Bayesian Hidden Markov modeling and model selection by Kalman filtering applied to multi-dimensional data of ion channels

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 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 epilepsia[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[5,6]. The gating of ion channels is usually interpreted by kinetic schemes which are inferred by macroscopic currents with rate equations (REs)[5,6] and single-channel currents with dwell time distributions[7–11] or hidden Markov models (HMMs)[12–15]. Recently, the advantage of Bayesian statistics has been studied for HMMs[16–22]. In ensemble patches, simultaneous orthogonal measurement of either conformational changes[23–29] or ligand binding itself[30–34] has increased insight into the complexity of the activation mechanism in channels.

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 yields a point estimate of the rate matrix or its eigenvalues[5,6,35–37]. The RE approach is based on a deterministic differential equation derived by averaging the chemical master equation (CME) for the underlying kinetic scheme[38–40]. Its accuracy can be improved by processing the information of the intrinsic noise[41,42]. Nevertheless, all deterministic approaches ignore to some degree the correlations of the intrinsic noise. 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. Rigorous attempts to incorporate the intrinsic noise of current data into the estimation[43] suffer from cubic computational complexity in the amount of data points, rendering the algorithm impractical for real data. Stepanyuk suggested a faster algorithm[44,45]. Realistic approaches to analyze single-molecule data solve the stochastic CME[46] such as HMMs[14,15]. 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[47]. HMM approaches are more accurate than fitting dwell time distributions for noisy recordings of rapidly gating channels[15] but the computational complexity limits this type of analysis in ensemble patches to no more than a few hundred channels per time trace[48].

To tame the computational complexity[46], we approximate the solution of the CME with a Kalman filter (KF), thereby remaining in a stochastic framework[49]. 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[50–53]. Our approach generalizes of Moffat[48] by including state-dependent fluctuations such as open-channel noise and Poisson noise in additional fluorescence data. For cPCF data our model takes into account the non-Poissonian elements of the noise which occur in the difference signal between bound ligand fluorescence and the bulk background signal.

Stochastic rather than deterministic modeling is gener-

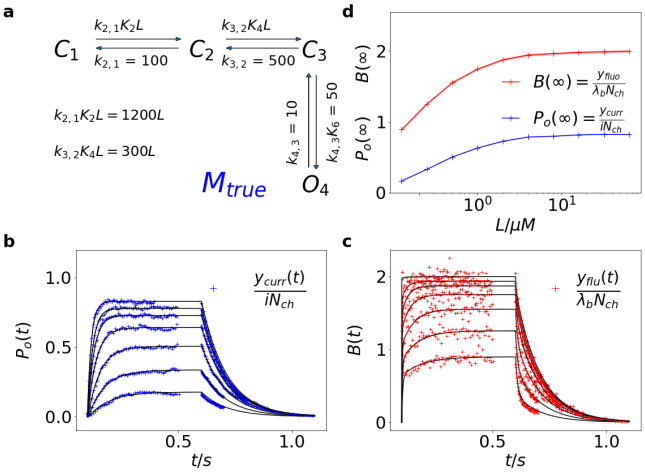

Figure 1: **HMM used for the simulations.** The HMM consists of two binding steps and one opening step. **a**, The rate matrix $\mathbf{K}$ is parametrized by the absolute rates $k_{i,j}$, the ratios K_i between on and off rates (i.e. equilibrium constants) and L, the ligand concentration in the solution. The units of the rates are $\frac{1}{s}$ and $\frac{1}{\mu M s}$ respectively. In order to make it more challenging for the algorithm, we did not assume any prior knowledge on stoichiometry among rates. The open state O_4 conducts a mean single-channel current $i = 1$. The liganded states are C_2, C_3. **b-c**, Normalized time traces of relaxation experiments with $N_{\mathrm{ch}} = 10^3$ channels, $\lambda_b = 0.375$ mean photons per bound ligand per frame and single-channel current $i = 1$. Simultaneous current y_{curr} and fluorescence y_{flu} time courses are simulated . 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 $\sigma_m^2 = i^2$ and open-channel noise $\sigma_{\mathrm{op}}^2 = 0.1i^2$. The fluctuation of fluorescence is caused by stochastic binding and Poisson counting noise of photons. The black lines are the theoretical open probabilities $P_{\mathrm{o}}(t)$ and the binding degree $B(t)$ for $N_{\mathrm{ch}} \to \infty$ of the used model. The ligand concentrations are $0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 64\,\mu\mathrm{M}$ **d**, Equilibrium binding and open probability as function of the ligand concentration L

