)}{\int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)\,\mathrm{d}\mathbf{n}_t}, \tag{6}$$

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

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

If the initial prior distribution is multivariate normal then due the mathematical properties of the normal distributions all priors and posteriors $\mathbb{P}(\cdot)$ in Eq. 3b and 4b become multivariate normal ***Chen et al.*** (***2003***). In this case one can derive algebraic equations for the prediction (Methods Eq. 33 and 34d) and correction (Methods Eq. 54 and Eq. 54) of the mean and covariance. 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. 7 is the normal distributed marginal likelihood $\mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})$ for each data point, which constructs by

$$\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \prod_{t=2}^{N_T} \mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta}) = \prod_{t=2}^{N_T} \int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})\,\mathrm{d}\mathbf{n}_t = \prod_{t=2}^{N_T} \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}[\mathbf{n}_t], \mathbf{H}\mathbf{P}_t\mathbf{H}^\mathsf{T} + \mathbf{\Sigma}_t), \tag{8}$$

a product marginal likelihood of normal distributions of the whole time trace $\mathcal{Y}_\mathrm{T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{N_\mathrm{T}}\}$ of length N_T for the KF. For the derivation of $\mathbf{P}_t$ and $\mathbf{\Sigma}_t$ see Methods Eq. 34d and Eq. 39. The likelihood for the data allows to ascribe a probability to the parameters $\boldsymbol{\theta}$, given the observed data (Methods Eq. 18). An illustration for the operation of the KF on the observation space (Fig. 2**b**).The predicted mean signal $\mathbf{H}\mathbb{E}[\mathbf{n}(t)]$ and the data are plotted as vector trajectories.

For signals with Poisson-distributed photon counting or open-channel noise Eq. 7 becomes intractable. By applying the theorem of total variance decomposition ***Weiss*** (***2005***), we derive the output statistics that approximate various forms of noise and cast them into the form of Eq. 3 (Methods Eq. 53). The Bayesian posterior distribution

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) \sim \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}) \tag{9}$$

encodes all information from model assumptions and experimental data used during model training (see Methods). Our algorithm reconstructs the posterior (Fig. 3**a**) by sampling from it with the Hamiltonian Monte Carlo (HMC) method ***Hoffman and Gelman*** (***2014***); ***Betancourt*** (***2017***) provided by the STAN software ***Gelman et al.*** (***2015***).

Benchmarking of the KF against REs

For the synthetic time traces the KF samples from the posterior (Fig. 3**a**,**b**). For realistic channel numbers as 10^3 per patch, the posterior of the KF contains the true parameter values within the bounds of its 5th and 95th percentile (Fig. 3**b**). However, for typical experimental settings the total parameter error of the RE estimates ***Moffatt*** (***2007***); ***Milescu et al.*** (***2005***), calculated as the Euclidean distance to the true values, is roughly 10 times larger than the corresponding error of the posterior median of the KF to the true values. (Fig. 3**c**). It is noteworthy that, even for 10^4 channels per patch, the precision of the KF is 4 times higher than that of the RE model on that data set. Dividing the error of all estimates from the REs approach for $N_\mathrm{ch} = 10^4$ by the error of the KF estimates for $N_\mathrm{ch} = 10^3$ gives a ratio $0.97 \approx 1$. This means that analysis of the same data with the KF yields an improvement of model quality that REs could only match with a tenfold increase in the numbers of channels analysed. This result confirms that the KF approach is superior to the RE approach as already discovered when comparing the two methods with current data alone ***Moffatt*** (***2007***). For small N_ch, the ratio of the errors decreases like $\sim 1/\sqrt{N_\mathrm{ch}}$ (Fig. 3**c**). Thus the RE approach scales like $\sim 1/N_\mathrm{ch}$ and does not simply scale like the inverse square root of the system

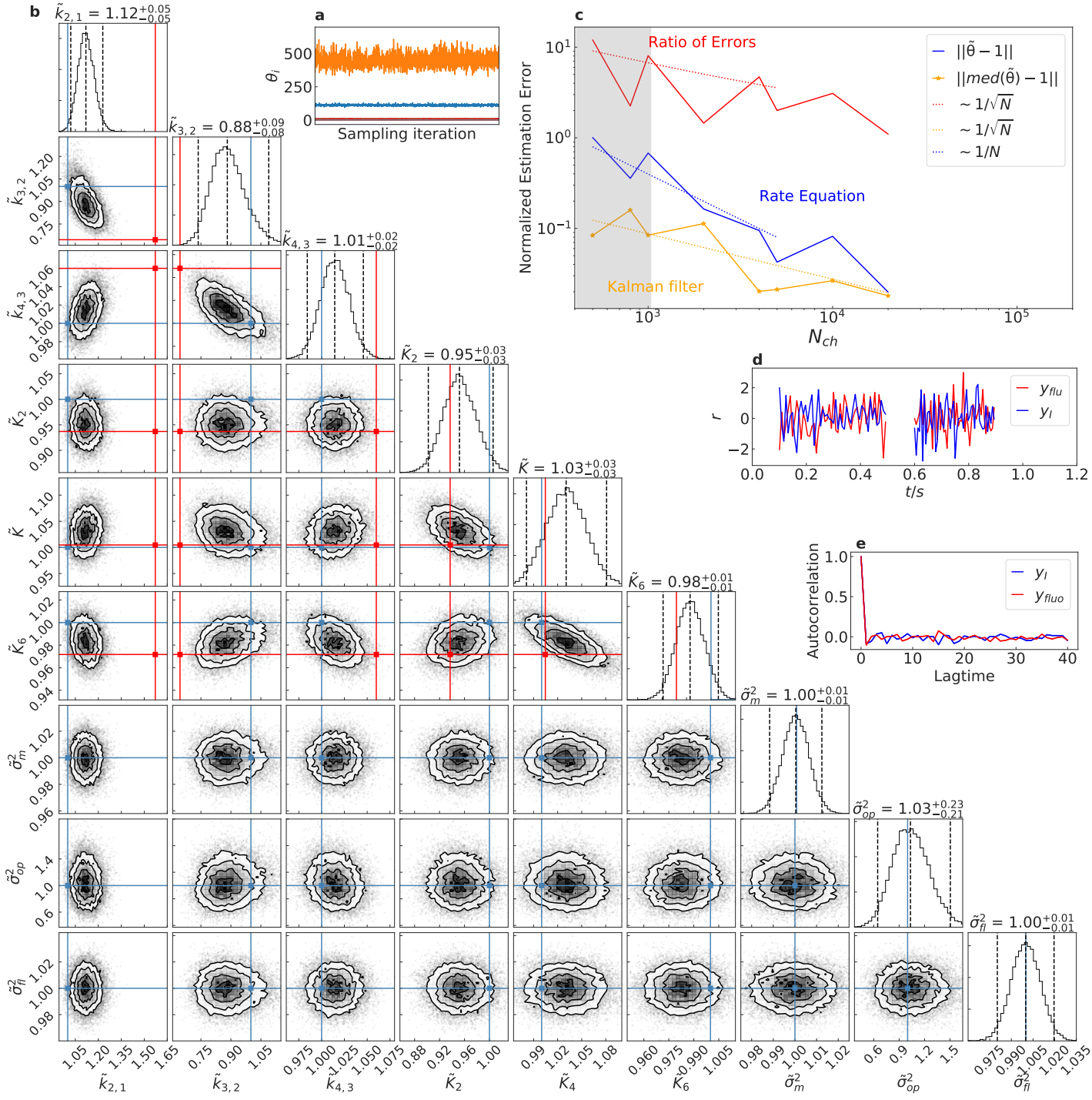


Figure 3. Benchmark of the KF against least squares fitting of REs. By $\tilde{\theta}_i$ we indicate that the samples from the posterior are scaled by their true value. **a,** Sample traces of 3 representative parameters from the posterior distribution of the KF algorithm created by Hamiltonian Monte Carlo sampling. The posterior is constructed by those samples. **b,** Posterior distribution plotted against the point estimate of a least squares fit with REs for $N_{\mathrm{ch}} = 10^3$. The blue lines represent the true values used for simulating the data, the red lines are their estimate from the RE fits. The dashed black lines show the quantiles (0.025, 0.5, 0.975) of the posterior. All values are normalized by their true value. Parameters which are not possible to infer from the mean values alone i, N_{ch} and λ are used as fixed input parameters to make both approaches comparable. **c,** Absolute errors of the median for the rate and equilibrium constants obtained by the KF (orange) and from the REs (blue) are plotted against N_{ch} . Error ratio (red) between both approaches scales according to $1/\sqrt{N_{\mathrm{ch}}}$ at least for smaller N_{ch} which is the expected scaling since the intrinsic noise divided by the mean signal scales in the same way. One expects an asymptotic equivalence for large N_{ch} between KF and REs since the signal to noise ratio diverges. The typical experimental situation $N_{\mathrm{ch}} \propto 10^2 - 10^3$ is indicated by the area shaded in gray. **d,** Time trace of median of the normalized residuals r_{curr} (blue) and r_{fluo} (red) for one ligand concentration after analyzing with the KF. **e,** The autocorrelation function of r from the KF shows the expected white-noise process.

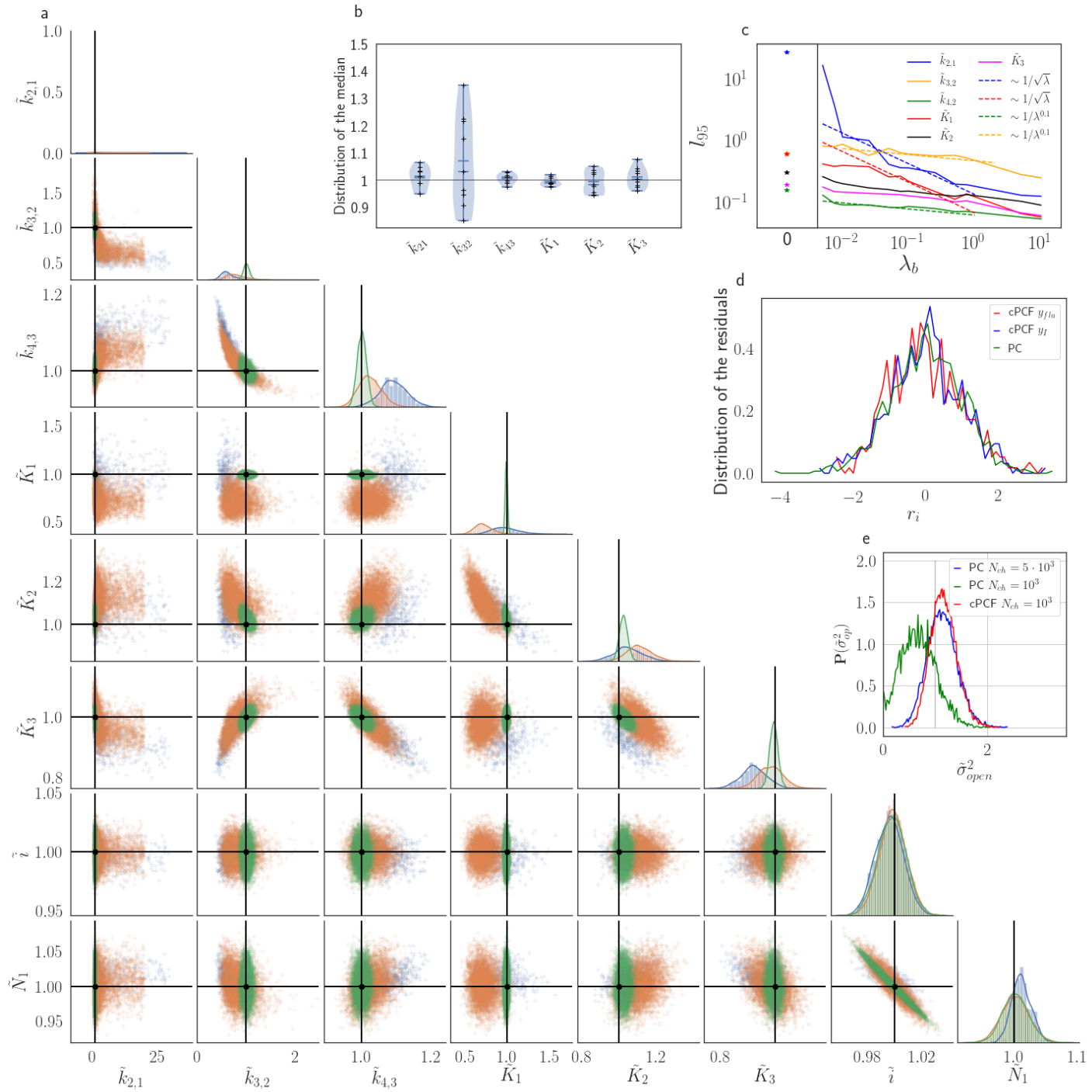


Figure 4. Benchmark of the KF for patch-clamp versus patch-clamp fluorometry data. a, Posteriors of PC data (blue), cPCF data with $\lambda_b = 0.00375$ (orange) and cPCF data with $\lambda_b = 0.375$ (green) but accounting for the superimposing fluorescence of unbound ligands in solution. The black 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 absolute error of the median for the parameters of the rate matrix for 9 different data sets, with $\lambda_b = 0.375$ and superimposing bulk signal **c**, The 95th percentile of the marginalized posteriors vs. λ_b normalised 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 are well described by $r_i \sim \text{normal}(0, \sigma^2_{\text{res}} = 1)$. The estimated variance is $\sigma^2_{\text{res}} = 0.98 + 0.26$. Note that the fluorescence signal per frame of is very low such that it is skewed. **e**, Posterior of the open-channel noise $\mathbb{P}(\sigma^2_{\text{op}}/\sigma^2_{\text{op,true}})$ for PC data with $N_{\text{ch}} \cdot 10^3$ (green) and $N_{\text{ch}} \cdot 10^5$ (blue) as well as for cPCF data with $N_{\text{ch}} \cdot 10^3$ (red) with $\lambda_b = 0.375$. Adding fluorescence data is roughly equal to five times more ion channels to estimate σ^2_{op}. We assumed as prior for the instrumental variance $\mathbb{P}(\sigma^2) = \mathcal{N}(1, 0.01)$.

