# Bayesian selection and analysis of Hidden Markov models of macroscopic multi-dimensional ion channel data

Jan Münch[1], Fabian Paul[2], Ralf Schmauder[1], and Klaus Benndorf[1]

[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.*

**Inferring the complex conformational dynamics of ion channels from ensemble currents is a daunting task due to insufficient information in the data. Model identification can be facilitated by including orthogonal fluorescence time courses. We present a parallelized Kalman filter for specifying Hidden Markov Models for patch-clamp fluorescence data. We demonstrate an enormous flexibility of this Bayes estimator by including the Poisson noise in fluorometry data and open-channel noise in current data. The accuracy of the parameter estimation is increased by tenfold if compared to fitting Rate Equations and the selection of the most likely kinetic model is improved. The algorithm allows to avoid overfitting by continuous model expansion, by cross-validation, by applying the 'widely applicable information criterion' or variance-based model selection. In contrast, detecting the true kinetic scheme with Rate Equations is only successful in combination with cross validation. Adding fluorescence data increases the accuracy of the model parameters by up to two orders of magnitude.**

***Keywords***– Predictive Accuracy, Bayesian cross validation, WAIC, AIC, continuous model expansion, stochastic analysis

Ion channels are essential proteins for the homeostasis of an organism. Disturbance of their function by mutations often causes severe diseases, such as epilepsy[1,2], sudden cardiac death[3] or sick sinus syndrome[4] indicating a medical need[2] to gain further insight into the biophysics of ion channels[5]. The gating of ion channels is usually interpreted by kinetic schemes which are inferred from macroscopic currents with rate equations (REs)[5] or from single-channel currents using dwell time distributions[6–10] or hidden Markov models (HMMs)[11–14]. It is becoming increasingly clear that the use of Bayesian statistics in HMM estimation constitutes a major advantage[15–21]. In ensemble patches, simultaneous orthogonal fluorescence measurement of either conformational changes[22–28] or ligand binding itself[29–33] has increased insight into the complexity of channel activation.

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[5,34–36]. The RE approach is based on a deterministic differential equation derived by averaging the chemical master equation (CME) for the underlying kinetic scheme[37–39]. Its accuracy can be improved by processing the information contained in the intrinsic noise (stochastic gating and binding)[40,41]. 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[42] suffer from cubic computational complexity in the amount of data points, rendering the algorithm impractical for real data. Stepanyuk suggested a faster algorithm[43,44]. Advanced approaches to analyze single-molecule data such as HMMs make use of solutions of the stochastic CME[45 13,14]. 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[46]. HMM approaches are more accurate than fitting dwell time distributions for noisy recordings of rapidly gating channels[14] but the computational complexity limits this type of analysis in ensemble patches to no more than a few hundred channels per time trace[47].

To tame the computational complexity[45], we approximate the solution of the CME with a Kalman filter (KF), thereby remaining in a stochastic framework[48]. 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[49–52]. Our approach generalizes the work of Moffat[47] 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 non-linear dynam-

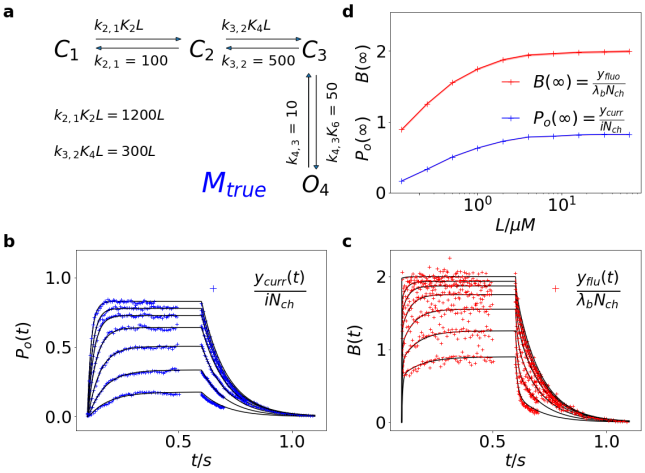


Figure 1: **HMM used for the simulations.** **a**, The kinetic scheme $M_{\text{true}}$ used for simulating the data. The Markov state model (kinetic scheme) consists of two binding steps and one opening step. The rate matrix $\mathbf{K}$ is parametrized by the absolute rates $k_{i,j}$, the ratios $K_i$ between on and off rates (i.e. equilibrium constants) and $L$, the ligand concentration in the solution. The units of the rates are $\text{s}^{-1}$ 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, $\lambda_b = 0.375$ mean photons per bound ligand per frame and single-channel current $i = 1$. The current $y_{\text{curr}}$ and fluorescence $y_{\text{flu}}$ 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 $L$

ics[38,53]. 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[47] 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 $N_{\text{ch}}$ for each time trace, the single-channel conductance $i$ and the mean number $\lambda_{\text{b}}$ 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[54,55]. In contrast, previous work on ensemble currents combined the KF only with maximum likelihood (ML) estimation[47] and did not derive model selection strategies. Difficulties in treating simple stochastic models by ML approaches in combination with the KF[56], 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. Second, the Bayesian approach is still applicable in situations where parameters are not identifiable[57,58] or posteriors are non-Gaussian, whereas ML fitting ceases to be valid[59,60]. Third, a Bayesian approach provides superior model selection tools for singular models such as HMMs[61]. 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[62]. Information criteria allow for model testing on training data instead of hold-out data by performing a bias-corrected computation of the predictive accuracy[62]. We use the recently developed Widely Applicable Information Criterion (WAIC)[63] 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[63] 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 and Supplementary Sec. 2 and Methods). The state evolution $s(t)$ of each individual channel in the patch was sampled with the Gillespie algorithm.[64] 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. At first we assume that the fluorescence signal originates only from bound ligands (Fig. 3), later signal of unbound ligand and the correction using a reference dye is included as well (see Figs. 4-7, Supplementary material, and Meth-

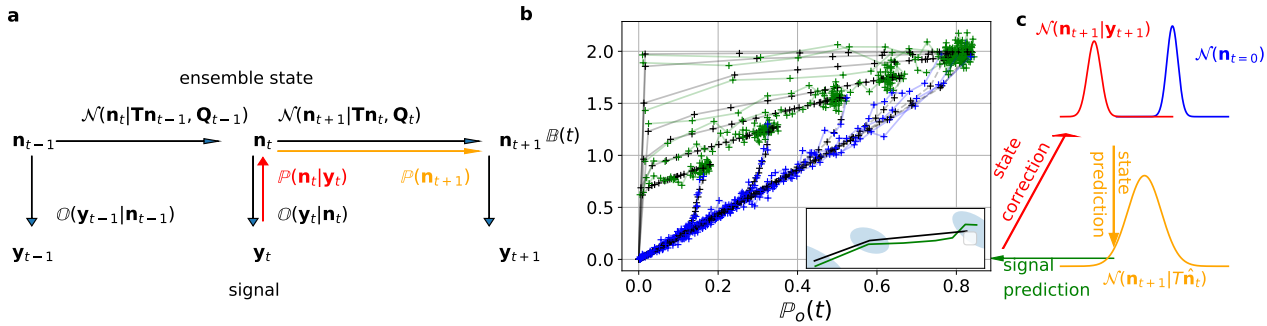


Figure 2: **KF as Bayesian network.** **a,** Graphical model of the conditional dependencies of the stochastic process. Horizontal black arrows represent the conditional multivariate normal transition probability $\mathcal{N}(\mathbf{n}_{t+1}|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t)$ of a continuous state 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 for each time point $\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t)$ in the correction step (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 $\lambda_{\mathrm{B}}$, $i$ and $N_{\mathrm{ch}}$ and the ratio of open-channels $\frac{y_{\mathrm{curr}}}{N_{\mathrm{ch}} i}$ which approximates the open probability $P_{\mathrm{o}}(t)$. 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 inset visualizes the predicted multivariate normal likelihood for three data points of the analysis cycle. The surface of blue ellipses are the points with equal Mahalanobis distance from the predicted mean signal. **c**, Starting from equilibrium assumptions the algorithm constructs a prior (blue). Then the analysis cycle predicts a prior for the next ensemble state (orange) and a likelihood for the signal (green). Subsequently, the algorithm corrects the prior state distribution with the received data (red) which calculates the posterior state distribution. The predicted prior distribution of $\mathbf{n}_{t+1}$ reflects what could be $\mathbf{n}_{t+1}$, given a model with specific parameter $\boldsymbol{\theta}$ set and what we know about $\mathbf{n}_t$. Then the algorithm repeats the cycle for the next data point.

ods 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[65] (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 $\mathbf{T}$

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

$$\mathbf{n}_{t+1} = \mathbf{T}\mathbf{n}_t + \boldsymbol{\omega}_t, \tag{1b}$$

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. 31). 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 $\mathbf{H}$.

$$\mathbf{y}_t \sim \mathbb{O}(\cdot|\mathbf{H}\mathbf{n}_t) := \mathcal{N}(\cdot|\mathbf{H}\mathbf{n}_t, \boldsymbol{\Sigma}_t) \tag{2a}$$

$$\mathbf{y}_t = \mathbf{H}\mathbf{n}_t + \boldsymbol{\nu}_t \tag{2b}$$

The noise of the measurement setup (Supplementary sec. 3 and Eq. 37) 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. 2a 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 $\mathbf{n}_t$

$$\mathbb{P}(\mathbf{n}_t) = \int \mathbb{P}(\mathbf{n}_t|\mathbf{n}_{t-1})\mathbb{P}(\mathbf{n}_{t-1}|\mathbf{y}_{t-1})\mathrm{d}\mathbf{n}_{t-1}, \tag{3a}$$

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. 1a

$$\mathbb{P}(\mathbf{n}_t) = \int \mathcal{N}(\mathbf{n}_t|\mathbf{T}\mathbf{n}_{t-1}, \mathbf{Q}_{t-1})\mathbb{P}(\mathbf{n}_{t-1}|\mathbf{y}_{t-1})\mathrm{d}\mathbf{n}_{t-1} \quad (3b)$$