ally preferable for small systems or non-linear dynamics[39,54]. 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[48] already demonstrated the advantage of the KF to learn absolute rates from time traces at equilibrium. Other benefits are the ability to infer the number of channels N_{ch} for each time trace, the single-channel conductance i and the mean number λ_{b} of photons from bound ligands per recorded frame. Thus there is no need for error-prone normalizations of the signal typical for deterministic (i.e. averaging) approaches. The KF obtains a likelihood which makes it possible to fit the data globally with every other data type whose model also delivers a likelihood for the respective data.
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[55,56]. In contrast, previous work on ensemble currents combined the KF only with maximum likelihood (ML) estimation[48] and did not derive model selection strategies. Difficulties in treating simple stochastic models by ML approaches in combination with the KF[57], especially with non-observable dynamics, justify the computational burden of Bayesian statistics. Bayesian inference provides outmatching tools for modeling: First, information from other experiments 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[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]. In particular, we use the recently developed Widely Applicable Information Criterion (WAIC)[63], which relies 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 confocal patch-clamp fluorometry (cPCF).

Results and Discussion

Simulation of relaxing cPCF data

As an exemplary HMM (Supplementary Sec. 2 and Methods) we assume a ligand-gated channel (Fig. 1a). The state evolution $s(t)$ of each channel in the patch was sampled[64] and then 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). We will drop this assumption later (Figs. 4-6) (see Supplementary material and Methods section). Example data are shown in Fig. 1b-d.

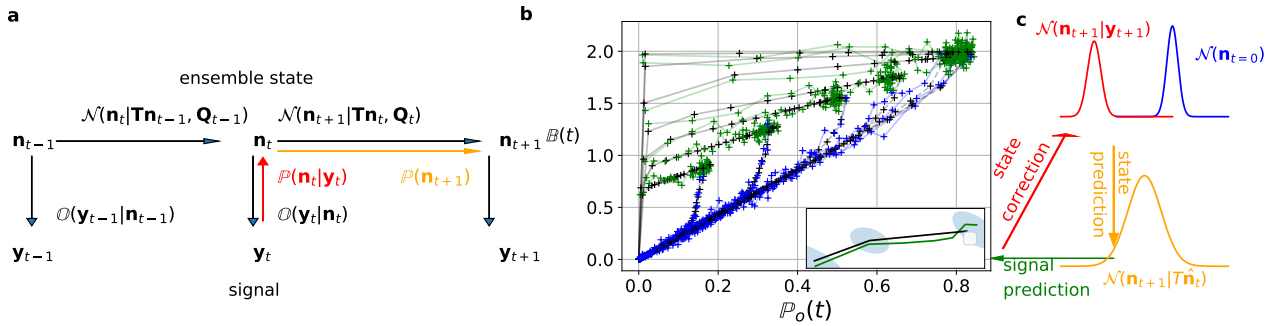

Figure 2: **KF as Bayesian network.** **a,** Graphical model of the conditional dependencies of the stochastic process. Horizontal black arrows represent the conditional multivariate normal transition probability $\mathcal{N}(\mathbf{n}_{t+1}|\mathbf{T}\mathbf{n}_t, \mathbf{Q}_t)$ of a continuous state 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 λ_{B}, i and N_{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 (blue multivariate normal distribution) the analysis cycle predicts a multivariate normal distribution for the ensemble state (orange) for the next time point, and a likelihood for the signal (green) given in the denominator of Eq. 3. Subsequently, the algorithm corrects the state distribution with the received data (red). In this moment, the state distribution reflects what is known about the channel ensemble state. The predicted distribution of the ensemble reflects what could be $\mathbf{n}(t+1)$, given a specific parameter set and what we know about $\mathbf{n}(t)$. Then the algorithm repeats the cycle for the next data point.