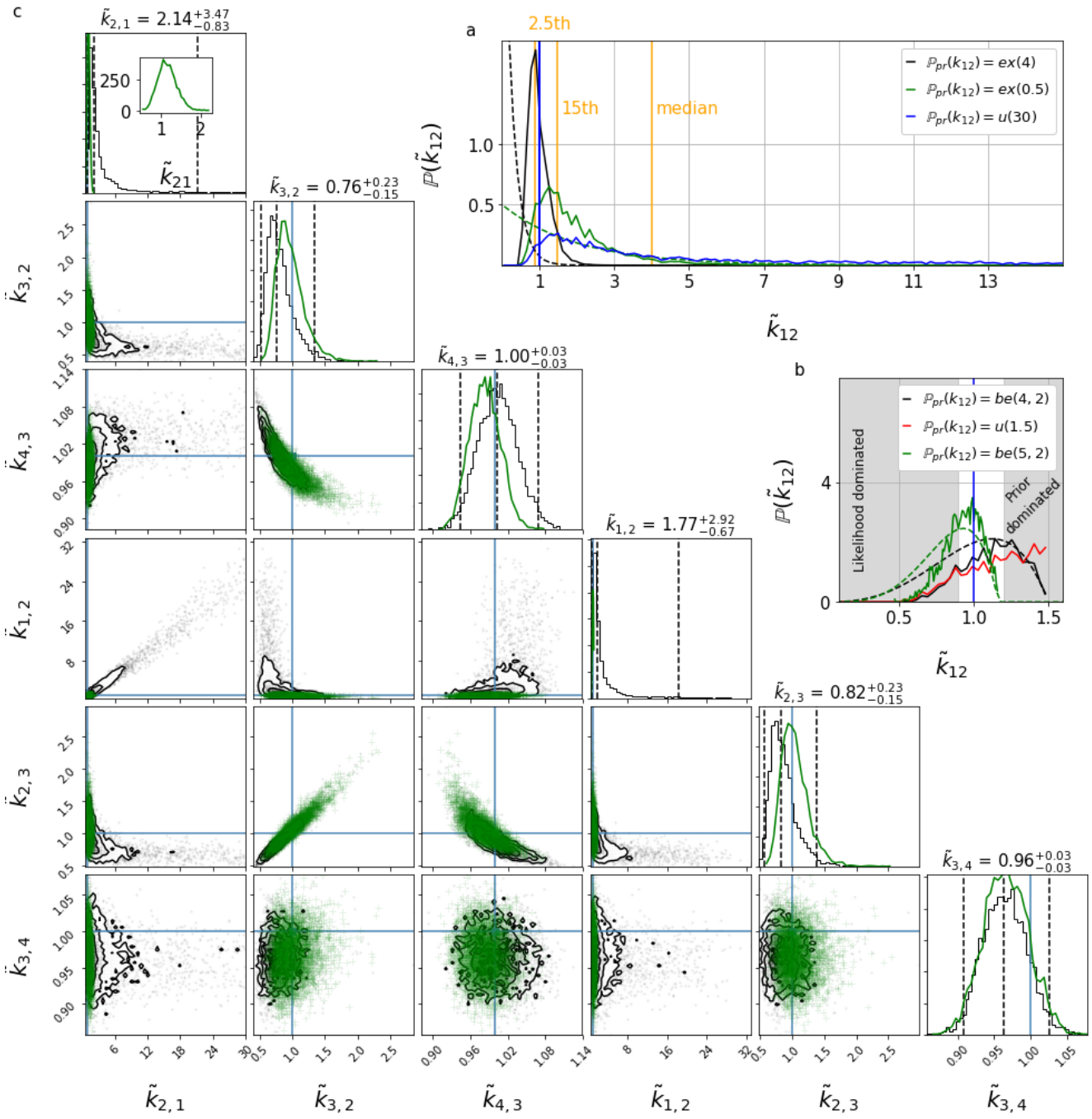


Figure 5. Variations of the posterior under different prior assumptions All posterior samples of θ_i are plotted scaled by their true values as $\tilde{\theta}_i$. True values are indicated by blue lines. The vertical orange lines represent the median and 2.5 and 15-percentile. **a,** Posterior $\mathbb{P}(k_{1,2})$ of the PC data from M_{true} as used throughout this article. Posterior derived from different priors. The priors are indicated by dashed curves. The uniform prior leads to a heavy tail in the posterior (blue). Note that the 97.5–percentile is at $\tilde{k}_{1,2} = 28.9$ and the median is at $\tilde{k}_{1,2} = 4$. In contrast, the exponential tails of the other priors dominate their posteriors in the tails. Even if we set the prior mean value 4 times smaller then the true rate (black dashed curve), the posterior (black) is still better centered around the true value than the posterior with the uniform prior. **b,** Since $k_{1,2}$ describes the ligand binding, theoretical predictions of a diffusion-limited binding rate can be used as a maximum in an informative prior. A beta distribution as the prior can be tuned $\frac{k_{1,2}}{1800} \cdot \text{s}\mu\text{M} = \tilde{k}_{1,2} \sim beta(4, 2)$ to favor ligand binding between $\tilde{k}_{1,2} \in [0.6, 1.4]$ (black, dashed curve). This results in a posterior which is likelihood-dominated below $\tilde{k}_{1,2} = 0.9$ while it is dominated by the prior above $\tilde{k}_{1,2} = 1.1$. A uniform prior with the same support as the beta prior results in a posterior with more weight above the theoretical possible range (red), where it is prior dominated. Thus, the difference between black and red posteriors indicates the information which is added by the beta prior. In this case it penalizes too high values and pushes the probability mass towards the true value. A stricter beta distribution $\frac{k_{1,2}}{1400} \cdot \text{s}\mu\text{M} = \tilde{k}_{1,2} \sim beta(5, 2)$ results in a narrower posterior (green). The Bayesian update concentrates for both priors the posterior mass towards the true value. **c,** The consequence of the more informative beta prior (green dashed line in b) on the posterior of the complete rate matrix. Green dots result from this prior, black lines are from a uniform prior $u(30)$. The inset zooms onto the collapsed posterior.

size, as one might assume. In Fig. 3**d**, the normalized residuals of one time trace are shown which are defined as

$$r_i := \frac{y_i - (\mathbf{H}\mathbb{E}[\mathbf{n}])_i}{\sqrt{\mathrm{var}[y_i]}}. \tag{10}$$

We normalize with respect to the predicted standard deviation for each data point $\sqrt{\mathrm{var}[y_i]}$ given by the KF. If the synthetic data are fitted with the true model, one expects to find a white-noise process for the residuals. Plots of the autocorrelation function of both signal components confirm our expectation (Fig. 3e). The estimated autocorrelation vanishes after one multiple of the lag time (the sampling interval), which means that the residuals are indeed a white-noise process. Estimating the residuals from RE would lead to correlated residuals ***Moffatt*** (***2007***), which is one reason for less precise parameter estimates.

cPCF versus patch clamp only

To evaluate the advantage of cPCF data ***Biskup et al.*** (***2007***) with respect to PC data only (Fig. 4), we compare different types of ligands: Idealized ligands with brightness λ_b, emitting light only when bound to the channels, and 'real' ligands which also produce background fluorescence when diffusing in the bath solution (Appendix 3). 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. 4**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. 4**a**). All median estimates of nine different cPCF data sets (Fig. 4**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. 4**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. 4**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 included (Fig. 4**c**). The estimated variance of r_i for PC or cPCF data is $\sigma^2(r_i) \approx 1$ (Fig. 4**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 pior information. Nevertheless, we show that for similar data sets the true underlying process can still be determined (Fig. 6**g** and Fig. 7**b**). Remarkably, the KF allows to determine in a macropatch the variance of the open-channel current noise for $\sigma_\mathrm{op} = 0.1i$, i.e. when the total noise is dominated by the much larger gating noise $(\mathbf{H}\mathbf{P}_t\mathbf{H}^\top)_{2,2}$ (Fig. 4**e**): For the saturating ligand concentration $p_{\mathrm{o,max}} = 0.833$, i.e. the expected open probability of the true process, the ratio at equilibrium is

$$\frac{(\mathbf{H}\mathbf{P}_t\mathbf{H}^\top)_{2,2}}{\sigma_\mathrm{op}^2} \approx \frac{1 - p_{\mathrm{o,max}}}{\sigma_\mathrm{op}^2} \geq 20. \tag{11}$$

Including theoretical limits and vague parameter knowledge into the analysis with different priors

One advantage of Bayesian statistics is that with prior distributions, one can account for partial knowledge about parameters and their uncertainties. While it is straightforward to use the posterior of a previous experiment as a prior for the data set, it is less obvious how to model notion of the plausible magnitude of a parameter into a prior. Here we propose some usable options for