Each prediction of $\mathbf{n}_t$ (Eq. 3b) is followed by a correction step,

$$\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t) = \frac{\mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)}{\int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t)\,\mathrm{d}\mathbf{n}_t}, \quad (4a)$$

that allows to incorporate the current data point into the estimate, based on the Bayesian theorem.[66] Additionally, the KF assumes[47,67] a multivariate normal observation distribution

$$\mathbb{P}(\mathbf{n}_t|\mathbf{y}_t) = \frac{\mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbf{n}_t, \mathbf{\Sigma}_t)\mathbb{P}(\mathbf{n}_t)}{\int \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbf{n}_t, \mathbf{\Sigma}_t)\mathbb{P}(\mathbf{n}_t)\,\mathrm{d}\mathbf{n}_t}, \quad (4b)$$

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[66]. In that case one can derive algebraic equations for the prediction (Methods Eq. 27 and 31) and correction (Supplementary Sec. 5) 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. 4b is the normal distributed marginal likelihood $\mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})$ for each data point, which constructs by

$$\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \prod_{t=2}^{N_T} \mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta}) \quad (5a)$$

$$= \prod_{t=2}^{N_T} \int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})\,\mathrm{d}\mathbf{n}_t \quad (5b)$$

$$= \prod_{t=2}^{N_T} \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}[\mathbf{n}_t], \mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \mathbf{\Sigma}_t), \quad (5c)$$

a product marginal likelihood of normal distributions of the whole time trace $\mathcal{Y}_\mathrm{T} = \{\mathbf{y}_1, ..., \mathbf{y}_{N_\mathrm{T}}\}$ of length $N_\mathrm{T}$ for the KF. For the derivation of $\mathbf{P}_t$ and $\mathbf{\Sigma}_t$ see Methods Eq. 31 and Eq. 37. The likelihood for the data allows to ascribe a probability to the parameters $\boldsymbol{\theta}$, given the observed data (Methods Eq. 11). An illustration for the operation of the KF on the observation space (Fig. 2**b**) and ensemble state space is shown in Fig. 2**c**. The 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. 4b becomes intractable. By applying the theorem of total variance decomposition[68], we derive the output statistics that approximate various forms of noise and cast them into the form of Eq. 2a (Methods Eq. 49). The Bayesian posterior distribution

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) \sim \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}) \quad (6)$$

encodes all information from model assumptions and experimental data used during model training (Supplementary Sec. 1). Our algorithm reconstructs the posterior (Fig. 3**a**) by sampling from it with the Hamilonian Monte Carlo (HMC) method[69,70] provided by the STAN software[71].

## 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[40,47], 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[47]. For small $N_\mathrm{ch}$, the ratio of the errors decreases like $\sim 1/\sqrt{N_\mathrm{ch}}$, (Fig. 3**c**), thus the RE approach scales $1/N_\mathrm{ch}$ and does not simply scale like the inverse square root of the system size, as one might naively assume. In Fig. 3**d**, the normalized residuals of one time trace are shown which are defined as

$$r_i := \frac{y_i - (\mathbf{H}\mathbb{E}[\mathbf{n}])_i}{\sqrt{\mathrm{var}[y_i]}}. \quad (7)$$

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[47], which is one reason for less precise parameter estimates.

## cPCF versus patch clamp only

To evaluate the advantage of cPCF data[29] with respect to PC data only (Fig. 4), we compare different types of ligands: Idealized ligands with brightness $\lambda_\mathrm{b}$, emitting

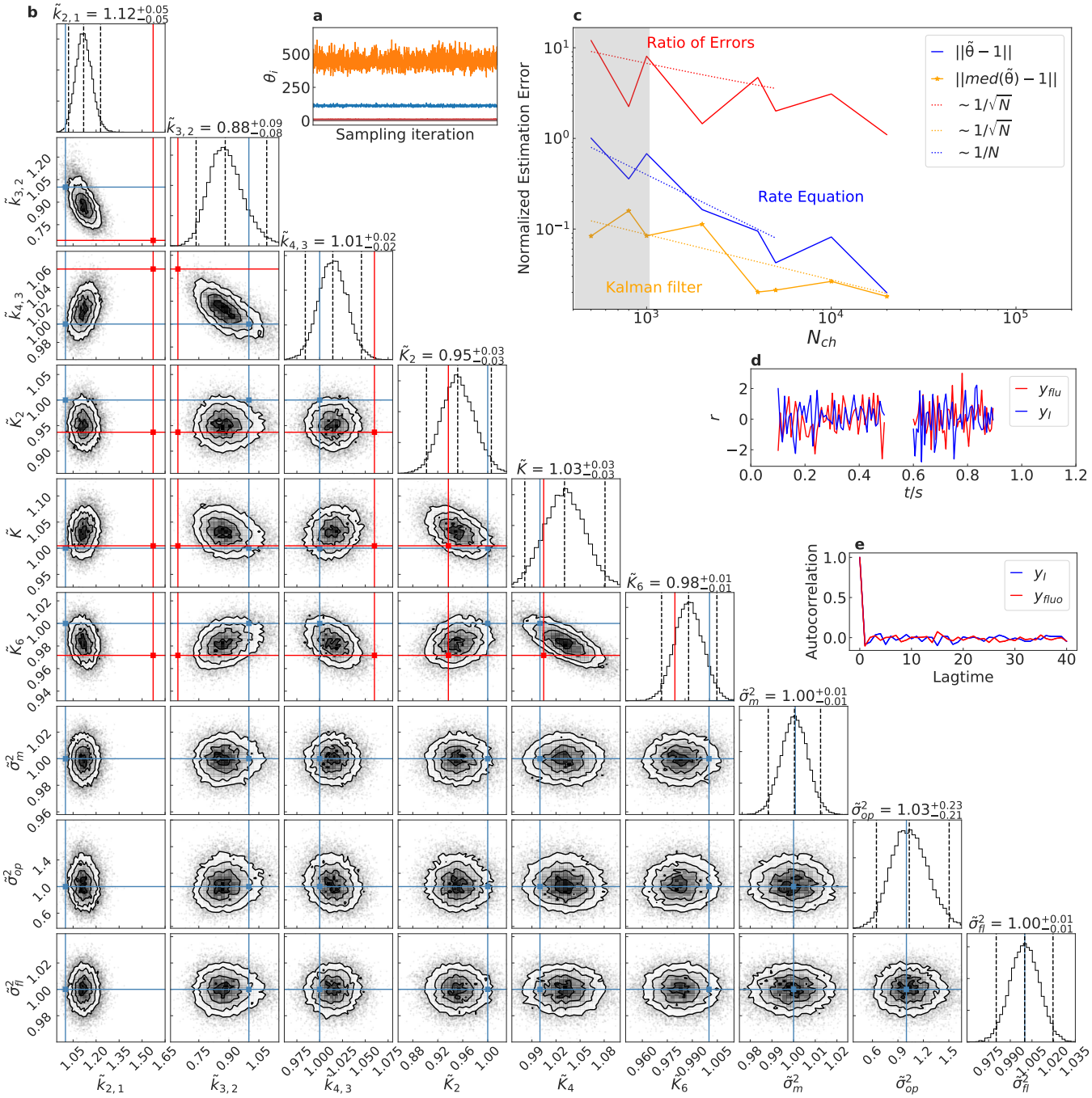


Figure 3: **Benchmark of the KF against least squares fitting of REs.** By $\tilde{\theta}_i$ we indicate that the samples from the posterior are scaled by their true value. **a,** Sample traces of 3 representative parameters from the posterior distribution of the KF algorithm created by Hamiltonian 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 $i$, $N_{\mathrm{ch}}$ and $\lambda$ 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 $N_{\mathrm{ch}}$ . Error ratio (red) between both approaches scales according to $1/\sqrt{N_{\mathrm{ch}}}$ at least for smaller $N_{\mathrm{ch}}$ 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 $N_{\mathrm{ch}}$ 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 $r_{\mathrm{curr}}$ (blue) and $r_{\mathrm{fluo}}$ (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.