Kalman filter derived from a Bayesian network

The KF is a Bayesian network, i.e. a continuous state HMM[65] (Fig. 2a). 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 $\mathbf{T}$

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

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 $\mathrm{cov}(\boldsymbol{\omega}) = \mathbf{Q}(\mathbf{T}, \mathbf{n}_{t-1})$ (Methods Eq. 30). The observations $\mathbf{y}(t)$ depend linearly on the hidden state $\mathbf{n}(t)$. The linear map is determined by an observation matrix $\mathbf{H}$.

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

The noise of the measurement setup (Supplementary sec. 3, 4 and Eq. 36) 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 $\mathbb{O}$.
For each element in a sequence of hidden states $\{\mathbf{n}_t : 0 < t < T\}$ and for a fixed set of parameters $\boldsymbol{\theta}$, the KF recursively predicts (prior) distributions $\mathbb{P}(\mathbf{n}_t)$(Eq. 26 and 30). Each prediction is followed by a correction step,

$$\tilde{\mathbb{P}}(\mathbf{n}_t|\mathbf{y}_t, \boldsymbol{\theta}) = \frac{\mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\boldsymbol{\theta})}{\int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\boldsymbol{\theta})\,\mathrm{d}\mathbf{n}_t}, \tag{3}$$

that allows to incorporate the current data point into the estimate[66] and which is based on the Bayesian theorem[48,67](Supplementary Sec. 5). Due to the recursiveness of its equations, the KF has a time complexity that is linear in the number of data points, allowing for a fast computer implementation. The denominator of Eq. 3 is the normal distributed marginal likelihood for each data

point, which constructs by

$$\mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) = \prod_{t=2}^{N_T} \mathbb{L}(\mathbf{y}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta}) \tag{4a}$$

$$= \prod_{t=2}^{N_T} \int \mathbb{O}(\mathbf{y}_t|\mathbf{n}_t)\mathbb{P}(\mathbf{n}_t|\mathcal{Y}_{t-1}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{n}_t \tag{4b}$$

$$= \prod_{t=2}^{N_T} \mathcal{N}(\mathbf{y}_t|\mathbf{H}\mathbb{E}(\mathbf{n}_t), \mathbf{H}\mathbf{P}_t\mathbf{H}^\top + \boldsymbol{\Sigma}_t), \tag{4c}$$

a product likelihood of normal distributions of the whole time trace $\mathcal{Y}_\mathrm{T} = \{\mathbf{y}_1, ..., \mathbf{y}_{N_\mathrm{T}}\}$ of length N_T for the KF. For the derivation of $\mathbf{P}_t$ and $\boldsymbol{\Sigma}_t$ see Methods Eq. 30 and Eq. 36. The likelihood for the data allows to ascribe probability to the parameters $\boldsymbol{\theta}$, given the observed data (Methods Eq. 10). An illustration for the operation of the KF on the observation space (Fig. 2**b**) and ensemble states space (Fig. 2**c**) is shown. We plot the mean signal $\mathbf{H}\mathbb{E}[\mathbf{n}(t)]$ and the data as vector trajectories.
For Poisson-distributed photon counting or open-channel noise Eq. 3 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.48). The Bayesian posterior distribution

$$\mathbb{P}(\theta|\mathcal{Y}_T) \sim \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}) \tag{5}$$

encodes all information from model assumptions and experimental data used during model training (Supplementary Sec. 1). The 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

We simulate synthetic time traces as described in Methods material. 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**). For typical experimental settings the total parameter error of the RE estimates[41,48], calculated as the Euclidean distance to the true values, is roughly 10 times larger than the corresponding parameter error of the KF (Fig. 3**c**). It is noteworthy that, even for 10^4 channels per patch, the precision of the KF algorithm is 4 times higher than that of the RE model on that data set. Dividing the error of all estimates from REs 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 would have been achieved with REs only if increasing the number of channels by tenfold. 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[48]. For small N_ch the ratio of the errors decreases like $\sim 1/\sqrt{N_\mathrm{ch}}$, (Fig. 3**c**), thus the RE approach scales $1/N_\mathrm{ch}$ 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 of the time traces are shown, that are defined as

$$r_i := \frac{y_i - \mathbf{H}\mathbb{E}(\mathbf{n}_i)}{\sqrt{\mathrm{var}(y_i)}}. \tag{6}$$