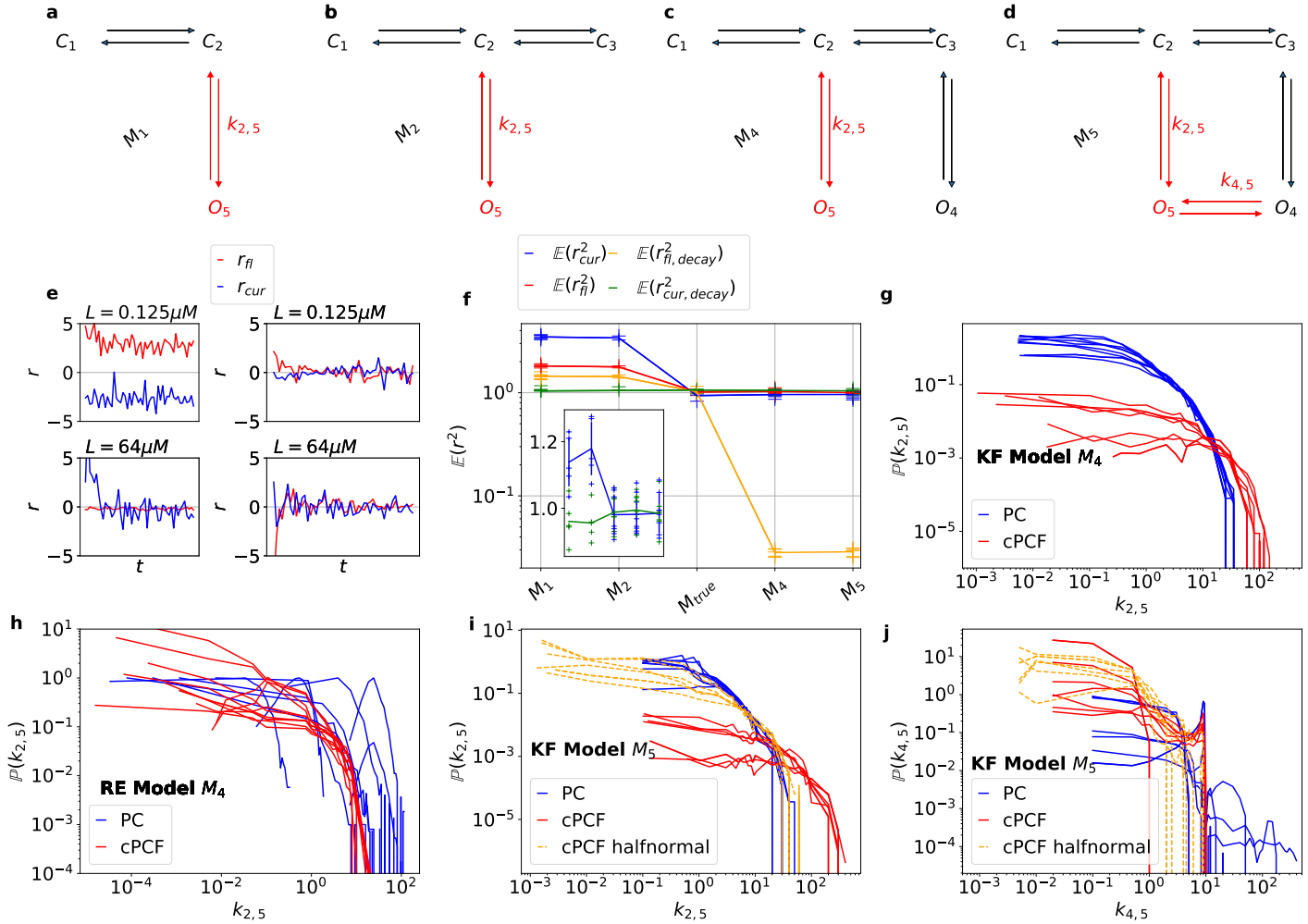


Figure 6. Model selection by the second moment of the residuals and by continuous model expansion. a-d, Model structures of the trained models differing from the true process (c.f. Fig. 1a). Red states and transitions have no counterpart in the true model. All models are nested in the most complex model M_5 (**d**). **e,** Time traces of residuals r of cPCF data from KF-fits for current data (blue) and fluorescence data (red) at two ligand concentrations L for the incorrect model M_2. Left: Jump to the ligand concentration. Right: Jump to zero. Systematic deviations from zero suggest that model M_2 is too simple. **f**, Second non-centralized moment of the residuals for all models. For the true model the second non-centralized moment becomes the variance with $\mathrm{var}[r_i] = 1$ because $\mathbb{E}[r_i] = 0$. The inset shows the second non-centralized moment of PC data for the five different models on a linear scale. While underfitting can be clearly detected, overfitting is not detected. cPCF Data increase the contrast of current residuals between underfitting and true process by ≈ 12.5. **g-j**, Continuous model expansion by model M_4 and M_5 from **c,d**. **g**, Posterior distribution $\mathbb{P}(k_{2,5})$ for a rate into a state which does not exist in the true process obtained by the KF. For current data only, 10 out of 10 data sets maximise the probability density for $k_{2,5} \to 0$. For cPCF data, 9 out of 10 data sets yielded the same result and only one data set has a finite maximum. Hence, the KF suggests to delete O_5. **h**, Same current data (blue) analyzed with a RE approach finds in 4 out of 7 data sets rates into the non-existing state. Adding fluorescence data (red) improves the analysis. Now in 8 out of 10 data sets the posterior is maximized for $k_{2,5} = 0$. **i,j**. The KF for various PC and cPCF data sets reduces M_5 to the true data-generating process M_{true} for most data sets by maximizing the posterior for a zero rate into O_5. For current only data 4 out of 7 data sets show this behavior. For cPCF 9 out 10 data sets maximise $\mathbb{P}(k_{2,5})$ for $k_{2,5} \to 0$ and $\mathbb{P}(k_{4,5})$ for $k_{4,5} \to 0$. Across M_4 and M_5 (**g** and **i**) the posteriors $\mathbb{P}(k_{2,5})$ show a similar shape. **j**, For cPCF data there is no probability mass for $k_{4,5} > 10$. Hence, for $k_{4,5}$ cPCF data constrain closer to zero then PC data but $k_{2,5}$ it is reverse. The dashed red lines belong to the same cPCF data but with a weakly informative halfnormal prior $k_{5,2} \sim \mathrm{halfnormal}(0, 6 \cdot 10^3)$ for one specific rate which we identified to be not confined by the data. This prior also reduces the magnitude of the peak in $k_{4,5}$ and thus suggests omitting state O_5.

prior distributions for $k_{1,2}$ that are related to the maximum entropy principle using diffusion limited binding as example.

Binding can not be faster than molecular encounters due to diffusion ***Smoluchowski*** (***1918***). For the first binding step we used an approximation for small ligand receptor interactions ***van Holde*** (***2002***)$k_{\text{binding}} = 600\,\mu\text{M}^{-1}\text{s}^{-1}$. Thus, for two available binding sites the stoichiometry increases the upper limit to $k_{1,2} = 1,200\,\mu\text{M}^{-1}\text{s}^{-1}$. Here we investigate priors with different information content to model the *a priori* plausibility for $k_{1,2}$ by making use of the mentioned diffusion limit. Traditional Bayesian or frequentist approaches, using uniform priors $\text{u}(l) = \frac{1}{l}$ if $k_{1,2} \in [0, l]$, perform well for the "strong data case". A uniform prior is a maximum entropy distribution under the condition that the only information available about the unknown rate is the interval of possible values, i.e. their support. A maximum entropy distribution adds the least information (is the most conservative assumption) to the posterior apart from the explicitly used conditions, which is in this case the support ***Jaynes*** (***1957***). In the strong data context, the posterior is dominated by the likelihood and the influence of the prior information is minor ***van der Vaart*** (***1998***). In contrast, for modeling situations with weakly informative data (Fig. 5 **a**) an educated prior selection influences the posterior to centre around the true values. For instance, the PC data are not informative enough to make the likelihood of $k_{1,2}$ contract in a small neighbourhood around the true value (see Fig. 4). Due to the uniform prior the corresponding posterior behaves accordingly (Fig. 5 **a**). The data are only weakly informative because larger $k_{1,2}$ can be partially compensated for by a larger $k_{2,1}$ and a smaller $k_{3,2}$ (Fig. 5 **c**). Nevertheless, all probability mass of $\mathbb{P}(k_{1,2})$ above the diffusion limit of binding (Fig. 5 **a**) is physically impossible, though plausibly given by the data, since the rates $k_{1,2}$ and $k_{2,3}$ are diffusion-limited or slower. Note that the estimated 15th-percentile is at $\tilde{k}_{1,2} = 1.47$ such that more than 85% of the probability mass lies in a physically impossible area.

The situation can be improved by supplying more plausible information about $k_{1,2}$ using an exponential distribution

$$\mathbb{P}(k_{1,2}) = \zeta \cdot \exp(-\zeta k_{1,2}). \tag{12}$$

The parameter ζ refers to the parameter which scales the statistics, $\mathbb{E}[k_{1,2}] = 1/\zeta$ and $\text{var}[k_{1,2}] = 1/\zeta^2$. Notably, the exponential distribution is a maximum entropy distribution if two conditions are met ***McElreath*** (***2018***). The parameter has to be positive and the mean of the parameter is known. On the one side the exponential prior succeeds in penalizing the heavy tails of the likelihood (Fig. 5 **b**). On the other side, even if we apply an exponential prior whose mean value is four times smaller than the true binding rates, the posterior (black) is still more concentrated around the true value than the posterior with the uniform prior.

Nevertheless, the exponential distribution is not well suited for our problem because it does not incorporate a hard upper limit. Even with the exponential prior, there is always some probability mass in areas which are physically impossible and, additionally, the exponential prior does not include that the response of the ion channel proceeds in a limited amount of time which means that $k_{1,2}$ and $k_{2,3}$ cannot be arbitrarily small. Thus it is unlikely that the true $k_{1,2}$, $k_{2,3}$ are by orders of magnitude slower than the diffusion limit. In fact, the exponential prior states the opposite: binding rates have the highest probability density at zero.

The beta distribution

$$\text{be}(a, b)\frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}, \tag{13}$$

where $B(\cdot, \cdot)$ defines the beta function, is a maximum entropy distribution derived from three conditions: that the support is $\theta_{1,2} \in [0, 1]$, that

$$E[\ln(\theta_{1,2})] = \Psi(a) - \Psi(a + b) \tag{14}$$

and that

$$E[\ln(1 - \theta_{1,2})] = \Psi(b) - \Psi(a + b), \tag{15}$$

where $\Psi(\cdot)$ symbolizes the digamma function. Since rescaling $k_{1,2} = l\theta_{1,2}$ by l adds only a constant term to the entropy and the entropy is translation invariant, we remain in the maximum entropy

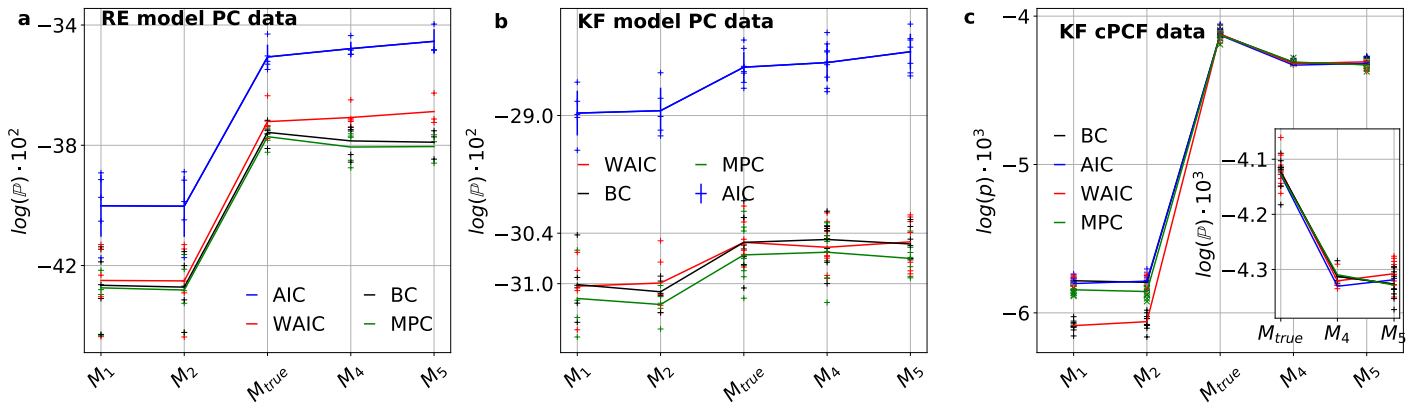


Figure 7. Bayesian Model selection Compared is the predictive accuracy of the indicated five models estimated by either the Akaike Information Criterion (AIC), the Widely Applicable Information Criterion (WAIC), Bayesian cross-validation (BC) or Maximum *a posteriori* cross-validation (MPC). We use BC on a single validation data set as an estimator of the predictive accuracy and evaluate the other criteria how they approximate BC. The solid lines represent the mean quality of the estimators over different data sets (crosses). **a**, Both information criteria fail to approximate the predictive accuracy if current-only data are modeled with deterministic REs. WAIC approximates the predictive accuracy better than AIC, though both information criteria suggest too complex models. Cross-validation (black and green) reveals the correct kinetic scheme. **b**, For current data analyzed with the KF, WAIC estimates the predictive accuracy obtained by Bayesian cross-validation (black) with high accuracy while AIC fails. WAIC predicts the BC value even better than MPC. **c**, Including the fluorescence, the difference in the predictive accuracy of the true model to the under-complex models increases strongly and all methods identify the right kinetic scheme. At the same time, the posterior has become multivariate normal by adding the second dimension to the data. Instead of an almost constant region in **b** for model M_{true} to M_5, there is a unique peak for model M_{true}. To identify this peak, we only needed to score the models by the data on the activation part of the time series. This is consistent with the observation of Fig. 6 **f** that the estimation of the decaying part of the fluorescence is very susceptible to overfitting. The residuals in the decaying part of the fluorescence data are smaller which results in a higher probability of those data points if fitted with M_4 or M_5. We did not observe this result in PC data. Note that ligand association happens over different trajectories in the observation space but ligand dissociation relaxes after a quick transition of a few data points onto a single trajectory (Fig. 2b). There is less diverse information about the deactivation about the rates. The inset shows a part of the diagram at an extended ordinate.

setting for every set of a, b, l. Therefore, we use the beta distribution to model the prior plausibility of $k_{1,2} \in [0, l]$ by setting hard constraints: positive but smaller than l. We then distribute the probability mass with respect to the vague idea of where to expect the binding rate within this interval. Thereby, we implicitly assume the conditions from above. $\tilde{k}_{1,2} := k_{1,2}/1,800 \cdot \mathrm{s}\mu\mathrm{M} \sim \mathrm{beta}(4,2)$ constrains the posterior such that $k_{1,2}$ cannot be larger than $1800 \cdot \mathrm{s}^{-1}\mu\mathrm{M}^{-1}$ and we expect $k_{1,2}$ to be between $700 \cdot \mathrm{s}^{-1}\mu\mathrm{M}^{-1}$ and $1,700 \cdot \mathrm{s}^{-1}\mu\mathrm{M}^{-1}$. Even though this beta prior (Fig. 5 **b**) is an informative prior, a lot of the information derives from the support of the beta distribution as revealed by comparison for $\tilde{k}_{1,2} < 0.9$ with the posterior obtained with the uniform prior having the same support. In this unconstrained area the data are really informative. In contrast, for $\tilde{k}_{1,2} > 1$ the prior is the most important source of information for the posterior. For the green posterior (Fig. 5 **b-c**) we assume a little bit stricter limits and plausibility within the possible interval. $k_{1,2}/1,400 \cdot \mathrm{s}\mu\mathrm{M} \sim \mathrm{beta}(5,2)$. The data support the *a priori* plausibility assumptions by concentrating the posterior within the area which contains most probability mass of the prior (Fig. 5 **b**). Note, that for the other rates (which have not been constrained by an informative prior), the beta prior for $k_{1,2}$ improves their parameter inference by concentrating the posterior (Fig. 5 **c**) around the true values. That effect is strongest for $k_{2,1}$ due to the strong correlation with $k_{1,2}$ induced by the likelihood. The posterior has now areas in the parameter space which are strongly influenced be the shape of the beta distribution whereas other areas are shaped by the likelihood. Remarkably, despite some arbitrariness of the shape of the beta prior, it provides profit for the inference of all ligand-related rates. Furthermore, the restricted range speeds up sampling and thus reduces the computation time.