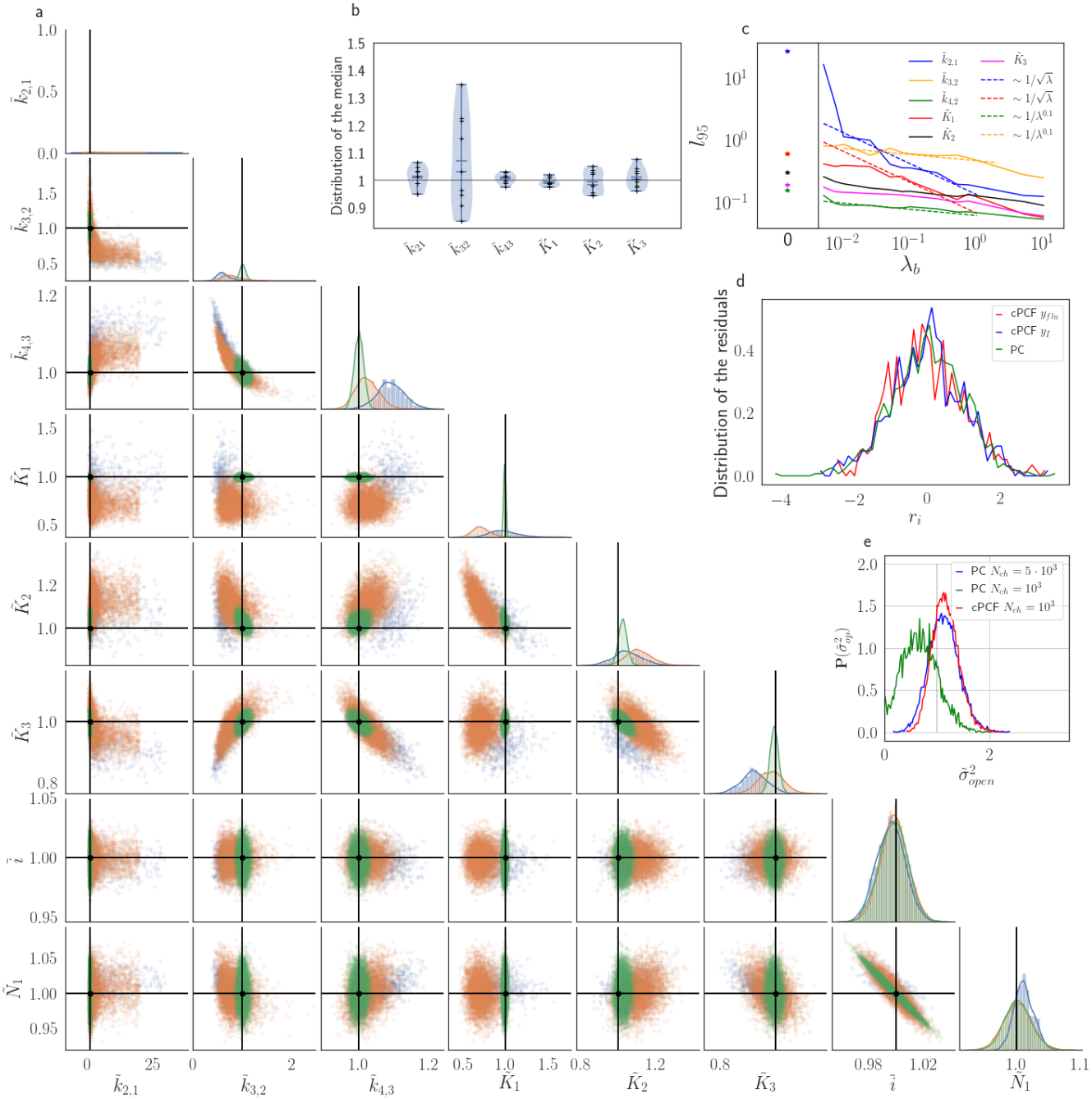


Figure 4: **Benchmark of the KF for patch-clamp versus patch-clamp fluorometry data.** **a**, Posteriors of PC data (blue), cPCF data with $\lambda_\mathrm{b} = 0.00375$ (orange) and cPCF data with $\lambda_\mathrm{b} = 0.375$ (green) but accounting for the superimposing fluorescence of unbound ligands 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 $\lambda_\mathrm{b} = 0.375$ and superimposing bulk signal **c**, The 95th percentile of the marginalized posteriors vs. $\lambda_\mathrm{b}$ 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 $\lambda_\mathrm{b} = 2.5 \cdot 10^{-3}$ data and PC data. The stocasticity of the normalized residuals of the cPCF or PC data are well discribed by $r_i \sim \mathrm{normal}(0, \sigma^2_\mathrm{res} = 1)$. The estimated variance is $\sigma^2_\mathrm{res} = 0.98 + 0.26$. Note that the fluorescence signal per frame of is very low such that it is skewed. **e**, Posterior of the open-channel noise $\mathbb{P}(\sigma^2_\mathrm{op}/\sigma^2_\mathrm{op,true})$ for PC data with $N_\mathrm{ch} \cdot 10^3$ (green) and $N_\mathrm{ch} \cdot 10^5$ (blue) as well as for cPCF data with $N_\mathrm{ch} \cdot 10^3$ (red) with $\lambda_\mathrm{b} = 0.375$. Adding fluorescence data is roughly equal to five times more ion channels to estimate $\sigma^2_\mathrm{op}$. We assumed as prior for the instrumental variance $\mathbb{P}(\sigma^2) = \mathcal{N}(1, 0.01)$.

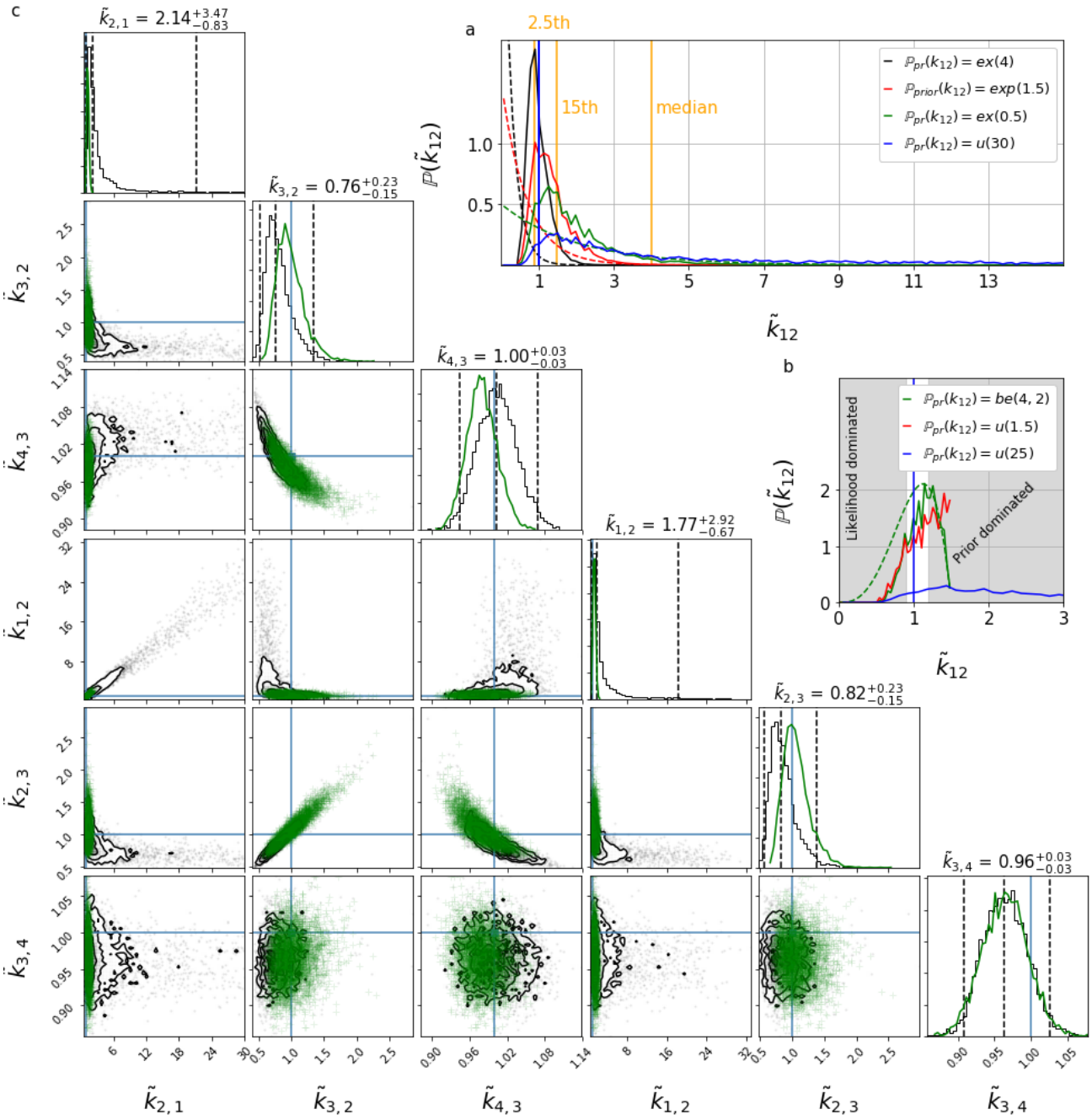


Figure 5: **Variations of the posterior under different prior assumptions** $\tilde{\theta}_i$ indicates that the samples from the posterior are scaled their true value. The vertical orange dashed lines represent the median. **a,** Posterior $\mathbb{P}(k_{1,2})$ of the PC data from $M_{\text{true}}$ as used throughout this article. Posterior rescaled by different exponential priors. The blue vertical line represents the true value. 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 centred 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 could be used as a maximum in an informative prior. A beta distribution as the prior (green dashed curve) 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 $\tilde{k}_{1,2} \in [0.6, 1.4]$. This results in a posterior which is likelihood-dominated below $\tilde{k}_{1,2} = 0.9$ while it is dominated by the prior above $\tilde{k}_{1,2} = 1.1$. The red posterior results from a uniform prior with the same support as the beta prior. Thus, the difference between blue and red indicates the information which is contained in the support. The difference between red and green encodes the information added by the beta prior. **c,** Consequence of the informative prior $k_{1,2}/1800 \cdot \text{s}\mu\text{M} = \tilde{k}_{1,2} \sim beta(4, 2)$ on the posterior of the complete rate matrix.

light only when bound to the channels, and 'real' ligands which also produce background fluorescence when diffusing in the bath solution (Supplementary Sec. 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 $N_{\mathrm{ch}}$ 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 $\lambda$, 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 $\lambda_{\mathrm{b}}$. 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

$$\frac{(\mathbf{H}\mathbf{P}_t\mathbf{H}^\top)_{2,2}}{\sigma_{\mathrm{op}}^2} \approx \frac{1 - p_{\mathrm{o,max}}}{\sigma_{\mathrm{op}}^2} \geq 20. \tag{8}$$