We normalize with respect to the standard deviation of the prediction $\sqrt{\mathrm{var}(y_i)}$ for each data point predicted 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 which means that the residuals are indeed a white-noise process. Estimating the residuals from RE would lead to correlated residuals[48], which is one reason for the biased parameter estimates.

cPCF versus patch clamp only

To evaluate the advantage of cPCF data[30] with respect to PC data only (Fig. 4), we compare different types of ligands: Idealized ligands with brightness λ_b, emitting light only when bound to the channels, and 'real' ligands which also produce background fluorescence when diffusing in the bath solution (Supplement Sec. 3). The increase in precision for the dissociation rate of the first ligand, $k_{2,1}$, is that strong that the variance of the posterior $\mathbb{P}(k_{2,1}, k_{3,2})$ nearly vanishes in the combined plot with the current data (nearly all probability mass is concentrated in a single point in Fig. 4**a**). The effect on the error of the equilibrium constants K_i is less strong. Additionally, the bias is reduced and even the estimation of N_ch is improved. The brighter the ligands are, the more $\mathbb{P}(k_{2,1}, k_{3,2})$ decorrelates. The influence of prior information such as the assumption of diffusion limited binding is investigated in (Supplementary Sec. 1.1) 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 λ, $k_{2,1}$ becomes better estimated than $k_{3,2}$ (Fig. 4**c**). The posterior of the binding and unbinding rates of the first ligand contracts with increasing λ_b. The l_{95} percentiles of other parameters exhibit a weaker dependency on the brightness ($l_{95} \sim \lambda^{-0.1}$). For $\lambda_\mathrm{b} = 0.01$ photons per bound ligand and frame, which corresponds to a maximum mean signal of 20 photons per frame, the normal approximation to the Poisson noise hardly captures the asymmetry of photon counting noise included in the time traces. Nevertheless, l_{95} decreases