315 ## Bayesian Model selection by continuous model expansion or predictive accuracy

316 We compare three methods of model selection: continuous model expansion, the statistics of $r(t)$,
317 and the predictive accuracy estimated either by cross-validation or information criteria on a set of
318 candidate models $M_1, M_2, M_{\text{true}}, M_4, M_5$ (Fig. 6**a-d**).
319 The KF enables to identify underfitting better by plotting the residuals $r(t)$ (Fig. 6**e**) rather than sig-
320 nal time traces because the large amplitude changes of the mean current obscures the relatively
321 small amplitude of the systematic errors. The estimated second moment of $r(t)$ is plotted for the
322 different models (Fig. 6**f**). For the true model M_{true} the estimated second moment equals the vari-
323 ance and, since we normalized the residual traces, the second non-centralized moment should be
324 close to 1. In fact all variance is explained by M_{true}. Overfitting models M_4 and M_5 are detected
325 by the decrease of the fluorescence variance in the decaying part of the traces. For PC data only,
326 underfitting can be detected. But as long as the modeler looks out for the simplest model which
327 does not underfit the detection of the true process is successful.
328 The conceptional idea of continuous model expansion is to sample from a model structure which
329 contains the true process and a lot of additional model structure whose rates are set to zero by
330 the algorithm when the data quality or quantity increases. In other words one assumes a complex
331 super model M_5 which includes all simpler models as a limiting case $k_{i,j} \to 0$. A simpler model can
332 be chosen if the posterior has a distinctive maximum for $k_{i,j} = 0$. Testing continuous model ex-
333 pansion by M_4 with the KF, the posterior for only one out of ten data sets shows a local maximum
334 for $k_{2,5} \neq 0$ (Fig. 6**g**). Thus, the KF switches off non-existing states for most data sets. By contrast,
335 the corresponding analysis of current-only data by REs reveals a peaking posterior for $k_{2,5} \neq 0$ in
336 several cases (Fig. 6**h**). Additional fluorescence data reduce the occurrence of those peaks.
337 If the PC or cPCF data are fitted with model M_5 by using the KF, for most data sets the rates into O_5
338 maximize the posterior if they vanish (Fig. 6**i-j**). For PC data, the posterior reveals multi-modality
339 with some data sets (Fig. 6**j**). Hence, point estimates of the parameters are not reliable while the
340 posterior of M_5 encodes all model uncertainties. Notably, this multimodality occurs also for cPCF
341 data though less pronounced. Thus, both experiments share the tendency to create a finite peak
342 around $k_{4,5} = 10$, indicating the false detection of an additional open state if not analyzed with cau-
343 tion.
344 Applying a weakly informative prior distribution supports the model determination in the contin-
345 uous model expansion case. The advantage of continuous model expansion is that it reduces the
346 risk of finding a local optimum on the discrete model space rather than a global optimum by trans-
347 lating the model space from a discrete to continuous model space. The disadvantage of having a
348 lot of possible structure in the model makes the model quickly too flexible to come to a conclusive
349 posterior with a limited amount of data. Many parameter sets can fit the data roughly equally likely.
350 Thus, prior distributions are needed to support the algorithm to select simpler base models, which
351 means concentrating for certain $k_{i,j}$ the posterior around zero ***Gelman et al.*** (***2017***). To exemplify
352 this, we use a weakly informative prior distribution on $k_{5,2}$ (see Appendix 1) to show how prior in-
353 formation alleviates model pathology due to excessive model flexibility. Heuristically, one should
354 be sceptical about rates which are faster than the sampling frequency because they could gener-
355 ate eigenvalues λ ***Sakmann*** (***2013***) of the rate matrix, which are smaller than the time between
356 two sampling points. Here we used a sampling frequency of $10\,\text{kHz}$ The frequency by which the
357 KF analysed the data ranged from $83.3\,\text{Hz}$ to $500\,\text{Hz}$ depending on the kinetics. The weakly infor-
358 mative half-normal prior $k_{5,2} \sim \text{half-normal}(0, 6 \cdot 10^3)$, which penalizes unrealistic high rates of $k_{5,2}$,
359 is necessary because the data are not able to constrain that rate. Applying this prior distribution
360 suppresses secondary peaks and probability mass in the distribution tails of the other rates $k_{2,5}$
361 and $k_{4,5}$. This further emphasizes to leave out O_5 in the final model (Fig. **i-j**).
362 We use this prior to argue that one loses a lot of descriptive power of a data set if one tries to be
363 objective by using uniform priors in particular with unrealistic large intervals. The notion of being
364 unbiased with a flat prior, where the likelihood does not dominate the prior, ends up in paradoxes

Zwickl and Holder (***2004***). A uniform prior on chemical rates leads to a non-uniform prior on the activation energies in the free energy landscape of the protein or any chemical reaction. Hence, the weakly informative prior acts as a guard against overfitting by suppressing fast rates beyond the experimental time resolution.

Notably, continuous model expansion fits many candidate models in one attempt with the exhaustive supermodel. This technique reduces the risk of getting into an impasse in a local optimum in the model space. Nevertheless, one should drop from the supermodel step by step the parts which are switched off and then refit. Optimally, this process should be accompanied by the estimation of the predictive accuracy ***Gelman et al.*** (***2014***). In particular in ambiguous situations with either multi-modality in the posterior or no clear maximum in the probability density at $k_{i,j} = 0$ the predictive accuracy is a well definied criterion.

Several other statistical approaches to identify the best fitting model have been reported ***Vehtari et al.*** (***2012***); ***Piironen and Vehtari*** (***2017***); ***McElreath*** (***2018***); ***Wallace*** (***2005***). Those approaches balance accuracy of the model's predictions with its simplicity. Note, the "simplicity" of a model is ultimately subjective because it depends on the choice of the (formal) language in which a model is described. ***Wallace*** (***2005***) Some of them, such as Maximum evidence or BIC, ***Vehtari et al.*** (***2012***) rely on the assumption that the true data-generating process is included in the set of models to be tested. This is called the M−closed situation ***Vehtari et al.*** (***2012***). These approaches perform well in simulation studies ***Bronson et al.*** (***2009***) if one is in fact in an M−closed setting.

Generally, when selecting a kinetic scheme for a protein, one reduces the high-dimensional continuous dynamics of the true data-generating process $\mathbb{F}(y)$ to a few discrete states and transition rates. The true data-generating process is therefore not included in the set of models from which the best fitting model is chosen which is the M−open setting ***Vehtari and Ojanen*** (***2012***). For this setting we define the best fitting model as the model which loses the least information, or adds the least entropy, if it is used as a proxy for $\mathbb{F}(y)$. In this way we are able to rank all models by their information loss. The information loss (or the increase of entropy) incurred by approximating one probability distribution (the true data-generating process) by another (the model) can be measured by the Kullback-Leibler divergence ***McElreath*** (***2018***) which is in principle not accessible. But the model with the minimum Kullback-Leibler divergence within a set of candidate models can be found asymptotically by maximizing the predictive accuracy ***Burnham and Anderson*** (***2004***); ***Gelman et al.*** (***2014***). The predictive accuracy for a specific data set $\tilde{\mathcal{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{N_{\text{data}}}\}$, which has not been used for model training ***Gelman et al.*** (***2014***), is defined as

$$\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}) = \log \mathbb{E}_{\theta}[\mathbb{L}(\tilde{\mathcal{Y}}|\boldsymbol{\theta})] \tag{16}$$

where $\mathbb{E}_{\theta}[\mathbb{L}(\tilde{\mathcal{Y}}|\boldsymbol{\theta})] = \int \mathbb{L}(\tilde{\mathcal{Y}}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y})\,\mathrm{d}\boldsymbol{\theta}$ means the average with regard to the posterior distribution. The mean predictive accuracy of a model for all possible data sets is the average over the unknown true data-generating process $\mathbb{F}(\tilde{y}, t)$

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})] = \int \log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})\mathbb{F}(\tilde{\mathcal{Y}}) \prod_{i=1}^{N_{\text{data}}} \mathrm{d}\tilde{y}_i \tag{17}$$

Maximizing Eq. 17 is equivalent to minimizing the Kullback-Leibler distance which specifies the information loss when approximating the unknown $\mathbb{F}(y)$ by a model ***Kullback and Leibler*** (***1951***); ***Burnham and Anderson*** (***2004***). Because $\mathbb{F}(y)$ is unknown, one has to estimate Eq. 17 by using the available samples from $\mathbb{F}(y)$. Therefore, different estimators of Eq. 17 are compared: AIC, WAIC, maximum *a posteriori* cross-validation and Bayesian cross-validation (see Methods). As a result, if current time series are modeled with REs, both information criteria fail to detect the overfitting model (Fig. 7**a**). This means, ignoring intrinsic fluctuations of macroscopic data, such as RE approaches do, leads to the inference of more states than present in the true model if the model performance is evaluated by the training data. One can exploit this weakness by choosing the kinetic scheme on cross-validated data, since too complex models derived from RE do not generalise as good to new data as a parsimonious model. Their additional states do not contribute positively

to the predictive accuracy of the model.

As expected from theory, WAIC ***Watanabe*** (***2010***) succeeds (Fig. 7**b**) while AIC fails to estimate the predictive accuracy distribution. It suggests a better predictive accuracy with more complex model structures (Fig. 7**b**). The failing AIC correlates with the occurrence of a non-normal posterior distribution ***Watanabe*** (***2007***)(see, Methods). The mean predictive accuracy of WAIC and BC (black and red) for M_4 and M_5 is only slightly smaller than that for M_{true} which can be explained by the observation that the KF automatically finds a sharp peak at $\mathbb{P}(k_{2,5} = 0)$ (Fig. 6**g-j**). This can be used as a model selection strategy: If two models have a similar predictive accuracy and are nested one should check whether the posterior of the larger model switches off certain rates. The predictive accuracy not only scores a kinetic scheme. It also evaluates how closely the whole algorithm mimics the true process. Comparing the predictive accuracy of the true kinetic scheme M_{true} in (Fig. 7**a** and **b**) reveals the higher precision of the KF in modeling the intrinsic noise compared to the RE approach. For multidimensional cPCF data all methods yield similar predictive accuracies and select the true data-generating process (Fig. 7**c**), as a unique peak for the true process is observed for all data sets.

Conclusion

We derived the prediction (Methods Eq. 33 and 34d), the output statistics (Methods Eq. 53) and correction equations (Appendix 4) of the KF for analyzing the gating and binding dynamics of ion channels including open-channel noise, photon-counting noise and background noise. For the correct kinetic scheme the parameter estimates obtained by the KF are ~ 10 times as good when applied to the same data set (Fig. 3**b,e**). Furthermore, enriching the data by fluorescence based ligand binding increases the accuracy of the parameter estimates up to $\sim 10^2$-fold (Fig. 4**a,c**). In the case of weakly informative data we show the superiority of informative priors to constrain the posterior to physically reasonable values. In this case the interaction between data and the prior information enables a much more meaningful model inference (Fig. 5**a-d**) compared to using flat priors. Moreover, we showed that overfitting can be detected by continuous model expansion (Fig. 6**g-i**). Usually the KF maximizes its posterior by abolishing a rate into an non-existing state. This is not the case if the current time traces are analyzed by REs (Fig. 6**h**). The potential weakly informative prior on one critical rate which increases the accuracy of continuous model expansion approach (Fig. 6**i-j**). We demonstrated that the information criterion WAIC performs much better in approximating the predictive accuracy than traditional information criteria based on point estimates such as AIC (Fig. 7**b**). We are even able to predict the correct kinetic network in cases were the data are insufficient for creating a multivariate normal posterior (Fig. 7**b**). Another relevant aspect is that both information criteria fail to predict the true kinetic scheme if the data are analyzed by deterministic REs (Fig. 7**a**). To select a model, one should apply WAIC and BC to multiple data sets, considering their dependency on noisy data. For the RE approach, only cross-validation revealed the true data-generating process. Model selection of kinetic schemes should not be done on training data if the analysis has been done by REs. For cPCF data we could detect the true kinetic scheme with the second moment of the residuals. For the true model the empirical $r(t)$ are close to the expected variance $\text{var}(r) = 1$ and overfitting is revealed distinctively by the variance of the fluorescence signal (Fig. 6**f**) given the noise sources are quantitatively described. Together this demonstrates the potential of a full Bayesian treatment of the state estimation, parameter estimation and model selection. This approach maximises the amount of information infered from stochastic time-courses. While developed for PC/cPCF data our approach is applicable to all time courses where the intrinsic noise of the studied system is governed by a first order kinetic scheme and the measuring apparatus can be quantitatively described.

Materials

The state evolution $s(t)$ of each individual channel in the patch was sampled with the Gillespie algorithm. ***Gillespie Daniel T.*** (***1977***) Then, 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.

Methods

In the Methods section we derive the equations for our Bayesian network for time series analysis of ion channels which are applicable for all linear chemical reaction networks (kinetic schemes). A detailed description of the experimental noise is provided in Appendix.

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

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) = \frac{1}{Z(\mathcal{Y}_T)} \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) \prod_j^{N_{\text{par}}} \mathbb{P}(\theta_j) \tag{18}$$

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 $\boldsymbol{\theta}$ (see Methods). The normalization constant

$$Z(\mathcal{Y}_T) = \int \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) \mathbb{P}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{19}$$

ensures that the posterior is a normalized distribution. The KF is a special class of models in the family of Bayesian networks ***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. 18 by the expression of Eq. 8.