## 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 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[72]. For the first binding step we used an approximation for small ligand receptor interactions[73] $k_{\mathrm{binding}} = 600\,\mu\mathrm{M}^{-1}\mathrm{s}^{-1}$. Thus, for two available binding sites the stoichiometry increases the upper limit to $k_{1,2} = 1,200\,\mu\mathrm{M}^{-1}\mathrm{s}^{-1}$. Here we investigate priors ,with different information content, to model from the mentioned knowledge resulting *a priori* plausibility for $k_{1,2}$. Reported binding rates, e.g. on isolated binding domains of CNGA2 channels[74], are indeed at or below this value. Traditional Bayesian or frequentist approaches, using uniform priors $\mathrm{u}(l) = \frac{1}{l}$ 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[75]. In the strong data context, the posterior is dominated by the likelihood and the prior information is minor[76]. In contrast, for modeling situations with weakly informative data (Fig. 5 **a**) a educated prior selection is influencing 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 a larger $k_{2,1}$ and a smaller $k_{3,2}$ (Fig. 5 **d**). Nevertheless, all probability mass of $\mathbb{P}(k_{1,2})$ above the diffusion limit of binding (Fig. 5 **a**) is physically impossible, though plausible 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

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

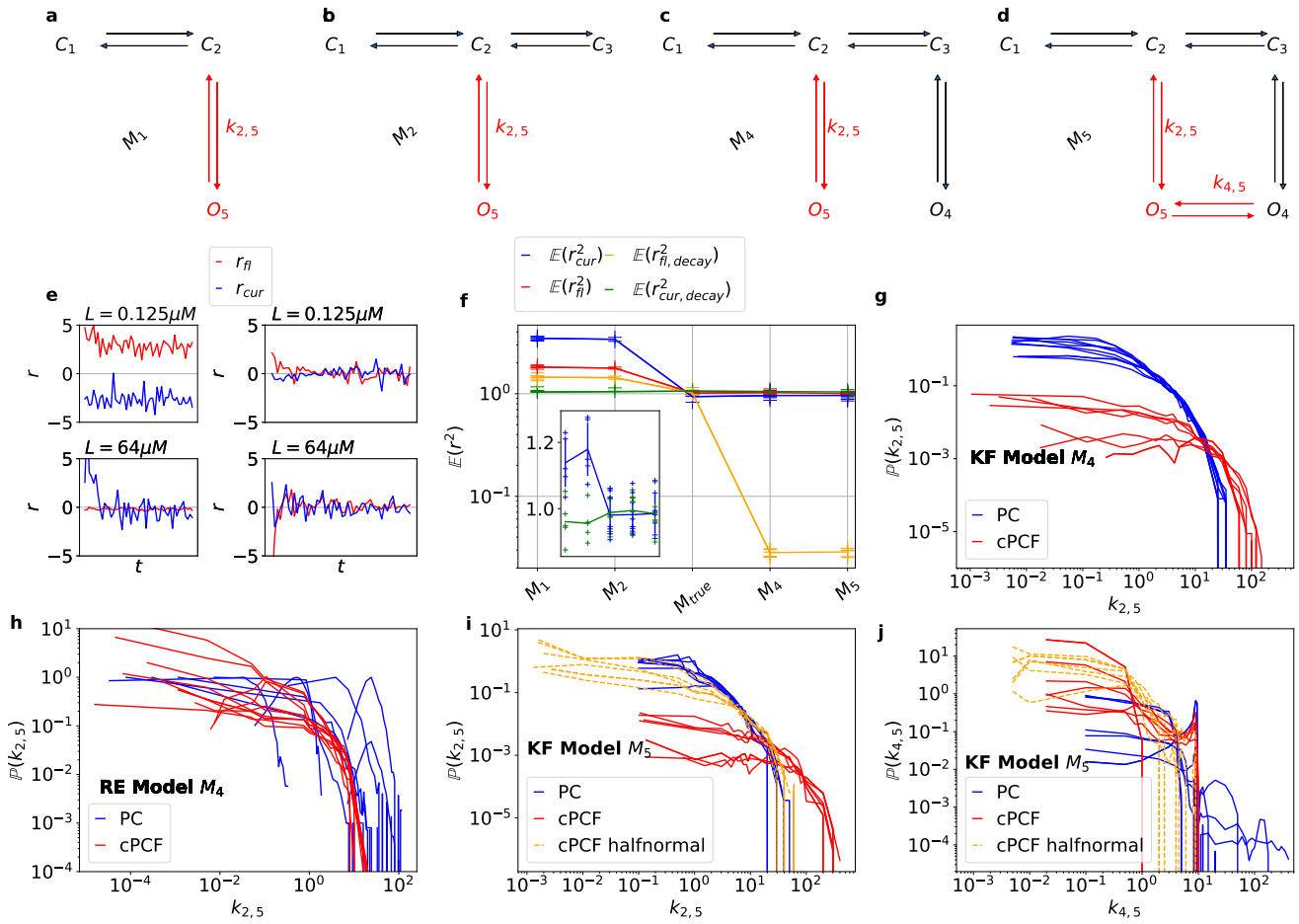


Figure 6: **Model selection by the second moment of the residuals and by continuous model expansion. a-d**, Model structures of the trained models differing from the true process (c.f. Fig. 1a). Red states and transitions have no counterpart in the true model. All models are nested in the most complex 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 $\approx 12.5$. **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 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 $M_{\mathrm{true}}$ 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 probabilty mass for $k_{4,5} > 10$. Hence, for $k_{4,5}$ cPCF data constrains 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 suggest omitting state $O_5$.

The parameter $\zeta$ refers to the parameter which scales the statistics, $\mathbb{E}[k_{1,2}] = 1/\zeta$ and $\mathrm{var}[k_{1,2}] = 1/\zeta^2$. Notably, the exponential distribution is a maximum entropy distribution if two conditions are met [77]. The parameter has to be positive and the mean of the parameter must be known. On the onside 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. 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 then the diffusion limit. The exponential prior states in fact the opposite it has the highest density at zero.

The beta distribution

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

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

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

and that

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

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 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. $k_{1,2}/1,800 \cdot \mathrm{s}\mu\mathrm{M} \sim \mathrm{beta}(4,2)$ 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 **d**) 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 area the data is really informative. In contrast, for $\tilde{k}_{1,2} > 1$ the prior is the most important source of information for the posterior. Note, that for the other rates which have not been given themselves any informative prior, the beta prior for $k_{1,2}$ improves their parameter inference by concentrating the posterior (Fig. 5 **d**)) around the true values. That effect is strongest for $k_{2,1}$ due to the strongest correlations between 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.

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

We compare three methods of model selection: continuous model expansion, the statistics of $r(t)$, and the predictive accuracy estimated either by cross-validation or information criteria on a set of candidate models $M_1, M_2, M_{\mathrm{true}}, M_4, M_5$ (Fig. 6**a-d**).

The KF enables to identify underfitting better by plotting the residuals $r(t)$ (Fig. 6**e**) than signal time traces because the large amplitude changes of the mean current obscures the relatively small amplitude of the systematic errors. The estimated second moment of $r(t)$ is plotted for the different models (Fig. 6**f**). For the true model $M_{\mathrm{true}}$ the estimated second moment equals the variance and, since we normalized the residual traces, the second non-centralized moment should be close to 1. In fact all variance is explained by $M_{true}$. Overfitting models $M_4$ and $M_5$ are detected by the decrease of the fluorescence variance in the decaying part of the traces. For PC data only underfitting can be detected as long as the modeler looks out for the simplest model which does not underfit the detection of a good model is still possible.

The conceptional idea of continuous model expansion is to sample from a model structure which contains the true process and a lot of additional model structure whose rates are set to zero by the algorithm when the data quality or quantity increases. In other words one assumes a complex super model $M_5$ which includes all simpler models as a limiting case $k_{i,j} \to 0$. A simpler model can be chosen if the posterior has a distinctive maximum for $k_{i,j} = 0$. Testing continuous model expansion by $M_4$ with the KF, the posterior for only one out of ten data sets shows a local maximum for $k_{2,5} \neq 0$ (Fig. 6**g**). Thus, the KF switches off non-existing states for most data sets. By contrast, the corresponding analysis of current-only data by REs reveals a peaking posterior for $k_{2,5} \neq 0$ in several cases (Fig. 6**h**). Additional fluorescence data reduce the occurrence of those peaks.

If the PC or cPCF data are fitted with model $M_5$ by using the KF, for most data sets the rates into $O_5$ maximize the posterior if they vanish (Fig. 6**i-j**). For PC data, the posterior reveals multi-modality with some data sets

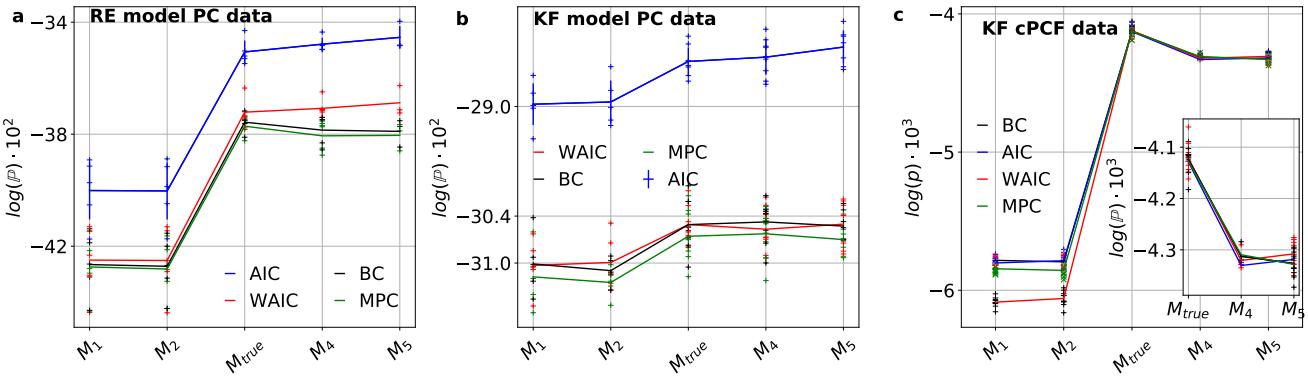


Figure 7: **Bayesian Model selection** Compared is the predictive accuracy of the indicated five models estimated by either the Akaike Information Criterion (AIC), the Widely Applicable Information Criterion (WAIC), Bayesian cross-validation (BC) or Maximum *a posteriori* cross-validation (MPC).We use 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 divers information about the deactivation about the rates. The inset shows a part of the diagram at an extended ordinate.