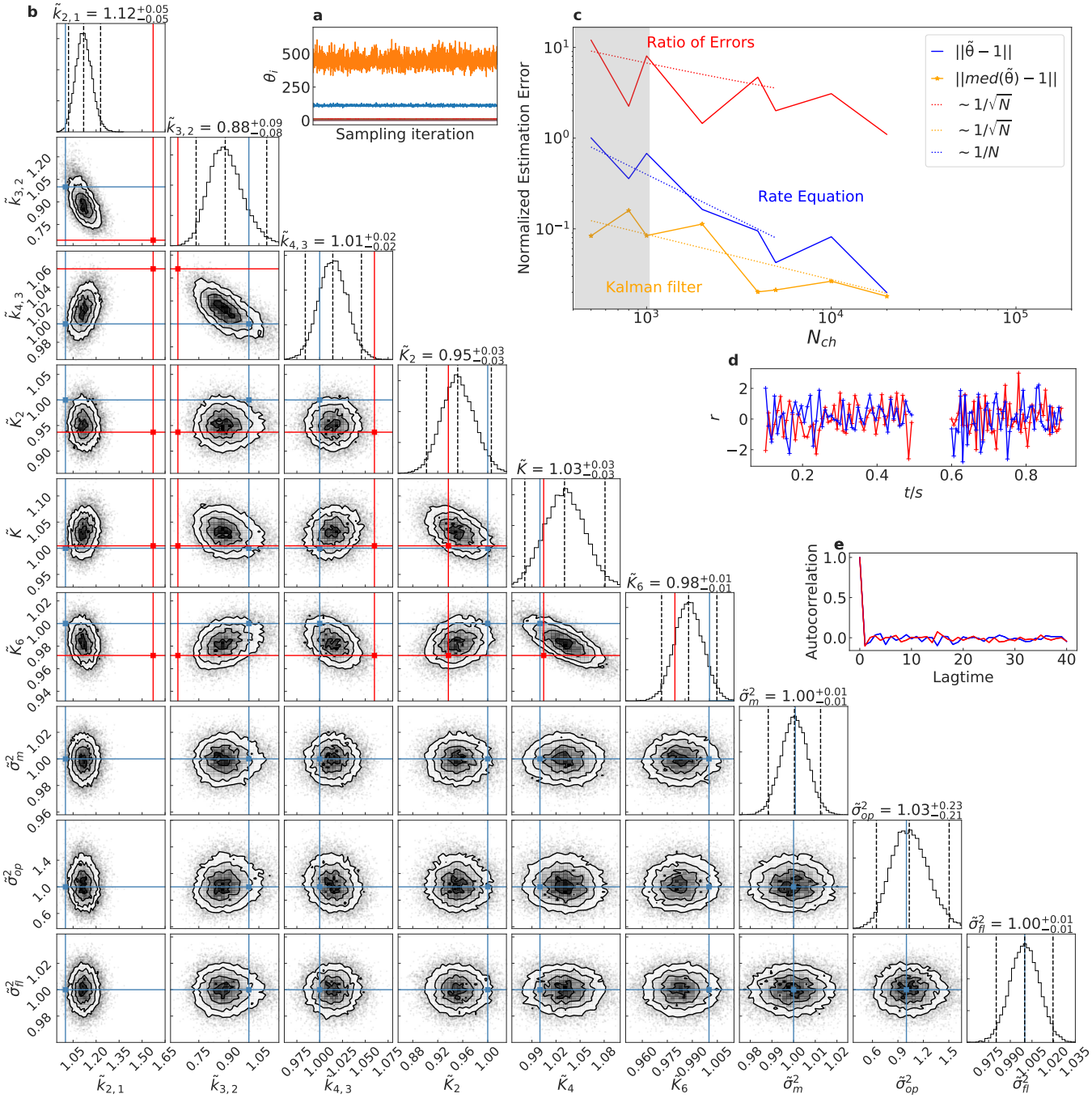

Figure 3: **Benchmark of the KF against least squares fitting of REs. a,** Sample traces of 3 representative parameters from the posterior distribution of the KF algorithm created by Hamiltonian Monte Carlo sampling. The posterior is constructed by those samples. **b,** Posterior distribution plotted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