The Bayesian framework (as demonstrated in this article) has various properties which makes it superior to maximum likelihood 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. 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-Roa 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 is not normally distributed. 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 maximum likelihood 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:

$$\mathbb{P}(\text{path}) = \mathbb{P}(\mathbf{s}(0), \mathbf{s}(1), \ldots, \mathbf{s}(T)) = \mathbb{P}(\mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(1) \mid \mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(2) \mid \mathbf{s}(1)) \cdots \mathbb{P}(\mathbf{s}(T) \mid \mathbf{s}(T-1)). \tag{20}$$

Markov models (MMs) and rate models are widely used for modeling molecular kinetics (Appendix Sec. 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.*** (***2013a***). 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:

$$\mathbf{T}_{\alpha,\beta} := \mathbb{P}(\mathbf{s}(t+1) = \mathbf{e}_\alpha \mid \mathbf{s}(t) = \mathbf{e}_\beta) = \exp(\mathbf{K} \cdot \Delta t)_{\alpha,\beta}, \tag{21}$$

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, is it 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 $\mathbf{s}$.

$$\mathbf{p}(t) = \mathbb{E}(\mathbf{s}(t)) \tag{22}$$

The probability vector obeys the discrete-time Master equation

$$\mathbf{p}(t+1) = \mathbf{T}\mathbf{p}(t)\mathbb{E}(\mathbf{s}(t+1)) = \mathbf{T}\mathbb{E}(\mathbf{s}(t)) \tag{23}$$

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 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:

$$\mathbf{n}(t) = \sum_{i=1}^{N_{\mathrm{ch}}} \mathbf{s}_i(t) \tag{24}$$

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. 24 in the channel dimension and from the single-channel CME Eq. 23 one can immediately derive the equation for the time evolution of the mean occupancy $\bar{\mathbf{n}}(t) = \mathbb{E}[\mathbf{n}(t)]$:

$$\bar{n}_\alpha(t+1) = \sum_\beta T_{\alpha,\beta}\bar{n}_\beta(t) \tag{25}$$

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
$\mathcal{Y}_T$	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
$\mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}\|\mathcal{Y})$	predictive distribution of the new data points $\tilde{\mathcal{Y}}$
$\mathbb{O}(\mathbf{y}\|\mathbf{n})$	distribution of observables for a single time step
$\mathcal{N}(\cdot)$	normal distribution
$\mathbb{E}[\cdot]$	mean value

Table 1. Important symbols

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

$$\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha) = \mathbb{P}(n_1, \dots, n_M \mid \mathbf{n}(t) = n_\alpha \mathbf{e}_\alpha) = \frac{n_\alpha!}{n_1! \cdots n_M!} T_{1,\alpha}^{n_1} \cdots T_{M,\alpha}^{n_M} \tag{26}$$

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.

$$\mathbf{n}(t+1) = \sum_\alpha \mathbf{n}^{(\alpha)}(t+1) \tag{27}$$

Evidently, the total number of channels is conserved during propagation. The distribution of $\mathbf{n}(t+1)$, defined by Eqs. 26 and 27, is called the *generalized multinomial distribution*:

$$\mathbf{n}(t+1) \sim \mathrm{general{-}multinomial}(\mathbf{n}(t), \mathbf{T}) \tag{28}$$

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 $\mathrm{general{-}multinomial}(\mathbf{n}(t), \mathbf{T})$ 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 (26) has the following mean and covariance matrix

$$\bar{\mathbf{n}}^{(\alpha)}(t+1) = n_\alpha \mathbf{T}_{:,\alpha} \tag{29}$$

$$\mathbf{\Sigma}^{(\alpha)}(t+1) = n_\alpha \mathrm{diag}(\mathbf{T}_{:,\alpha}) - n_\alpha \mathbf{T}_{:,\alpha,:} \mathbf{T}_{:,\alpha}^{\top} \tag{30}$$

where $\mathbf{T}_{:,\alpha}$ denotes the column number α of the transition matrix and $\mathrm{diag}(\mathbf{T}_{:,\alpha})$ describes the diagonal matrix with $\mathbf{T}_{:,\alpha}$ on its diagonal. Combining Eq. 27 with Eqs. 29 and 30 we deduce the mean and variance of the generalized multinomial distribution:

$$\mathbb{E}\left[\mathbf{n}(t+1) \mid \mathbf{n}(t)\right] = \sum_\alpha n_\alpha(t) \mathbf{T}_{:,\alpha} = \mathbf{T}\mathbf{n}(t) \tag{31}$$

$$\mathrm{cov}\left[\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right] = \sum_\alpha n_\alpha(t) \left(\mathrm{diag}(\mathbf{T}_{:,\alpha}) - \mathbf{T}_{:,\alpha} \mathbf{T}_{:,\alpha}^{\top}\right) = \mathrm{diag}\left(\mathbf{T}\mathbf{n}(t)\right) - \mathbf{T}\,\mathrm{diag}\left(\mathbf{n}(t)\right) \mathbf{T}^{\top} \tag{32}$$

Note that Eqs. 31 and 32 are conditional expectations that depend on the random state $\mathbf{n}$ at the previous time t and not only on the previous mean $\bar{\mathbf{n}}$. To find the absolute mean, the law of total expectation is applied to Eq. 31, giving

$$\bar{\mathbf{n}}(t+1) = \mathbb{E}\left[\mathbb{E}\left[\right]\mathbf{n}(t+1) \mid \mathbf{n}(t)\right)] = \mathbf{T}\bar{\mathbf{n}}(t), \tag{33}$$

in agreement with the simple derivation of Eq. 25. 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. 32 and 31, giving

$$\begin{aligned}
\mathbf{P}(t+1) &= \mathbb{E}\left[\mathrm{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] + \mathrm{cov}\left[\mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t)), \mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t))\right] && \text{(34a)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathrm{cov}(\mathbf{T}\mathbf{n}(t), \mathbf{T}\mathbf{n}(t)) && \text{(34b)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathbf{T}\mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))\mathbf{T}^{\top} && \text{(34c)}\\
&= \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} + \mathbf{T}\mathbf{P}(t)\mathbf{T}^{\top} && \text{(34d)}
\end{aligned}$$

where we have introduced the shorthand $\mathbf{P}(t) = \mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))$ in the last line. Eqs. 33, 34d 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 of the KF ***Moffatt*** (***2007***); ***Anderson and Moore*** (***2012***)(see Appendix 4). 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. 33, 34d. Comparing Eq. 34d with the KF prediction equation ***Anderson and Moore*** (***2012***) for $\mathbf{P}(t)$ we obtain the state-dependent covariance matrix of Eq. 2 as

$$\mathbf{Q}(\mathbf{T}, \overline{\mathbf{n}}(t)) = \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} \tag{35}$$

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 3) 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.

$$\mathbf{y}(t) = \begin{pmatrix} \text{fluorescence intensity}(t) \\ \text{ion current}(t) \end{pmatrix} = \begin{pmatrix} y_{\mathrm{flu}}(t) \\ y_{\mathrm{curr}}(t) \end{pmatrix} \tag{36}$$

As outlined in the introduction part, in Eq. 3 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. 34d). The observation matrix $\mathbf{H} \in \mathbb{R}^{N_{\mathrm{obs}} \times M}$ projects the hidden state vector $\mathbf{n}(t)$ onto $\mathbf{H}\mathbf{n}(t) \in \mathbb{R}^{N_{\mathrm{obs}}}$, the observation space. The observation distribution is

$$\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t)) = \mathcal{N}\left(\mathbf{y}(t)|\mathbf{H}\mathbf{n}(t), \mathbf{\Sigma}(\mathbf{n}(t))\right) \Leftrightarrow \mathbf{y}(t) = \mathbf{H}\mathbf{n}(t) + \boldsymbol{\nu}(t). \tag{37}$$

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 scaled Skellam distribution of the noise.

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

$$\mathbf{H} = \mathbf{H}_{\text{I}} + \mathbf{H}_{\text{binding}} \tag{38}$$

$$\mathbf{\Sigma}(t) = \mathbf{\Sigma}_{\text{open}}(t) + \mathbf{\Sigma}_{\text{meas.}} + \mathbf{\Sigma}_{\text{binding}}(t) + \mathbf{\Sigma}_{\text{back}} \tag{39}$$

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 3).

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 $\mathbb{E}[\nu_I(t)] = 0$ and variance $\mathbb{E}[\nu_I^2(t)] = \sigma_{\text{op}}^2 n_4(t) + \sigma_{\text{m}}^2$. By making the open-channel noise dependent on the hidden state population $n_4(t)$, we fully take advantage of the flexibility of Bayesian networks 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

$$\mathbf{H}_{\text{I}} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \end{pmatrix} \tag{40}$$

$$\mathbf{\Sigma}_{\text{open}}(t) + \mathbf{\Sigma}_{\text{meas.}} := \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{\text{op}}^2 n_4(t) + \sigma_{\text{m}}^2 \end{pmatrix} \tag{41}$$

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. 34d. By using the law of total variance the signal variance follows as:

$$\begin{aligned} \text{var}(\mathbf{y}(t)) &= \mathbb{E}[\text{var}[\mathbf{y}(t)|\mathbf{n}(t)]] + \text{var}[\mathbb{E}[\mathbf{y}(t)|\mathbf{n}(t)]] && \text{(42a)} \\ &= \mathbb{E}[\sigma_{\text{op}}^2 n_4(t) + \sigma_{\text{m}}^2] + \text{var}[\mathbf{H}_{\text{I}}\mathbf{n}(t)] && \text{(42b)} \\ &= \sigma_{\text{op}}^2 \mathbb{E}[n_4(t)] + \sigma_{\text{m}}^2 + (\mathbf{H}_{\text{I}}\mathbf{P}(t)\mathbf{H}_{\text{I}}^{\top})_{2,2} && \text{(42c)} \end{aligned}$$

See, Appendix Sec. 4.1 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}

$$y_{\mathrm{Fl}}(t) \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}(t)). \tag{43}$$

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. 37), we invoke the central limit theorem to approximate

$$y_{\mathrm{Fl}} \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}) \approx \mathcal{N}(\lambda_{\mathrm{Fl}}(t), \lambda_{\mathrm{Fl}}(t)) \tag{44}$$

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

$$\mathbf{H}_{\mathrm{binding}} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{45}$$

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. 44. We find

$$\mathbf{\Sigma}_{\mathrm{binding}}(t) := \begin{pmatrix} (\mathbf{Hn}(t))_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{46}$$

Under these assumptions the observation matrix can be written as follows

$$\mathbf{H} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & i \end{pmatrix} \tag{47}$$

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

$$\mathbf{y}(t) = \mathbf{Hn}(t) + \boldsymbol{\nu}_{\mathrm{m}}(t) + \begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix} \Leftrightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{Hn}(t), \mathbf{\Sigma}(t)). \tag{48}$$

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

$$\mathbb{E}\left[\begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}\right] = 0 \tag{49}$$

and

$$\mathrm{cov}\left(\begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}, \begin{pmatrix} \nu_{\mathrm{pois}}(t) \\ \nu_{\mathrm{op}}(t) \end{pmatrix}\right) = \mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{binding}}(t) \tag{50}$$

. 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

$$\mathrm{var}[y_{\mathrm{fluo}}(t)] = \mathbb{E}[\mathrm{var}(y(t)|\mathbf{n}(t))] + \mathrm{var}[\mathbb{E}(y(t)|\mathbf{n}(t)] \tag{51a}$$

$$= \mathbb{E}[\sigma^2_{\mathrm{back}} + (\mathbf{Hn}(t))_1] + \mathrm{var}(\mathbf{Hn}(t)) \tag{51b}$$

$$= \sigma^2_{\mathrm{back}} + (\mathbf{Hn}(t))_1 + (\mathbf{HP}(t))\mathbf{H}^{\mathsf{T}})_{1,1} \tag{51c}$$

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

$$\begin{aligned}\mathrm{cov}(y_{\mathrm{patch}}, y_{\mathrm{fluo}}) &= \mathbb{E}[\mathrm{cov}(y_{\mathrm{patch}}, y_{\mathrm{fluo}}|\mathbf{n})] + \mathrm{cov}(\mathbb{E}(y_{\mathrm{patch}}|\mathbf{n}), \mathbb{E}(y_{\mathrm{fluo}}|\mathbf{n})) && \text{(52a)}\\ &= 0 + \mathrm{cov}(\mathbf{H}_{2,:}\mathbf{n}, \mathbf{H}_{1,:}\mathbf{n}) && \text{(52b)}\\ &= \mathbf{H}_{2,:}\,\mathrm{cov}(\mathbf{n},\mathbf{n})\mathbf{H}_{1,:}^{\top} = \mathbf{H}_{2,:}\mathbf{P}(t)\mathbf{H}_{1,:}^{\top} && \text{(52c)}\end{aligned}$$

yielding the matrix

$$\mathrm{cov}(\mathbf{y},\mathbf{y}) = \mathbf{H}\mathbf{P}(t)\mathbf{H}^{\top} + \mathbf{\Sigma}(t) \tag{53}$$

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. 31) and for the covariance matrix (Eq. 34d) 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***).