(Fig. 6**j**). Hence, point estimates of the parameters are not reliable while the posterior of $M_5$ encodes all model uncertainties. Notably, this multimodality occurs also for cPCF data though less pronounced. Thus, both experiments share the tendency to create a finite peak around $k_{4,5} = 10$, indicating the false detection of an additional open state if not analyzed with caution.

Applying a weakly informative prior distribution supports the model determination in the continuous model expansion case. The advantage of continuous model expansion is that it reduces the risk of finding a local optimum on the discrete model space rather than a global optimum by translating the model space from a discrete to continuous model space. The disadvantage of having a lot of possible structure in the model makes the model quickly too flexible to come to a conclusive posterior with a limited amount of data. Many parameter sets can fit the data roughly equally likely. Thus, prior distributions are needed to support the algorithm to select simpler base models, which means concentrating for certain $k_{i,j}$ the posterior around zero[78]. To exemplify this, we use a weakly informative prior distribution on $k_{5,2}$ (Supplementary Sec. 1.1) to show how prior information alleviates model pathology due to excessive model flexibility. Heuristically, one should be sceptical about rates which are faster than the sampling frequency because they could generate eigenvalues $\lambda^5$ of the rate matrix, which are smaller than the time between two sampling points. Here we used a sampling frequency of $10k$Hz The frequency by which the KF analysed the data ranged from 83.3Hz to 500Hz depending on the kinetics. The weakly informative half-normal prior $k_{5,2} \sim \text{half-normal}(0, 6 \cdot 10^3)$, which penalizes unrealistic high rates of $k_{5,2}$, is necessary because the data are not able to constrain that rate. Applying this prior distribution suppresses secondary peaks and probability mass in the distribution tails of the other rates $k_{2,5}$ and $k_{4,5}$. This further emphasizes to leave out $O_5$ in the final model (Fig. **i-j**).

We use this prior to argue that one loses a lot descriptive power of a data set if one tries to be objective by using uniform priors in particular with unrealistic large intervals. The notion of being unbiased with a flat prior, where the likelihood does not dominate the prior, ends up in paradoxes[79]. 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 local optimum in the model space. Nevertheless, one should drop from the super model 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[62]. 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[77,80–82]. 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.[82] Some of them, such as Maximum evidence or BIC,[80] rely on the assumption that the true data-generating process is included in the set of models to be tested. That is called the M−closed situation[80]. These approaches perform well in simulation studies[83] 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[84]. 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[77] which is in principle not accessible. But the model with the minimum Kullback-Leibler divergence within a set candidate models can be found asymptotically by maximizing the predictive accuracy[62,85]. The predictive accuracy for a specific data set $\tilde{\mathcal{Y}} = \{\tilde{y}_1, ..., \tilde{y}_{N_{\text{data}}}\}$, which has not been used for model training[62], is defined as

$$\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}) = \log \mathbb{E}_\theta[\mathbb{L}(\tilde{\mathcal{Y}}|\boldsymbol{\theta})] \tag{9}$$

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 $\mathbb{F}(\tilde{y}, t)$

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})] = \int \log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})\mathbb{F}(\tilde{\mathcal{Y}}) \prod_{i=1}^{N_{\text{data}}} \mathrm{d}\tilde{y}_i \tag{10}$$

Maximizing Eq. 10 is equivalent to minimizing the Kullback-Leibler distance which specifies the information loss when approximating the unknown $\mathbb{F}(y)$ by a model[86?] . Because $\mathbb{F}(y)$ is unknown, one has to estimate Eq. 10 by using the available samples from $\mathbb{F}(y)$. Therefore, different estimators of Eq. 10 are compared:

AIC, WAIC, maximum *a posteriori* cross-validation and Bayesian cross-validation (see Methods). As 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 to 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[63] succeeds (Fig. 7**b**) while AIC fails to estimate the predictive accuracy distribution. It suggests a better predictive accuracy with complexer model structures (Fig. 7**b**). The failing AIC correlates with the occurrence of a non-normal posterior distribution[60](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 $M_{\text{true}}$ 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 $M_{\text{true}}$ 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. 27 and 31), the output statistics (Methods Eq. 49) and correction equations (Supplementary Sec. 5) 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 $\sim 10$ 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.

# Acknowledgements

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.


# Author contributions

Conceptualization: J.M, K.B.; Methodology: J.M., R.S., F.P.; Software, investigation, formal analysis: J.M.; Writing: J.M. F.P. K.B., Supervision: K.B.

# Competing Interests

The authors declare no competing interests.

# 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 Supplementary materials.

## 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 [54,55], 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 information that we possess about the parameter values before or after analysing the data, respectively. According to the Bayesian theorem the posterior distribution

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) = \frac{1}{Z(\mathcal{Y}_T)} \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) \prod_j^{N_{\text{par}}} \mathbb{P}(\theta_j) \tag{11}$$

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 stochasticity 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 $\boldsymbol{\theta}$ (Supplementary Sec. 1). The normalization constant

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

ensures that the posterior is a normalized distribution. Our proposed Bayesian network [65], which is a generalisation of the classical KF, delivers a set of recursive algebraic equations [66,67] 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. That allows us to replace $\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})$ of Eq. 11 by the expression of Eq. 5c.

The Bayesian framework has various properties which makes it superior to maximum likelihood estimation (MLE) [77]. 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. 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 (Supplementary Sec. 1).

## 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 $\alpha$ at time $t$. Mathematically, $\mathbf{e}_\alpha$ stands for the $\alpha$-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:

$$\mathbb{P}(\text{path}) = \mathbb{P}(\mathbf{s}(0), \mathbf{s}(1), \ldots, \mathbf{s}(T)) = \mathbb{P}(\mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(1) \mid \mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(2) \mid \mathbf{s}(1)) \cdots \mathbb{P}(\mathbf{s}(T) \mid \mathbf{s}(T-1)). \tag{13}$$

Markov models (MMs) and rate models are widely used for modeling molecular kinetics (Supplementary 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[87]. 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:

$$\mathbf{T}_{\alpha,\beta} := \mathbb{P}(\mathbf{s}(t+1) = \mathbf{e}_\alpha \mid \mathbf{s}(t) = \mathbf{e}_\beta) = \exp(\mathbf{K} \cdot \Delta t)_{\alpha,\beta}, \tag{14}$$

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[64]. 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 $\alpha$. One can interpret the probability vector $\mathbf{p}$ as the instantaneous expectation value of the state vector $\mathbf{s}$.

$$\mathbf{p}(t) = \mathbb{E}(\mathbf{s}(t)) \tag{15}$$

The probability vector obeys the discrete-time Master equation

$$\mathbf{p}(t+1) = \mathbf{T}\mathbf{p}(t) \tag{16}$$

$$\mathbb{E}(\mathbf{s}(t+1)) = \mathbf{T}\mathbb{E}(\mathbf{s}(t)) \tag{17}$$

## 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 $t$:

$$\mathbf{n}(t) = \sum_{i=1}^{N_{\text{ch}}} \mathbf{s}_i(t) \tag{18}$$

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. 18 in the channel dimension and from the single-channel CME Eq. 17 one can immediately derive the equation for the time evolution of the mean occupancy $\overline{\mathbf{n}}(t) = \mathbb{E}[\mathbf{n}(t)]$:

$$\overline{n}_\alpha(t+1) = \sum_\beta T_{\alpha,\beta} \overline{n}_\beta(t) \tag{19}$$

| 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 |
| $\sigma^2_{\mathrm{m}}$ | variance of the current including all noise from the patch and the recording system |
| $\sigma^2_{\mathrm{op}}$ | variance of the current noise generated by a single open-channel |
| $\lambda_{\mathrm{b}}$ | mean brightness of a bound ligand |
| $\lambda_{\mathrm{Fl}}$ | mean brightness of the fluorescence signal from bulk and bound ligands |
| $\sigma^2_{\mathrm{bulk}}$ | 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 |
| $N_{\mathrm{obs}}$ | dimensions of the observational space |
| $\mathbb{F}(\mathcal{Y})$ | true probability density of $\mathcal{Y}$, i.e. the true data-generating process |
| $\mathbb{L}(\mathcal{Y}\|\boldsymbol{\theta})$ | likelihood function of the model parameters |
| $\mathbb{P}(\boldsymbol{\theta}\|\mathcal{Y})$ | posterior distribution of the model parameters |
| $\mathbb{P}_{\mathrm{pred}}(\widetilde{\mathcal{Y}}\|\mathcal{Y})$ | predictive distribution of the new data points $\widetilde{\mathcal{Y}}$ |
| $\mathbb{O}(\mathbf{y}\|\mathbf{n})$ | distribution of observables for a single time step |
| $\mathcal{N}(\cdot)$ | normal distribution |
| $\mathbb{E}[\cdot]$ | mean value |

Table 1: Important symbols

with 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 $\alpha$. 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 $\alpha$ of the single-channel transition matrix $\mathbf{T}$. According to the theory of Markov models, the final channel distribution of channels originating from state $\alpha$ is the multinomial distribution

$$\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha) = \mathbb{P}(n_1, ..., n_M \mid \mathbf{n}(t) = n_\alpha \mathbf{e}_\alpha) = \frac{n_\alpha!}{n_1! \cdots n_M!} T_{1,\alpha}^{n_1} \cdots T_{M,\alpha}^{n_M} \tag{20}$$

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 $\alpha$ gives rise to its own final multinomial distribution that contributes $n_\beta^{(\alpha)}$ transitions into state $\beta$ 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 $\alpha$.