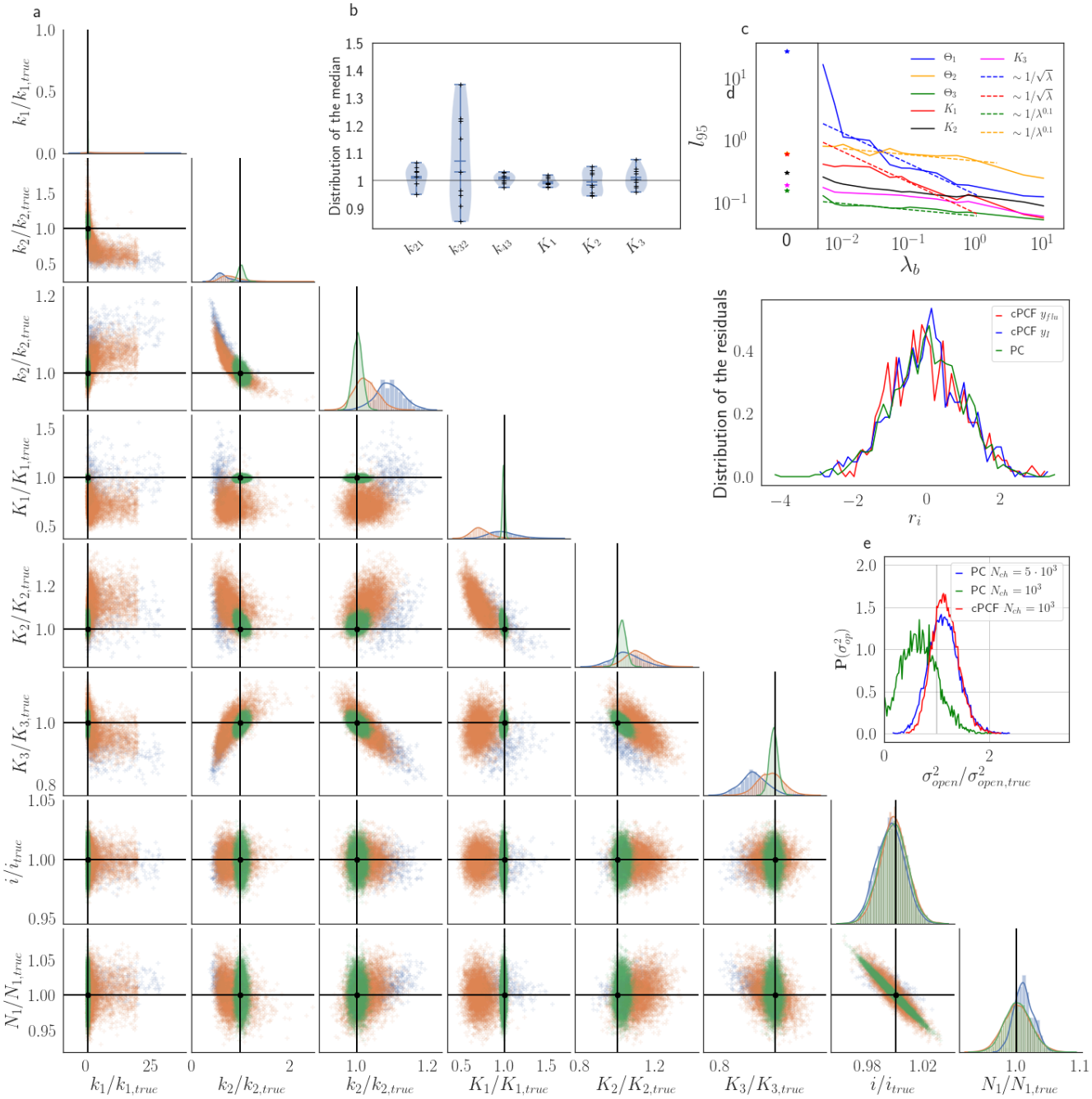

Figure 4: **Benchmark of the KF for patch-clamp versus patch-clamp fluorometry data.** **a**, Posteriors of PC data (blue), cPCF data with $\lambda_\mathrm{b} = 0.00375$ (orange) and cPCF data with $\lambda_\mathrm{b} = 0.375$ (green) but with superimposing bulk signal. The black lines represent the 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. λ_b normalised by the true value of each parameter. A regime with $l_{95} \sim 1/\sqrt{\lambda}$ is shown for $k_{2,1}$ and K_1, while other parameters show a weaker dependency on the ligand brightness. **d**, Histograms of the residuals r of cPCF with $\lambda_\mathrm{b} = 2.5 \cdot 10^{-3}$ 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 σ^2_op. We assumed as prior for the instrumental variance $\mathbb{P}(\sigma^2) = \mathcal{N}(1, 0.01)$.