Algorithm.		Prediction	Correction	
	Mean	Covariance	Mean	Covariance
RE	Yes	No	No	No
KF	Yes	Yes	Yes	Yes

Table 3. Comparison of algorithms: The RE approach predicts the next mean ensemble state, estimates probabilities of occupying a certain state by $\mathbf{p} \approx \frac{1}{N_{\mathrm{ch}}}\mathbb{E}[\mathbf{n}(t+1)]$ and constructs a likelihood by a multinomial assumption ***Milescu et al.*** (***2005***). The multinomial distribution is then approximated by a normal distribution and the variance from the experimental noise is added. There is neither a prediction of $\mathbf{P}(t)$ nor any correction step, thus the random fluctuations and the hidden structure of an ion channel ensemble of finite size is ignored. In contrast, the KF accounts correctly for all aspects of the hidden stochastic dynamics of the ion channels as long as all involved distributions can be approximated by multivariate normal distributions. This is a much less restrictive assumption then assuming that the ensemble is fully determined just by its mean value. Additionally, the KF includes the information from the data in each state estimation in an optimal manner.

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

$$\mathbf{n}(t)_{\mathrm{posterior}} = +\mathbf{n}(t)_{\mathrm{prior}} + \mathbf{K}\left(\mathbf{y}(t) - \mathbf{H}\mathbf{n}(t)_{\mathrm{prior}}\right) \tag{54}$$

Where Kalman gain matrix $K := \mathbf{P}(t)_{\mathrm{prior}}\mathbf{H}^{\top}\mathbf{\Sigma}^{-1}$ 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

$$\mathbf{P}(t)_{\mathrm{posterior}} = \mathbf{P}(t)_{\mathrm{prior}} - \mathbf{K}\left(\mathbf{H}\mathbf{P}(t)_{\mathrm{prior}}\mathbf{H} + \mathbf{\Sigma}(t)\right)\mathbf{K}^{\top} \tag{55}$$

Appendix Eq. 28 and 29 form with Methods Eq. 26 and 30 the filtering equations which summarize the algorithm. One initialises the first $\mathbf{n}(0)$ and $\mathbf{P}(0)$ and with an equilibrium assumption.

Bayesian Model selection via predictive accuracy

The minimum of the Kullback-Leibler divergence can be found asymptotically by maximizing the predictive accuracy ***Burnham and Anderson*** (***2004***); ***Gelman et al.*** (***2014***) The predictive accuracy is defined as

$$\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T) := \log \mathbb{E}_{\theta,\text{post}}[\mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})] = \log \int \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})\mathbb{P}_{\text{post}}(\boldsymbol{\theta}|\mathcal{Y}_T)\,\mathrm{d}\boldsymbol{\theta} \tag{56}$$

for a specific new (held-out) data set $\tilde{\mathcal{Y}}_T$, which has not been used for training the model ***Gelman et al.*** (***2014***). $\mathbb{E}_{\text{post}}[\cdot]$ denotes the average of some function found by integrating it over the posterior distribution $\mathbb{P}_{\text{post}}(\boldsymbol{\theta}|\mathcal{Y}_T)$. The difference between maximum likelihood (or maximum *a posteriori* cross-validation, MPC) to Bayesian cross-validation in Eq. 56 is that the mentioned point estimates in MPC would yield a Dirac distribution $\mathbb{P}_{\text{post}}(\boldsymbol{\theta}|\mathcal{Y}_T) = \delta(\boldsymbol{\theta}_{\text{MLE}} - \boldsymbol{\theta})$ as the posterior. Doing so collapses the integral in Eq. 56 to $\log \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta}_{\text{MLE}})$. Note, that Eq. 56 can also be used to selected the prior distribution for the parameters if the average is taken with respect to the prior distribution instead of to the posterior distribution. In our application to ion channel dynamics, we generate from each patch that was used for the training data at least a second hold-out time trace to validate the model. Here we explicitly use the term *model* in a way that includes the observation model and is not restricted to the kinetic scheme. A model of all unknowns considered to be relevant for the data. Moreover, assuming the availability of multiple time traces from the same patch allows to avoid difficulties of applying cross-validation within one time series. Since we wish to know the predictive performance of the model for all possible unseen data sets, we have also to average over the unknown true data generating process $\mathbb{F}(\tilde{\mathbf{y}}(t))$. The objective of the experiment is then to sample sequences of data $\mathcal{Y}_T$ which are as representative as possible for $\mathbb{F}$. The expected predictive accuracy for a full unseen set of time series of data is then

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] = \int \cdots \int \log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)\mathbb{F}(\tilde{\mathcal{Y}}_T) \prod_i^T \mathrm{d}\tilde{\mathbf{y}}_i \tag{57}$$

Unfortunately, $\mathbb{F}(\tilde{\mathcal{Y}}_T)$ is unknown. In practice the expectation value in Eq. 57 is approximated summing over independent realizations generated by the experiment.

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] \approx \sum_{k=1}^{K}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T^{(k)}|\mathcal{Y}_T)] \tag{58}$$

If no independent realizations are available e.g. if experiments are expensive, the estimation can be performed over the training data instead, i.e. setting $\tilde{\mathcal{Y}}_T = \mathcal{Y}_T$. But this leads to an over-optimistic biased predictive accuracy estimate***Gelman et al.*** (***2014***). To compensate for that optimistic bias one needs to penalize the model complexity***Gelman et al.*** (***2014***) which is done by scores called information criteria. The first non-Bayesian information criterion was found by Akaike ***Akaike*** (***1998***). It states that for linear models with Gaussian noise, with no hidden structures, asymptotically one can compensate the too optimistic bias from the training sample by subtracting the number of parameters $\dim(\boldsymbol{\theta})$ of the model

$$\text{AIC} = \log\, \mathbb{P}(\mathcal{Y}_T|\hat{\boldsymbol{\theta}}_{\text{mle}}) - \dim(\boldsymbol{\theta}). \tag{59}$$

Under those conditions AIC is asymptotically equivalent to cross-validation ***Stone*** (***1977***). These very restrictive model conditions are not satisfied by models with hidden or latent structures ***Watanabe*** (***2007***). Thus AIC has no mathematical justification for any Biophysical data model, whose experimental base is a signal which probes some hidden dynamics. We show (Fig. 7 b), that AIC fails to predict the true data-generating process unless the data are strong enough to create a multivariate normal posterior. Recently, Watanabe***Watanabe*** (***2010***) showed for muc broader class of models, including singular models, that asymptotically WAIC is equivalent for large data sets to Bayesian cross-validation Thus the predictive accuracy of the model Eq. 57 can be estimated by

$$\text{WAIC} := \log\, \mathbb{P}_{\text{post}}(\mathcal{Y}_T) - p_{\text{WAIC}} \tag{60}$$

Algori.		Selec.	Strategy			
	AIC	WAIC	MPC	BC	CMEXP	$\mathbb{E}[\mathbf{r}^2]$
$\mathrm{RE}_{\mathrm{PC}}$	No	No	Yes	Yes	No	No
$\mathrm{KF}_{\mathrm{PC}}$	No	Yes	Yes	Yes	Yes	Yes
$\mathrm{KF}_{\mathrm{cPCF}}$	No(Yes)	Yes	Yes	Yes	Yes	Yes

Table 4. Model selection strategies by estimating the predictive accuracy (columns 1-4), by continuous model expansion (CMEXP) (column 5) and by residua r^2 (column 6). The No(Yes) means that for the used example we were successful but for AIC there is no asymptotic guarantee that it converges with large N_{data} to the true value.

The bias correction $p_{\mathrm{WAIC}} = \mathrm{var}_{\mathrm{post}}(\log \mathbb{P}(\mathcal{Y}|\boldsymbol{\theta}))$ is asymptotically correct even for singular models and reduces to the from Akaike expected $\dim(\boldsymbol{\theta})$ for regular normal models ***Gelman et al.*** (***2014***). We show that in order to reliably detect overfitting and determine the best generalizing model on the training data, it is inevitable to use the KF instead of REs, see Fig. 7 a, b. Notably, WAIC exploits the full posterior thus model selection for singular model should usually be done within the Bayesian framework. The obvious way to estimate Eq. 57 by Eq. 56 used through out this study as a reference is with hold out data. In order to decide upon a parsimonious model, predictive accuracy methods should be combined with the continuous model expansion technique to interpret the cross-validation and information criteria correctly.

Acknowledgments

The authors are grateful to Dr. E. Schulz and Dr. T. Eick for designing a software to simulate channel activity in ensemble patches and computing time traces, respectively. 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

Akaike H. A New Look at the Statistical Model Identification. New York, NY: Springer New York; 1998.

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; .

741 **Biskup C**, Kusch J, Schulz E, Nache V, Schwede F, Lehmann F, Hagen V, Benndorf K. Relating ligand binding
742 to activation gating in CNGA2 channels. Nature. 2007 feb; 446:440. doi: 10.1038/nature05596; 10.1038/na-
743 ture05596.

744 **Bronson JE**, Fei J, Hofman JM, Gonzalez Jr RL, Wiggins CH. Learning rates and states from biophysical time series:
745 a Bayesian approach to model selection and single-molecule FRET data. Biophys J. 2009; 97(12):3196–3205.

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

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

751 **Burnham KP**, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociological
752 methods & research. 2004; 33(2):261–304.

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

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

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

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

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

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

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

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

769 **Deuflhard P**, Weber M. Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl. 2005;
770 398(Supplement C):161–184. doi: 10.1016/j.laa.2004.10.026, special Issue on Matrices and Mathematical
771 Biology.

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

774 **Fearnhead P**, Giagos V, Sherlock C. Inference for reaction networks using the linear noise approximation.
775 Biometrics. 2014; 70(2):457–466.

776 **Finkenstädt B**, Woodcock DJ, Komorowski M, Harper CV, Davis JR, White MR, Rand DA, et al. Quantifying
777 intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to
778 single cell data. Ann Appl Stat. 2013; 7(4):1960–1982.

779 **Folia MM**, Rattray M. Trajectory inference and parameter estimation in stochastic models with temporally
780 aggregated data. Statistics and Computing. 2018 Sep; 28(5):1053–1072. doi: 10.1007/s11222-017-9779-x.

781 **Frauenfelder H**, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science. 1991;
782 254(5038):1598–1603. doi: 10.1126/science.1749933.

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

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

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

789 **Gelman A**, Simpson D, Betancourt M. The prior can often only be understood in the context of the likelihood.
790 Entropy. 2017; 19(10):555.

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

793 **Gillespie CS**, Golightly A. Bayesian inference for the chemical master equation using approximate models. In:
794 *Ninth International Workshop on Computational Systems Biology, WCSB 2012, June 4-6, Ulm, Germany*, vol. 4;
795 2012. p. 23.

796 **Gillespie Daniel T** . Exact stochastic simulation of coupled chemical reactions. The J of Phys Chem. 1977;
797 81(25):2340–2361.

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

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

802 **de Gunst MM**, Künsch H, Schouten J. Statistical analysis of ion channel data using hidden Markov models
803 with correlated state-dependent noise and filtering. Journal of the American Statistical Association. 2001;
804 96(455):805–815.

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

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

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

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

812 **van Holde K**. A hypothesis concerning diffusion-limited protein–ligand interactions. Biophys Chem. 2002;
813 101:249–254.

814 **Horn R**, Lange K. Estimating kinetic constants from single channel data. Biophys J. 1983; 43(2):207–223. doi:
815 10.1016/S0006-3495(83)84341-0.

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

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

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

822 **Jaynes ET**. Information theory and statistical mechanics. Phys rev. 1957; 106(4):620.

823 **Kalman RE**. A new approach to linear filtering and prediction problems. J of basic Engineering. 1960; 82(1):35–
824 45.

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

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

829 **Kienker P**. Equivalence of aggregated Markov models of ion-channel gating. P of the Roy Soc of London B
830 Biological Sciences. 1989; 236(1284):269–309.

831 **Komorowski M**, Finkenstädt B, Harper CV, Rand DA. Bayesian inference of biochemical kinetic parameters
832 using the linear noise approximation. BMC Bioinformatics. 2009; 10:343–343.

833 **Kullback S**, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951; 22(1):79–
834 86.

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

837 **Kusch J**, Biskup C, Thon S, Schulz E, Nache V, Zimmer T, Schwede F, Benndorf K. Interdependence of Re-
838 ceptor Activation and Ligand Binding in HCN2 Pacemaker Channels. Neuron. 2010; 67(1):75–85. doi:
839 10.1016/j.neuron.2010.05.022.

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

843 **McElreath R**. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC;
844 2018.

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

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

849 **Moffatt L**. Estimation of Ion Channel Kinetics from Fluctuations of Macroscopic Currents. Biophys J. 2007;
850 93(1):74–91. doi: 10.1529/biophysj.106.101212.

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

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

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

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

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

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

864 **Piironen J**, Vehtari A. Comparison of Bayesian predictive methods for model selection. Statistics and Comput-
865 ing. 2017; 27(3):711–735.