$$\mathbf{n}(t+1) = \sum_\alpha \mathbf{n}^{(\alpha)}(t+1) \tag{21}$$

Evidently, the total number of channels is conserved during propagation. The distribution of $\mathbf{n}(t+1)$, defined by Eqs. 20 and 21, is called the *generalized multinomial distribution*:

$$\mathbf{n}(t+1) \sim \text{general-multinomial}(\mathbf{n}(t), \mathbf{T}) \tag{22}$$

While no simple expression exists for the generalized multinomial distribution, closed form expressions for its moments can be readily derived. For large $N_{\text{ch}}$ 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 general-multinomial$(\mathbf{n}(t), \mathbf{T})$ has an 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 (20) has the following mean and covariance matrix

$$\overline{\mathbf{n}}^{(\alpha)}(t+1) = n_\alpha \mathbf{T}_{:,\alpha} \tag{23}$$

$$\mathbf{\Sigma}^{(\alpha)}(t+1) = n_\alpha \text{diag}(\mathbf{T}_{:,\alpha}) - n_\alpha \mathbf{T}_{:,\alpha,:} \mathbf{T}_{:,\alpha}^\top \tag{24}$$

where $\mathbf{T}_{:,\alpha}$ denotes the column number $\alpha$ of the transition matrix and diag$(\mathbf{T}_{:,\alpha})$ describes the diagonal matrix with $\mathbf{T}_{;,\alpha}$ on its diagonal. Combining Eq. 21 with Eqs. 23 and 24 we deduce the mean and variance of the generalized multinomial distribution:

$$\mathbb{E}\left(\mathbf{n}(t+1) \mid \mathbf{n}(t)\right) = \sum_\alpha n_\alpha(t) \mathbf{T}_{:,\alpha} = \mathbf{T}\mathbf{n}(t) \tag{25}$$

$$\text{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right) = \sum_\alpha n_\alpha(t) \left(\text{diag}(\mathbf{T}_{:,\alpha}) - \mathbf{T}_{:,\alpha}\mathbf{T}_{:,\alpha}^\top\right) = \text{diag}\left(\mathbf{T}\mathbf{n}(t)\right) - \mathbf{T}\text{diag}\left(\mathbf{n}(t)\right)\mathbf{T}^\top \tag{26}$$

Note that Eqs. 25 and 26 are conditional expectations that depend on the random state $\mathbf{n}$ at the previous time $t$ and not only on the previous mean $\overline{\mathbf{n}}$. To find the absolute mean, the law of total expectation is applied to Eq. 25, giving

$$\overline{\mathbf{n}}(t+1) = \mathbb{E}\left[\mathbb{E}\left(\mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] = \mathbf{T}\overline{\mathbf{n}}(t), \tag{27}$$

in agreement with the simple derivation of Eq. 19. Similarly, the absolute covariance matrix of $\mathbf{n}(t+1)$ can be found by applying the law of total variance decomposition[68] to Eqs. 26 and 25, giving

$$\mathbf{P}(t+1) := \text{cov}(\mathbf{n}, \mathbf{n})(t+1) = \mathbb{E}\left[\text{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] + \text{cov}\left[\mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t)), \mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t))\right] \tag{28}$$

$$= \text{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\text{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^\top + \text{cov}(\mathbf{T}\mathbf{n}(t), \mathbf{T}\mathbf{n}(t)) \tag{29}$$

$$= \text{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\text{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^\top + \mathbf{T}\text{cov}(\mathbf{n}(t), \mathbf{n}(t))\mathbf{T}^\top \tag{30}$$

$$= \text{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\text{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^\top + \mathbf{T}\mathbf{P}(t)\mathbf{T}^\top \tag{31}$$

where we have introduced the shorthand $\mathbf{P}(t) = \mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))$ in the last line. Eqs. 27, 31 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 [47,67](Supplementary Sec. 5). 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. 27, 31. Comparing Eq. 31 with the KF prediction equation [67] for $\mathbf{P}(t)$ we obtain the state-dependent covariance matrix of Eq. 1 as

$$\mathbf{Q}(\mathbf{T}, \overline{\mathbf{n}}(t)) = \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} \tag{32}$$

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 $\mathbb{O}(\mathbf{y}|\mathbf{n}(t))$, (Supplementary Sec. 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.

$$\mathbf{y}(t) = \begin{pmatrix} \text{fluorescence intensity}(t) \\ \text{ion current}(t) \end{pmatrix} = \begin{pmatrix} y_{\text{flu}}(t) \\ y_{\text{curr}}(t) \end{pmatrix} \tag{33}$$

As outlined in the introduction part, in Eq. 2b 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. 31). The observation matrix $\mathbf{H} \in \mathbb{R}^{N_{\text{obs}} \times M}$ projects the hidden state vector $\mathbf{n}(t)$ onto $\mathbf{H}\mathbf{n}(t) \in \mathbb{R}^{N_{\text{obs}}}$, the observation space. The observation distribution is

$$\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t)) = \mathcal{N}\big(\mathbf{y}(t)|\mathbf{H}\mathbf{n}(t), \mathbf{\Sigma}(\mathbf{n}(t))\big) \tag{34}$$

$$\Leftrightarrow \mathbf{y}(t) = \mathbf{H}\mathbf{n}(t) + \boldsymbol{\nu}(\boldsymbol{t}). \tag{35}$$

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

| | 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.

signals and four noise sources one by one. For this we decompose the observation matrix and the observation noise covariance matrix into the individual terms:

$$\mathbf{H} = \mathbf{H}_{\mathrm{I}} + \mathbf{H}_{\mathrm{binding}} \tag{36}$$

$$\mathbf{\Sigma}(t) = \mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{meas.}} + \mathbf{\Sigma}_{\mathrm{binding}}(t) + \mathbf{\Sigma}_{\mathrm{back}} \tag{37}$$

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 $\mathbf{\Sigma}_{\mathrm{back}}$ refer to (Supplementary Sec. 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 $\sigma^2_{\mathrm{open}}$. 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 $\mathbb{E}[\nu_I^2(t)] = \sigma^2_{\mathrm{op}} n_4(t) + \sigma^2_{\mathrm{m}}$. 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

$$\mathbf{H}_{\mathrm{I}} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \end{pmatrix} \tag{38}$$

$$\mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{meas.}} := \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{\mathrm{op}} n_4(t) + \sigma^2_{\mathrm{m}} \end{pmatrix} \tag{39}$$

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

$$\begin{aligned} \mathrm{var}(\mathbf{y}(t)) &= \mathbb{E}[\mathrm{var}[\mathbf{y}(t)|\mathbf{n}(t)]] + \mathrm{var}[\mathbb{E}[\mathbf{y}(t)|\mathbf{n}(t)]] && \text{(40a)} \\ &= \mathbb{E}[\sigma^2_{\mathrm{op}} n_4(t) + \sigma^2_{\mathrm{m}}] + \mathrm{var}[\mathbf{H}_{\mathrm{I}}\mathbf{n}(t)] && \text{(40b)} \\ &= \sigma^2_{\mathrm{op}} \mathbb{E}[n_4(t)] + \sigma^2_{\mathrm{m}} + (\mathbf{H}_{\mathrm{I}}\mathbf{P}(t)\mathbf{H}_{\mathrm{I}}^{\top})_{2,2} && \text{(40c)} \end{aligned}$$

See Supplementary 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 $\lambda_{\mathrm{Fl}}$

$$y_{\mathrm{Fl}}(t) \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}(t)). \tag{41}$$

The total emission rate $\lambda_{\mathrm{Fl}}$ can be modeled as a weighted sum of the specific emission rates $\lambda_b$ 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. 34), we invoke the central limit theorem to approximate

$$y_{\mathrm{Fl}} \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}) \approx \mathcal{N}(\lambda_{\mathrm{Fl}}(t), \lambda_{\mathrm{Fl}}(t)) \tag{42}$$

The larger $\lambda_{\mathrm{Fl}}$ 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

$$\mathbf{H}_{\text{binding}} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{43}$$

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 $\lambda_{\text{Fl}}(t) = \lambda_b[n_2(t) + 2(n_3(t) + n_4(t))]$. The variance-covariance matrix $\mathbf{\Sigma}_{\text{binding}}$ can be derived along the same lines using Eq. 42. We find

$$\mathbf{\Sigma}_{\text{binding}}(t) := \begin{pmatrix} (\mathbf{Hn}(t))_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{44}$$

Under these assumptions the observation matrix can be written as follows

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

## 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

$$\mathbf{y}(t) = \mathbf{Hn}(t) + \boldsymbol{\nu}_{\text{m}}(t) + \begin{pmatrix} \nu_{\text{pois}}(t) \\ \nu_{\text{op}}(t) \end{pmatrix} \Leftrightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{Hn}(t), \mathbf{\Sigma}(t)). \tag{46}$$

The vector $\boldsymbol{\nu}_{\text{m}}$ denotes the experimental noise, with $\mathbb{E}[\boldsymbol{\nu}_{\text{m}}] = 0$ and variance given by the diagonal matrix $\mathbf{\Sigma}_{\text{meas}} + \mathbf{\Sigma}_{\text{back}}$. The second noise term arises from Poisson-distributed photon counting statistics and the open-channel noise. It has the properties $\mathbb{E}\left[\begin{pmatrix} \nu_{\text{pois}}(t) \\ \nu_{\text{op}}(t) \end{pmatrix}\right] = 0$ and $\text{cov}\left(\begin{pmatrix} \nu_{\text{pois}}(t) \\ \nu_{\text{op}}(t) \end{pmatrix}, \begin{pmatrix} \nu_{\text{pois}}(t) \\ \nu_{\text{op}}(t) \end{pmatrix}\right) = \mathbf{\Sigma}_{\text{open}}(t) + \mathbf{\Sigma}_{\text{binding}}(t)$. The matrix $\mathbf{\Sigma}$ is a diagonal matrix. To derive the covariance matrix $\text{cov}(\mathbf{y}(t))$ we need to additionally calculate $\text{var}(y_{\text{fluo}}(t))$ and $\text{cov}(y_{\text{fluo}}(t), y_{\text{patch}}(t))$. By the same arguments as above we get

$$\begin{aligned} \text{var}[y_{\text{fluo}}(t)] &= \mathbb{E}[\text{var}(y(t)|\mathbf{n}(t))] + \text{var}[\mathbb{E}(y(t)|\mathbf{n}(t)] && \text{(47a)} \\ &= \mathbb{E}[\sigma^2_{\text{back}} + (\mathbf{Hn}(t))_1] + \text{var}(\mathbf{Hn}(t)) && \text{(47b)} \\ &= \sigma^2_{\text{back}} + (\mathbf{Hn}(t))_1 + (\mathbf{HP}(t))\mathbf{H}^\top)_{1,1} && \text{(47c)} \end{aligned}$$