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

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

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

872 **Rosales RA**. MCMC for hidden Markov models incorporating aggregation of states and filtering. Bull Math Biol.
873 2004 Sep; 66(5):1173–1199. doi: 10.1016/j.bulm.2003.12.001.

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

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

877 **Siekmann I**, Sneyd J, Crampin EJ. MCMC Can Detect Nonidentifiable Models. Biophys J. 2012; 103(11):2275–
878 2286. doi: 10.1016/j.bpj.2012.10.024.

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

881 **Smoluchowski Mv**. Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen.
882 Zeitschrift für physikalische Chemie. 1918; 92(1):129–168.

883 **Stepanyuk A**, Borisyuk A, Belan P. Maximum likelihood estimation of biophysical parameters of synaptic re-
884 ceptors from macroscopic currents. Frontiers in Cellular Neuroscience. 2014; 8:303.

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

887 **Stone M**. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. J of the
888 Roy Statistical Soc: Series B (Methodological). 1977; 39(1):44–47. doi: 10.1111/j.2517-6161.1977.tb01603.x.

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

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

893 **van der Vaart A**. Bayes procedures. In: *Asymptotic Statistics* Cambridge University Press; 1998.p. 138–152.

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

895 **Varga RS**, et al. Minimal Gerschgorin sets. Pacific Journal of Mathematics. 1965; 15(2):719–729.

896 **Vehtari A**, Ojanen J. A survey of Bayesian predictive methods for model assessment, selection and comparison.
897 Statist Surv. 2012; 6:142–228. doi: 10.1214/12-SS102.

898 **Vehtari A**, Ojanen J, et al. A survey of Bayesian predictive methods for model assessment, selection and com-
899 parison. Statistics Surveys. 2012; 6:142–228.

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

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

904 **Wallace CS**. Statistical and Inductive Inference by Minimum Message Length. Information Science and Statistics,
905 Berlin, Germany: Springer-Verlag; 2005.

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

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

910 **Watanabe S**. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in
911 singular learning theory. J of machine learning research. 2010; .

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

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

915 **Wulf M**, Pless SA. High-Sensitivity fluorometry to resolve ion channel conformational dynamics. Cell reports.
916 2018; 22(6):1615–1626.

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

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

Appendix 1

An example emphasizing the importance of weakly informative priors for complex models

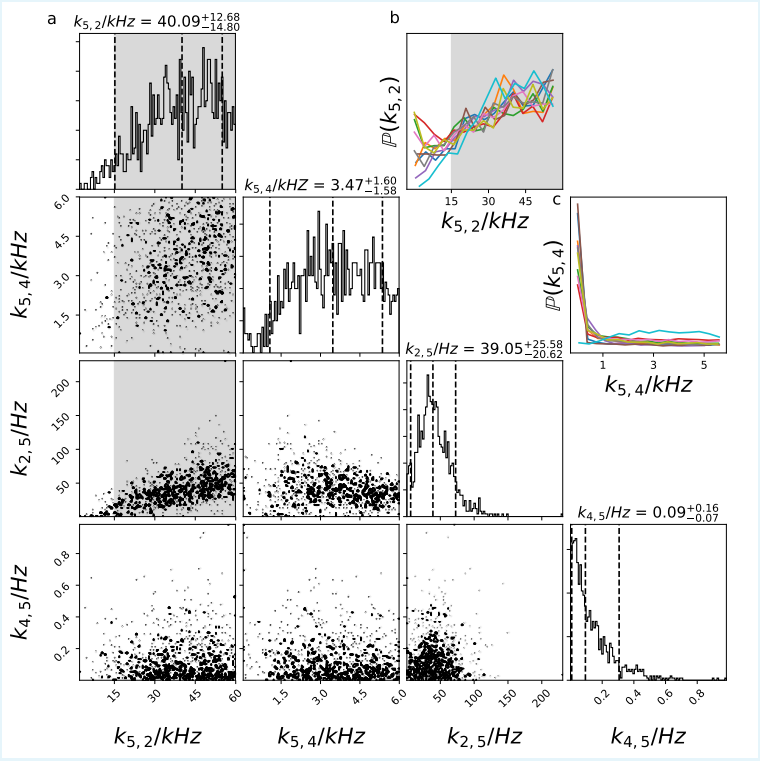


Figure 1. a, Posterior of the rates which are not present in the true process equipped with a uniform prior on $k_{5,2}$. The uniform prior puts too much probability mass into regions (gray shaded) where one should distrust the likelihood value due to limitations of the experiment. The likelihood cannot confine $k_{5,2}$ by the data such that the only limitation is the sampling box. **b,** This lack of information in the data is true for 10 out of 10 data sets though some of the data sets show a local maximum for $k_{5,2} \to 0$. **c,** For $k_{5,4}$ 9 out of 10 data sets show a global maximum at $k_{5,4} = 0$, only the showcased data set has significant probability mass for $k_{5,4} > 0$. Notice, that the first and second columns and rows are displayed in kHz while the others are in Hz.

We demonstrate the influence of prior information for the model selection by continuous model expansion (Fig. 6 **i-j**). As argued in the main article, weakly informative prior distributions can support the algorithm to select simpler base models, which means concentrating the posterior for certain $k_{i,j}$ around zero ***Gelman et al.*** (***2017***). Here we show the effect of the halfnormal prior $\mathbb{P}(k_{5,2})$ for the other rates and eigenvalues. We compare now (Appendix Fig. 1) posteriors of the rates of in model M_5 which are not present in the true process. First, we use a prior $k_{5,2} \sim \text{uniform}(0, 60\text{kHz})$ which places too much probability mass to high frequencies, if we consider that the frequency by which the KF analysed the data ranged from 83.3 Hz to 500 Hz. At a first glance one might see a finite peak in $\mathbb{P}(k_{2,5})$, implying that there is a second open state which opens at a similar rate as the true opening step. Looking at $\mathbb{P}(k_{2,5}, k_{5,2})$ one realizes that the posterior has a ridge which appears in the marginal distribution $\mathbb{P}(k_{2,5})$ as a peak. Most of that ridge lies in regions $k_{5,2} > 10\,\text{kHz}$ where we should distrust the experimental data due to the noise and limited time resolution to reasonably constrain the model. Note that the used prior places 6 times more probability mass higher then the sampling frequency, which in case of weakly informative data is a strong statement for mag-

nitude of $k_{5,2}$ being greater than would could have been measured. As argued in the main text that would not be an issue if the data would be informative on $k_{5,2}$. Indeed, $\mathbb{P}(k_{2,5}, k_{5,2})$ is bounded by the sampling box rather than by the data.

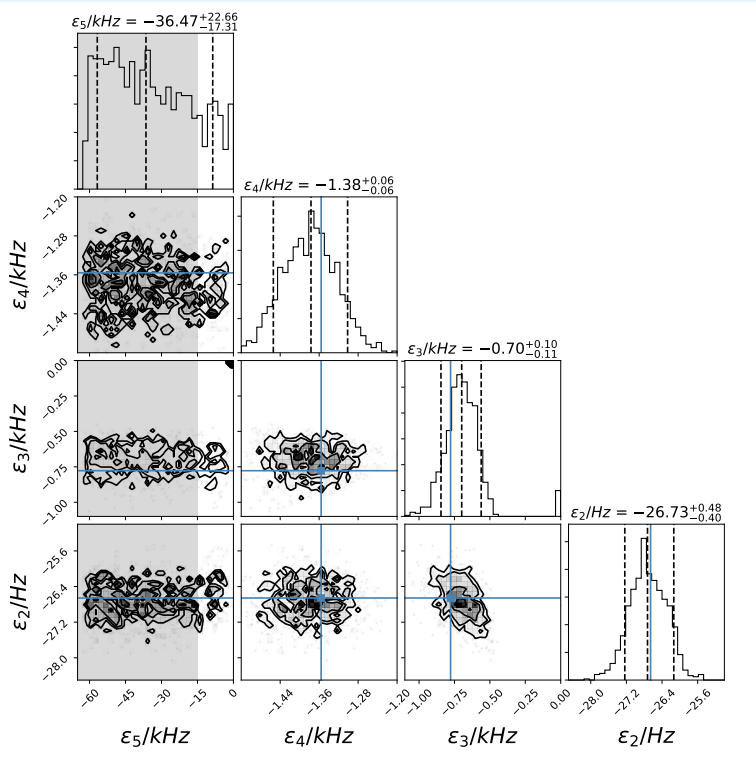


Figure 2. Posterior of the eigenvalues derived from the posterior of the rate matrix. We left out the equilibrium eigenvalue $\epsilon_1 = 0$, since by construction of the rate matrix it is always zero. The vertical dashed lines on the diagonal show the quantiles $\{0.1, 0.5, 0.9\}$. The posterior of the three slower eigenvalues cover the eigenvalues of the true process. Examining ϵ_5 reveals that roughly 90% of the probability mass belongs to eigenvalues faster than the sampling frequency. The upper limit of ϵ_5 is only confined by the sampling box but not by the data.

Closely related to the rate matrix is its spectrum of eigenvalues ϵ. A kinetic scheme consisting of M states has M eigenvalues. The largest one is always zero ***Colquhoun and Hawkes*** (***(1995***). This eigenvalue corresponds to the equilibrium solution of the chemical network ***Colquhoun and Hawkes*** (***(1995***). All other $M - 1$ eigenvalues are the negative inverse of the timescales on which deviations from the equilibrium distributions decay in time ***Colquhoun and Hawkes*** (***(1995***). The Gershgorin-circle theorem justifies this heuristic ***Varga et al.*** (***1965***). It states, that the spectrum of a matrix is inside the union of Gershgorin-circles. For a matrix with real valued eigenvalues such as the rate matrix this statement simplifies to real valued intervals R. For the j-th column the Gershgorin-interval R_j is

$$R_j = [k_{j,j} - \sum_{i,i\neq j}^{M} k_{j,i}, k_{j,j} + \sum_{i,i\neq j}^{M} k_{j,i}] \tag{61}$$

We use the properties $k_{i,j} > 0$, for $i \neq j$ and $k_{j,j} = -\sum_{i,i\neq j}^{M} k_{j,i} < 0$ of a rate matrix and derive

$$R_j = [2k_{j,j}, 0] \tag{62}$$

Since the interval from the column with the smallest diagonal element $k_{min} < 0$ covers all other intervals the union is

$$\epsilon_i \in R = [2k_{\min}, 0] \tag{63}$$

Each eigenvalue is always larger than 2 times the smallest diagonal element but smaller then 0.

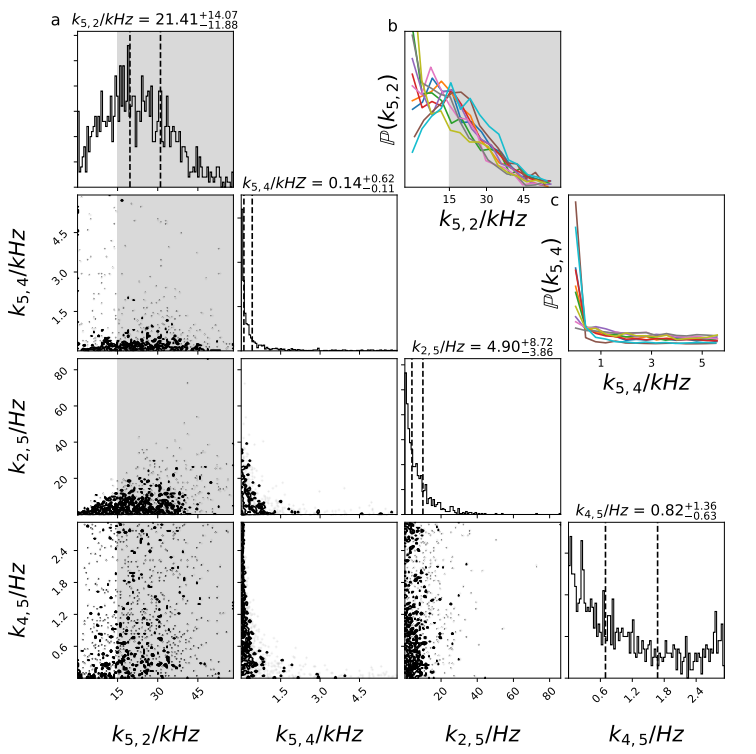


Figure 3. a, Posterior of the rates which are not present in the true process equipped with a weakly informative prior $k_{5,2} \sim \text{halfnormal}(0, 6\,\text{kHz})$. Still there is much probability mass in unrealistic rate regions (gray shaded) and one can still identify the correlated structure of $\mathbb{P}(k_{5,2}, k_{2,5})$ but the distribution, in particular the rates into the overfitting state O_5, show a distinct maximum for $k \to 0$. **b,** The posteriors for all data sets of $\mathbb{P}(k_{5,2})$ show now the tendency to develop a peak for $k_{5,2} \to 0$. **c,** All data sets indicate that also $k_{5,4}$ describes a process which is either slow beyond the time scales of the experiment or does not exist.

The unbounded rate $\mathbb{P}(k_{5,2})$ creates a posterior for the eigenvalues (Appendix Fig. 2) whose 90% probability mass for ϵ_5 covers areas faster than what could have been measured. The algorithm places most of the probability mass where it does not harm the fit. Since disturbances corresponding to that eigenvalue ***Colquhoun and Hawkes*** (***1995***) suffered already, a strong decay before the next data point is measured. This should raise the concern that the model would do better without the fifth eigenvalue thus one of the fife states should be left out. Not surprisingly, the likelihood benefits from increasing a rate which empties a state which does not exist in the true process. The posterior of the slower time scales $\epsilon_4 - \epsilon_2$ (Appendix Fig. 2) covers the true values. From Appendix Eq. 5 it is clear that there is close correspondence between ϵ_5 and $k_{5,2}$. A prior such as $k_{5,2} \sim \text{halfnormal}(0, 6\,\text{kHz})$ is still vague but values $k_{5,2} > 10\,\text{kHz}$ are strongly penalized. The prior states that we only except a rate close to the sampling rate if the data indicates it by a sharp likelihood peak which dominates the weakly regularising prior. The effect of that prior on the posterior of rates can be seen in (Appendix Fig. 3). $\mathbb{P}(k_{2,5}$. Develops a peak for $k_{2,5} = 0$ and is much more concentrated

close to zero. It further emphasizes that O_5 is a state which should be left out (Fig. 5 **i-j**). Additionally, with this weak constraint some of 10 data sets gain the tendency to develop a maximum at $\mathbb{P}(k_{5,2} = 0)$ (Appendix Fig. 3)**b**. As a side effect, the prior helps to sample from the posterior in limited time because it suppresses the correlations in the high-dimensional tails.

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.*** (***2013b***).

Appendix 3

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 network is that $\mathrm{var}[y_{\mathrm{fl}}] = \mathbb{E}[y_{\mathrm{fl}}]$

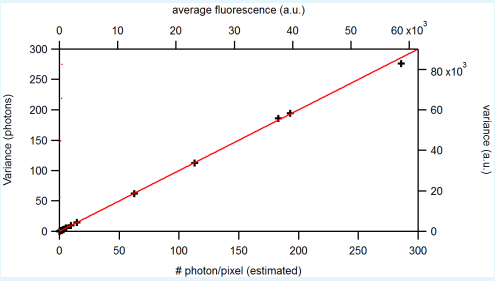
Here we show that this assumption for the detectors used in our system (Ch1 and Ch2, LSM 710, Carl Zeiss) under typical cPCF conditions is fulfilled by re-scaling to photon-numbers, The measured variance obeys $\mathrm{var}[y_{\mathrm{fl}}] = a \cdot \mathbb{E}[y_{\mathrm{fl}}]$. It depends linearly on the mean signals (Appendix Fig. 1a).

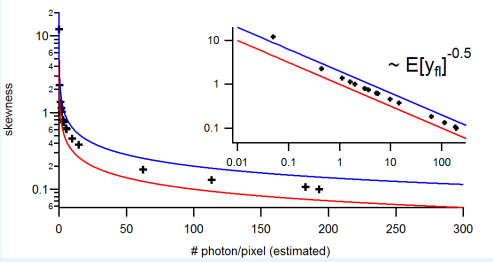
$$\mathrm{var}[y] = a\mathbb{E}[y] \tag{64}$$

$$\mathrm{var}[\mu x] = a\mathbb{E}[\mu x] \tag{65}$$

$$\mu^2 \,\mathrm{var}[x] = a\mu\,\mathrm{E}[x] \tag{66}$$

For the scaled signal x being Poisson distributed follows $\mu = a$. Then 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.





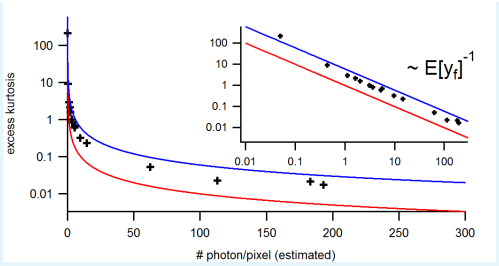


Figure 1. Benchmark of the signal statistics for experimental solution data recorded under cPCF conditions: The concentrations of the fluorescent ligand were 0.25, 3 and 15 μM and a reference dye was present. The laser intensities covered 1.6 orders of magnitude at constant detection settings. The data 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.

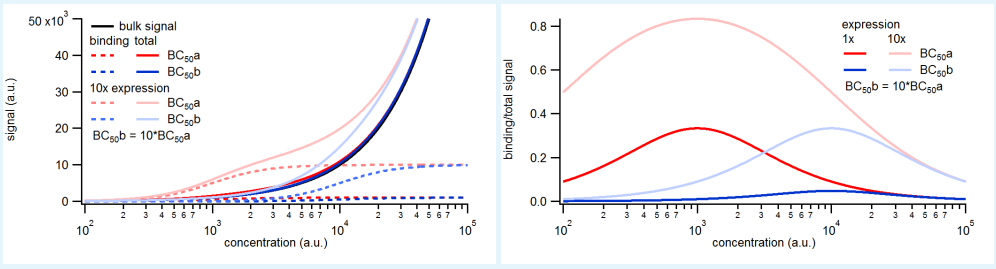


Figure 2. Simulated binding signals. **a,** Comparison of binding of a labeled ligand at two concentrations. A simple two-ligand binding process is simulated with the Hill equation for the two expression levels of 1,000 or 10,000 binding sites and a BC_{50} of 1,000 (BC_{50}a) or 10,000(BC_{50}b), respectively, 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

$$y_{\mathrm{fl,k}} = y_{\mathrm{lig,total}} - \hat{\lambda}_{\mathrm{lig,back}} - (y_{\mathrm{fl,ref}} - \hat{\lambda}_{\mathrm{ref,back}}) \frac{\hat{\lambda}_{\mathrm{lig}} - \hat{\lambda}_{\mathrm{lig,back}}}{\hat{\lambda}_{\mathrm{ref}} - \hat{\lambda}_{\mathrm{ref,back}}}, \tag{67}$$

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 were 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

$$\sigma_\zeta^2 = \lambda_{\mathrm{lig}} + \frac{\lambda_{\mathrm{lig}}^2}{\lambda_{\mathrm{ref}}} \tag{68}$$

is derived from simulated data in the Appendix 3. 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

$$\Sigma_{back} = \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{69}$$

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

$$y_{\mathrm{bulk}} = y_{\mathrm{lig,bulk}} - y_{\mathrm{ref,bulk}} \frac{\lambda_{\mathrm{lig,bulk}}}{\lambda_{\mathrm{ref,bulk}}} \tag{70}$$

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

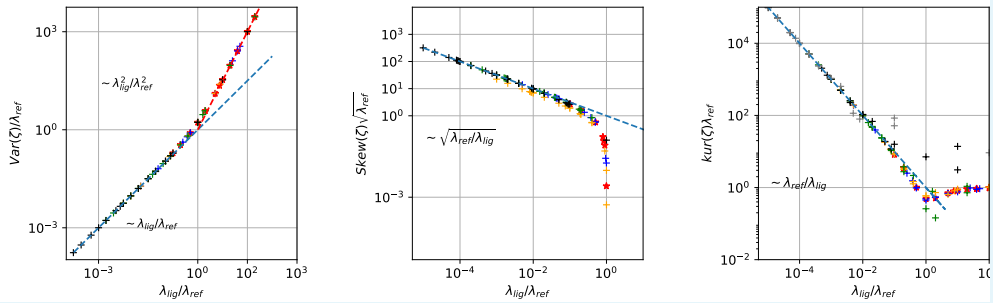


Figure 3. Master curves of 2nd till 4th centralized moment of photon counting noise ζ arising from the difference signal of fluorescent ligands and the dye in the bulk. The curves are created from $4 \cdot 10^5$ draws from Poisson distributions with different combinations of intensities for the reference 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 $\rho_{\text{ref}}\lambda_{\text{bulk}}$ 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. 70. The variance is derived empirically to be

$$\frac{\sigma_\zeta^2}{\lambda_{\text{ref}}} = \frac{\lambda_{\text{lig}}}{\lambda_{\text{ref}}} + \frac{\lambda_{\text{lig}}^2}{\lambda_{\text{ref}}^2}. \tag{71}$$

In Appendix Fig. S. 3 **a**, we confirm the intuition $\lambda_{\text{ref}} \to \infty \Rightarrow \text{var}(\zeta) = \lambda_{\text{lig}}$. 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

$$\text{skew}(\zeta)\sqrt{\lambda_{\text{ref}}} = \sqrt{\frac{\lambda_{\text{ref}}}{\lambda_{\text{lig}}}} \tag{72}$$

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

$$\text{skew}(\zeta)\sqrt{\lambda_{\text{ref}}} \leq \sqrt{\frac{\lambda_{\text{ref}}}{\lambda_{\text{lig}}}} \tag{73}$$

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

$$\text{kur}(\zeta)\lambda_{\text{ref}} \leq \frac{\lambda_{\text{ref}}}{\lambda_{\text{lig}}} \tag{74}$$

for $\lambda_{\text{ref}} \leq \lambda_{\text{lig}}$. 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 should be $\lambda_{\text{lig}}/\lambda_{\text{ref}}$.

Appendix 4

Output statistics of Bayesian networks

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

$$y(t) = \mathbf{H}\mathbf{n}(t) + \nu(t) \Leftrightarrow y \sim \mathbb{O}(y|\mathbf{n}) = \mathrm{normal}(\mathbf{H}\mathbf{n}(t), \sigma_{\mathrm{m}}^2) \tag{75}$$

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 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

$$\begin{aligned}
\mathrm{var}(y(t), y(t)) &= \mathbb{E}[(y(t) - \mathbb{E}[y(t)])(y(t) - \mathbb{E}[y(t)])^{\mathsf{T}}] && \text{(76a)}\\
&= \mathbb{E}[(y(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])(y(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])^{\mathsf{T}}] && \text{(76b)}\\
&= \mathbb{E}[(\mathbf{H}\mathbf{n}(t) + \nu(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])(\mathbf{H}\mathbf{n}(t) + \nu(t) - \mathbf{H}\mathbb{E}[\mathbf{n}(t)])^{\mathsf{T}}] && \text{(76c)}\\
&= \mathbf{H}\mathbb{E}[(\mathbf{n}(t) - \mathbb{E}[\mathbf{n}(t)])(\mathbf{n}(t) - \mathbb{E}[\mathbf{n}(t)]^T]\mathbf{H}^{\mathsf{T}} + \mathbb{E}[\nu(t)^2] && \text{(76d)}\\
&= \mathbf{H}\mathbf{P}(t)\mathbf{H}^{\mathsf{T}} + \sigma_{\mathrm{m}} && \text{(76e)}\\
& && \text{(76f)}
\end{aligned}$$

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. 76 is the classical KF variance prediction of a signal. The first term in Eq. 76, 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

Additional 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

$$y(t) = \mathbf{H}\mathbf{n}(t) + \nu_{\mathrm{m}}(t) + \nu_{\mathrm{op}}(t) \Leftrightarrow y \sim p(y|\mathbf{n}) = \mathrm{normal}(y|\mathbf{H}\mathbf{n}(t), \sigma_{\mathrm{m}}^2 + n_4(t)\sigma_{\mathrm{op}}^2) \tag{77}$$

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

$$\begin{aligned}
\mathbb{P}(y(t)) &= \int \mathrm{normal}(y|\mathbf{H}\mathbf{n}, \sigma_{\mathrm{m}}^2 + n_4\sigma_{\mathrm{op}}^2)\ \mathrm{normal}(\mathbf{n}|\bar{\mathbf{n}}(t), \mathbf{P}(t))\,\mathrm{d}n && \text{(78a)}\\
&= \frac{1}{const}\int \exp\left(\frac{(y - \mathbf{H}\mathbf{n})^2}{2(\sigma_{\mathrm{m}}^2 + n_4\sigma_{\mathrm{op}}^2)}\right)\exp\left(\frac{1}{2}(\mathbf{n} - \bar{\mathbf{n}}(t))\mathbf{P}^{-1}(\mathbf{n} - \bar{\mathbf{n}}(t))^{\mathsf{T}}\right)\mathrm{d}n && \text{(78b)}
\end{aligned}$$

When assuming that the relative fluctuations of $\mathbf{n}(t)$ are small on average 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, just with the difference that the measurement noise is the sum of two components.

$$y(t) \sim \text{normal}(\mathbf{H}\bar{\mathbf{n}}(t), \sigma_{\text{m}}^2 + \sigma_{\text{op}}^2 \bar{n}_4(t) + \mathbf{HPH}^\top) \tag{79}$$

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

$$\begin{aligned} \text{var}(y(t)) &= \mathbb{E}[\text{var}[y(t)|\mathbf{n}(t)]] + \text{var}[\mathbb{E}[y(t)|\mathbf{n}(t)]] && \text{(80a)} \\ &= \mathbb{E}[\Sigma + \sigma_{\text{op}}^2 n_4(t)] + \text{var}[\mathbf{Hn}(t)] && \text{(80b)} \\ &= \sigma_m^2 + \sigma_{\text{op}}^2 \mathbb{E}[n_4(t)] + \mathbf{HP}(t)\mathbf{H}^\top && \text{(80c)} \end{aligned}$$

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 about the current $\mathbf{n}(t)$. The information in the noise enters in two ways the likelihood of the data. 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 $\mathbf{K}$ matrix defined in the next section.