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

$$\begin{aligned} \text{cov}(y_{\text{patch}}, y_{\text{fluo}}) &= \mathbb{E}[\text{cov}(y_{\text{patch}}, y_{\text{fluo}}|\mathbf{n})] + \text{cov}(\mathbb{E}(y_{\text{patch}}|\mathbf{n}), \mathbb{E}(y_{\text{fluo}}|\mathbf{n})) && \text{(48a)} \\ &= 0 + \text{cov}(\mathbf{H}_{2,:}\mathbf{n}, \mathbf{H}_{1,:}\mathbf{n}) && \text{(48b)} \\ &= \mathbf{H}_{2,:}\,\text{cov}(\mathbf{n}, \mathbf{n})\mathbf{H}^\top_{1,:} = \mathbf{H}_{2,:}\mathbf{P}(t)\mathbf{H}^\top_{1,:} && \text{(48c)} \end{aligned}$$

yielding the matrix

$$\text{cov}(\mathbf{y}, \mathbf{y}) = \mathbf{HP}(t)\mathbf{H}^\top + \mathbf{\Sigma}(t) \tag{49}$$

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 $N_{\text{ch,FL}} = N_{\text{ch,I}}$. For real data further refinement might be necessary to model the stochasticity of the sub-ensemble in the summed voxels. With the time evolution equations for the mean (Eq. 25) and for the covariance matrix (Eq. 31) as well as with the expressions for the signal variance we possess all parameters that are needed in the correction equation of the KF[48,67].

## Bayesian Model selection via predictive accuracy

The minimum of the Kullback-Leibler divergence can be found asymptotically by maximizing the predictive accuracy[62,85] The predictive accuracy is defined as

$$\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}_T) := \log \mathbb{E}_{\theta,\text{post}}[\mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})] = \log \int \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})\mathbb{P}_{\text{post}}(\boldsymbol{\theta}|\mathcal{Y}_T)\,\mathrm{d}\boldsymbol{\theta} \tag{50}$$

| 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[40]. 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 stochasticity and the hidden structure of an ion channel ensemble of finite size is ignored. The RE approach does not predict $\mathbf{P}(t)$ nor has any correction step, thus the stochasticity and the hidden structure of an ion channel ensemble of finite size is ignored. It only predicts the mean ensemble states, 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[40]. The multinomial distribution is then approximated by a normal distribution and the variance from the experimental noise is added. 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.

for a specific new (held-out) data set $\tilde{\mathcal{Y}}_T$, which has not been used for training the model[62]. $\mathbb{E}_{\mathrm{post}}[\cdot]$ denotes the average of some function found by integrating it over the posterior distribution $\mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta}|\mathcal{Y}_T)$. The difference between maximum likelihood (or maximum *a posteriori* cross-validation, MPC) to Bayesian cross-validation in Eq. 50 is that the mentioned point estimates in MPC would yield a Dirac distribution $\mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta}|\mathcal{Y}_T) = \delta(\boldsymbol{\theta}_{\mathrm{MLE}} - \boldsymbol{\theta})$ as the posterior. Doing so collapses the integral in Eq. 50 to $\log \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta}_{\mathrm{MLE}})$. Note, that Eq. 50 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

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] = \int \cdots \int \log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)\mathbb{F}(\tilde{\mathcal{Y}}_T) \prod_i^T \mathrm{d}\tilde{\mathbf{y}}_i \tag{51}$$

Unfortunately, $\mathbb{F}(\tilde{\mathcal{Y}}_T)$ is unknown. In practice the expectation value in Eq. 51 is approximated summing over independent realizations generated by the experiment.

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] \approx \sum_{k=1}^{K}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T^{(k)}|\mathcal{Y}_T)] \tag{52}$$

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[62]. To compensate for that optimistic bias one needs to penalize the model complexity[62] which is done by scores called information criteria. The first non-Bayesian information criterion was found by Akaike[88]. 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

$$\mathrm{AIC} = \log\, \mathbb{P}(\mathcal{Y}_T|\hat{\boldsymbol{\theta}}_{\mathrm{mle}}) - \dim(\boldsymbol{\theta}). \tag{53}$$

Under those conditions AIC is asymptotically equivalent to cross-validation[89]. These very restrictive model conditions are not satisfied by models with hidden or latent structures[60]. Thus AIC has no mathematical justification

| 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 $\mathrm{r}^2$ (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 $N_{\mathrm{data}}$ to the true value.

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 is strong enough to create a multivariate normal posterior. Recently, Watanabe[63] 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. 51 can be estimated by

$$\mathrm{WAIC} := \log \mathbb{P}_{\mathrm{post}}(\mathcal{Y}_T) - p_{\mathrm{WAIC}} \tag{54}$$

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[62]. 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. 51 by Eq. 50 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.

# References


[1] Oyrer, J. *et al.* Ion channels in genetic epilepsy: from genes and mechanisms to disease-targeted therapies. *Pharmacol. Rev.* **70**, 142–173 (2018).

[2] Goldschen-Ohm, M. P., Wagner, D. A., Petrou, S. & Jones, M. V. An epilepsy-related region in the GABAA receptor mediates long-distance effects on GABA and benzodiazepine binding sites. *Mol. Pharmacol.* **77**, 35–45 (2010).

[3] Clancy, C. E. & Rudy, Y. Cellular consequences of HERG mutations in the long QT syndrome: precursors to sudden cardiac death. *Cardiovascular research* **50**, 301–313 (2001).

[4] Verkerk, A. O. & Wilders, R. Pacemaker activity of the human sinoatrial node: effects of HCN4 mutations on the hyperpolarization-activated current. *Europace* **16**, 384–395 (2014).

[5] Sakmann, B. *Single-channel recording* (Springer Science & Business Media, 2013).

[6] Neher, E. & Sakmann, B. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* **260**, 799–802 (1976).

[7] Colquhoun, D., Hawkes, G. A. & Bernard, K. On the stochastic properties of single ion channels. *P. of the Roy. Soc. of London. Series B. Biological Sciences* **211**, 205–235 (1981).

[8] Horn, R. & Lange, K. Estimating kinetic constants from single channel data. *Biophys J.* **43**, 207–223 (1983).

[9] Epstein, M., Calderhead, B., Girolami, M. A. & Sivilotti, L. G. Bayesian Statistical Inference in Ion-Channel Models with Exact Missed Event Correction. *Biophys J.* **111**, 333–348 (2016).

[10] Siekmann, I., Fackrell, M., Crampin, E. J. & Taylor, P. Modelling modal gating of ion channels with hierarchical markov models. *Proc. of the Roy. Soc. A- Math. Phys.* **472**, 20160122 (2016).

[11] Chung, S.-H., Moore, J. B., Xia, L., Premkumar, L. & Gage, P. W. Characterization of single channel currents using digital signal processing techniques based on hidden markov models. *Philos. T. of the Roy. Soc. of Lond. Series B Bio. Sci.* **329**, 265–285 (1990).

[12] Fredkin, D. R. & Rice, J. A. Maximum likelihood estimation and identification directly from single-channel recordings. *P. of the Roy. Soc. of London. Series B: Biological Sciences* **249**, 125–132 (1992).

[13] Qin, F., Auerbach, A. & Sachs, F. Hidden Markov Modeling for Single Channel Kinetics with Filtering and Correlated Noise. *Biophys J.* **79**, 1928–1944 (2000).

[14] Venkataramanan, L. & Sigworth, F. Applying hidden Markov models to the analysis of single ion channel activity. *Biophys J.* **82**, 1930–1942 (2002).

[15] Ball F. G., C. Y. K. J. B. & 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).

[16] Rosales, R., Stark, J. A., Fitzgerald, W. J. & Hladky, S. B. Bayesian Restoration of Ion Channel Records using Hidden Markov Models. *Biophys J.* **80**, 1088–1103 (2001).

[17] Rosales, R. A. MCMC for hidden Markov models incorporating aggregation of states and filtering. *Bulletin of Mathematical Biology* **66**, 1173–1199 (2004).

[18] Gin, E., Falcke, M., Wagner, L. E., Yule, D. I. & Sneyd, J. Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors. *J. of Theoretical Biology* **257**, 460–474 (2009).

[19] Siekmann, I. *et al.* MCMC Estimation of Markov Models for Ion Channels. *Biophys J.* **100**, 1919–1929 (2011).

[20] Siekmann, I., Sneyd, J. & Crampin, E. J. MCMC Can Detect Nonidentifiable Models. *Biophys J.* **103**, 2275–2286 (2012).

[21] Hines, K. E., Bankston, J. R. & Aldrich, R. W. Analyzing Single-Molecule Time Series via Nonparametric Bayesian Inference. *Biophys J.* **108**, 540–556 (2015).

[22] Zheng, J. & Zagotta, W. N. Gating rearrangements in cyclic nucleotide-gated channels revealed by patch-clamp fluorometry. *Neuron* **28**, 369–374 (2000).

[23] Taraska, J. W. & Zagotta, W. N. Structural dynamics in the gating ring of cyclic nucleotide–gated ion channels. *Nature structural & molecular biology* **14**, 854 (2007).

[24] Taraska, J. W., Puljung, M. C., Olivier, N. B., Flynn, G. E. & Zagotta, W. N. Mapping the structure and conformational movements of proteins with transition metal ion FRET. *Nature methods* **6**, 532 (2009).

[25] Bruening-Wright, A., Elinder, F. & Larsson, H. P. Kinetic Relationship between the Voltage Sensor and the Activation Gate in spHCN Channels. *The Journal of General Physiology* **130**, 71–81 (2007).

[26] Kalstrup, T. & Blunck, R. Dynamics of internal pore opening in kv channels probed by a fluorescent unnatural amino acid. *Proc. of the Nat. Academy of Sci.* **110**, 8272–8277 (2013).

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

[28] Wulf, M. & Pless, S. A. High-sensitivity fluorometry to resolve ion channel conformational dynamics. *Cell reports* **22**, 1615–1626 (2018).

[29] Biskup, C. *et al.* Relating ligand binding to activation gating in cnga2 channels. *Nature* **446**, 440 (2007).

[30] Kusch, J. *et al.* Interdependence of Receptor Activation and Ligand Binding in HCN2 Pacemaker Channels. *Neuron* **67**, 75–85 (2010).

[31] Kusch, J. *et al.* How subunits cooperate in cAMP-induced activation of homotetrameric HCN2 channels. *Nature Chemical Biology* **8**, 162 (2011).

[32] Biskup, C. *et al.* Relating ligand binding to activation gating in CNGA2 channels. *Nature* **446**, 440 (2007).

[33] Wu, S. *et al.* State-Dependent cAMP Binding to Functioning HCN Channels Studied by Patch-Clamp Fluorometry. *Biophys J.* **100**, 1226–1232 (2011).

[34] Colquhoun, D., Hawkes, G. A. & Bernard, K. Relaxation and fluctuations of membrane currents that flow through drug-operated channels. *Proc. R. Soc. Lond. B* (1997).

[35] Alcantara, P., Cardenas, L. M., Swillens, S. & Scroggs, R. S. Reduced Transition between Open and Inactivated Channel States Underlies 5HT Increased INa+ in Rat Nociceptors. *Biophys J.* **83**, 5–21 (2002).

[36] Wang, W. *et al.* Optimal Estimation of Ion-Channel Kinetics from Macroscopic Currents. *PLOS ONE* **7**, 1–12 (2012).

[37] Kurtz, T. G. The relationship between stochastic and deterministic models for chemical reactions. *The J. of Chem. Phys.* **57**, 2976–2978 (1972).

[38] Van Kampen, N. G. *Stochastic processes in physics and chemistry*, vol. 1 (Elsevier, 1992).

[39] Jahnke, T. & Huisinga, W. Solving the chemical master equation for monomolecular reaction systems analytically. *J. of Mathematical Biology* **54**, 1–26 (2007).

[40] Milescu, L. S., Akk, G. & Sachs, F. Maximum Likelihood Estimation of Ion Channel Kinetics from Macroscopic Currents. *Biophys J.* **88**, 2494–2515 (2005).

[41] Munsky, B., Trinh, B. & Khammash, M. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology* **5**, 318 (2009).

[42] Celentano, J. J. & Hawkes, A. G. Use of the Covariance Matrix in Directly Fitting Kinetic Parameters: Application to GABAA Receptors. *Biophys J.* **87**, 276–294 (2004).

[43] Stepanyuk, A. R. & Borisyuk, P. V., Anya L.and Belan. Efficient maximum likelihood estimation of kinetic rate constants from macroscopic currents. *Plos one* **6**, 1–18 (2011).

[44] Stepanyuk, A., Borisyuk, A. & Belan, P. Maximum likelihood estimation of biophysical parameters of synaptic receptors from macroscopic currents. *Frontiers in Cellular Neuroscience* **8**, 303 (2014).

[45] Jahnke, T. & Huisinga, W. Solving the chemical master equation for monomolecular reaction systems analytically. *J. of mathematical biology* **54**, 1–26 (2007).

[46] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE* **77**, 257–286 (1989).

[47] Moffatt, L. Estimation of Ion Channel Kinetics from Fluctuations of Macroscopic Currents. *Biophys J.* **93**, 74–91 (2007).

[48] Kalman, R. E. A new approach to linear filtering and prediction problems. *J. of basic Engineering* **82**, 35–45 (1960).

[49] Komorowski, M., Finkenstädt, B., Harper, C. V. & Rand, D. A. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**, 343–343 (2009).

[50] Finkenstädt, B. *et al.* Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. *The Annals of Applied Statistics* **7**, 1960–1982 (2013).

[51] Fearnhead, P., Giagos, V. & Sherlock, C. Inference for reaction networks using the linear noise approximation. *Biometrics* **70**, 457–466 (2014).

[52] Folia, M. M. & Rattray, M. Trajectory inference and parameter estimation in stochastic models with temporally aggregated data. *Statistics and Computing* **28**, 1053–1072 (2018).

[53] Gillespie, C. S. & Golightly, A. Bayesian inference for the chemical master equation using approximate models. In *Ninth International Workshop on Computational Systems Biology, WCSB 2012, June 4-6, Ulm, Germany*, vol. 4, 23 (2012).

[54] Hines, K. E. A Primer on Bayesian Inference for Biophysical Systems. *Biophys J.* **108**, 2103–2113 (2015).

[55] Ball, F. MCMC for Ion-Channel Sojourn-Time Data: A Good Proposal. *Biophys J.* **111**, 267–268 (2016).

[56] Auger-Methe Marie *et al.* State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. *Scientific Reports* **6**, 26677 (2016).

[57] Hines, K. E., Middendorf, T. R. & Aldrich, R. W. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *The J. of General Physiology* **143**, 401–416 (2014).

[58] Middendorf, T. R. & Aldrich, R. W. The structure of binding curves and practical identifiability of equilibrium ligand-binding parameters. *J. of General Physiology* **149**, 121–147 (2017).

[59] Calderhead, B., Epstein, M., Sivilotti, L. & Girolami, M. *Bayesian Approaches for Mechanistic Ion Channel Modeling*, 247–272 (Humana Press, Totowa, NJ, 2013).

[60] Watanabe, S. Almost All Learning Machines are Singular. In *2007 IEEE Symposium on Foundations of Computational Intelligence*, 383–388 (2007).

[61] Kienker, P. Equivalence of aggregated markov models of ion-channel gating. *P. of the Roy. Soc. of London. B. Biological Sciences* **236**, 269–309 (1989).

[62] Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24**, 997–1016 (2014).

[63] Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. of machine learning research* (2010).

[64] Gillespie Daniel T. Exact stochastic simulation of coupled chemical reactions. *The J. of Phys. Chem.* **81**, 2340–2361 (1977).

[65] Ghahramani, Z. Learning dynamic bayesian networks. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, 168–197 (Springer, 1997).

[66] Chen, Z. *et al.* Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics* **182**, 1–69 (2003).

[67] Anderson, B. D. & Moore, J. B. *Optimal filtering* (Courier Corporation, 2012).

[68] Weiss, N. A. *A Course in Probability*, 385–386 (Addison-Wesley, Boston, U.S.A., 2005).

[69] Hoffman, M. D. & Gelman, A. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. of Machine Learning Research* **15**, 1593–1623 (2014).

[70] Betancourt, M. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434* (2017).

[71] Gelman, A., Lee, D. & Guo, J. Stan: A probabilistic programming language for bayesian inference and optimization. *J. of Educational and Behavioral Statistics* **40**, 530–543 (2015).

[72] Smoluchowski, M. v. Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. *Zeitschrift für physikalische Chemie* **92**, 129–168 (1918).

[73] van Holde, K. A hypothesis concerning diffusion-limited protein–ligand interactions. *Biophys. Chem* **101**, 249–254 (2002).

[74] Peuker, S. *et al.* Kinetics of ligand-receptor interaction reveals an induced-fit mode of binding in a cyclic nucleotide-activated protein. *Biophys J.* **104**, 63–74 (2013).

[75] Jaynes, E. T. Information theory and statistical mechanics. *Phys. rev.* **106**, 620 (1957).

[76] van der Vaart, A. Bayes procedures. In *Asymptotic Statistics*, 138–152 (Cambridge University Press, 1998).

[77] McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan* (Chapman and Hall/CRC, 2018).

[78] Gelman, A., Simpson, D. & Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy* **19**, 555 (2017).

[79] Zwickl, D. J. & Holder, M. T. Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics. *Systematic Biology* **53**, 877–888 (2004).

[80] Vehtari, A., Ojanen, J. *et al.* A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6**, 142–228 (2012).

[81] Piironen, J. & Vehtari, A. Comparison of bayesian predictive methods for model selection. *Statistics and Computing* **27**, 711–735 (2017).

[82] Wallace, C. *Statistical and Inductive Inference by Minimum Message Length.* Information Science and Statistics (Springer-Verlag, Berlin, Germany, 2005).

[83] Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez Jr, R. L. & Wiggins, C. H. Learning rates and states from biophysical time series: a bayesian approach to model selection and single-molecule FRET data. *Biophys J.* **97**, 3196–3205 (2009).

[84] Vehtari, A. & Ojanen, J. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.* **6**, 142–228 (2012).

[85] Burnham, K. P. & Anderson, D. R. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research* **33**, 261–304 (2004).

[86] Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **22**, 79–86 (1951).

[87] Noé, F., Wu, H., Prinz, J.-H. & Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *The J. of Chem. Phys.* **139**, 184114 (2013).

[88] Akaike, H. *A New Look at the Statistical Model Identification*, 215–222 (Springer New York, New York, NY, 1998).

[89] Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *J. of the Roy. Statistical Soc.: Series B (Methodological)* **39**, 44–47 (1977).