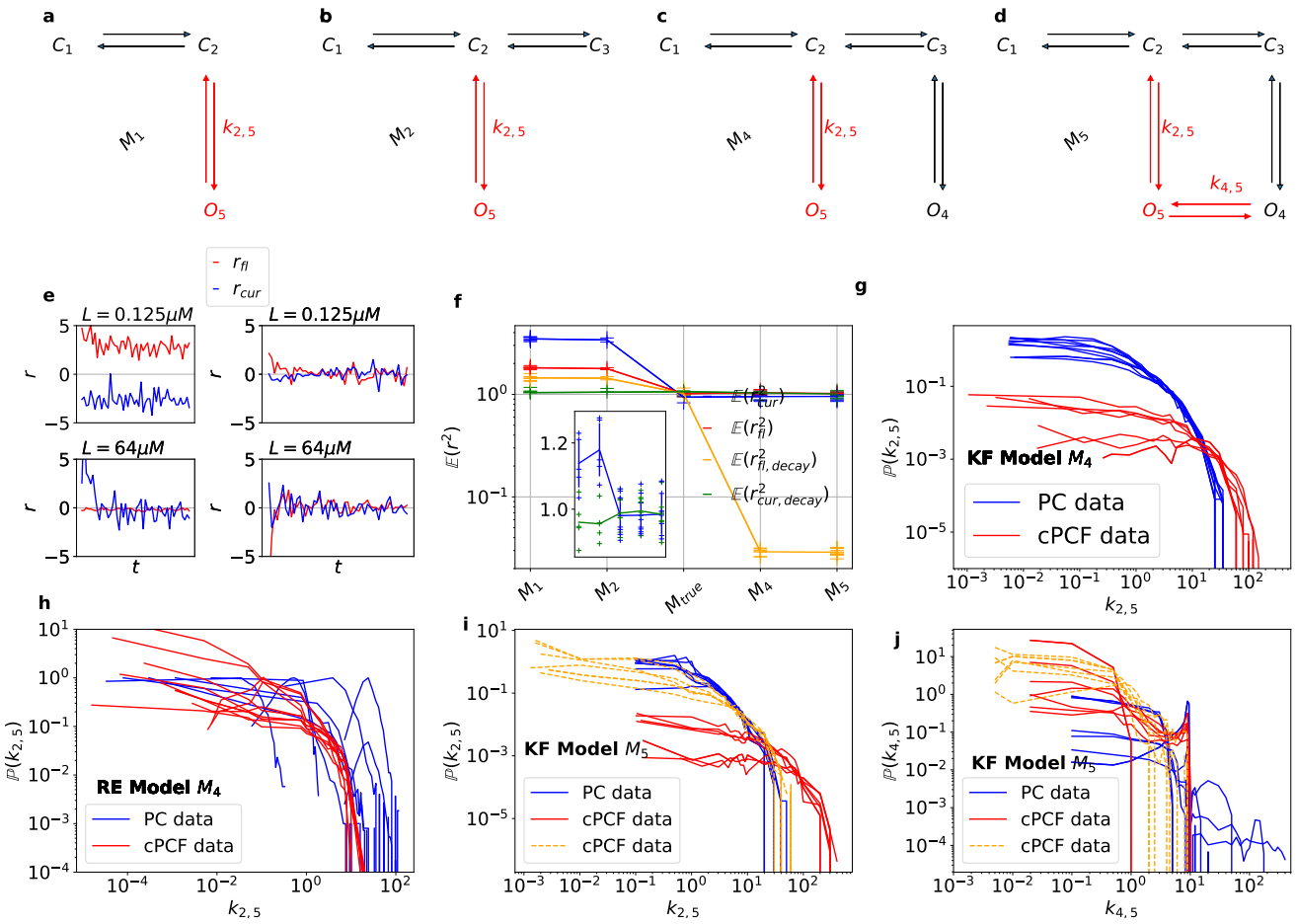

Figure 5: **Model selection by the second moment of the residuals and by continuous model expansion. a-d**, Model structures of the trained models differing from the true process (c.f. Fig. 1a). Red states and transitions have no counterpart in the true model. All models are nested in the most complex 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 $\mathbb{E}[r_i] = 0$. All variance is explained by the model. The overfitting models M_4 and M_5 can be easily identified by the decaying fluorescence time trace. The inset shows the second non-centralized moment of PC data for the five different models on a linear scale. While underfitting can be clearly detected, overfitting is not detected. cPCF Data increase the contrast of current residuals between underfitting and true process by ≈ 12.5. **g-j**, Continuous model expansion by model M_4 and M_5 from **c,d**. **g**, Posterior distribution $\mathbb{P}(k_{2,5})$ for a rate into a state which does not exist in the true process obtained by the KF. For current data only, 10 out of 10 data sets maximise the probability density for $k_{2,5} \to 0$. For cPCF data, 9 out 10 data sets yielded the same result and only one data set has a finite maximum. Hence, the KF suggests to delete O_5. **h**, Same current data (blue) analyzed with a RE approach finds in 4 out of 7 data sets rates into the non-existing state. Adding fluorescence data (red) improves the analysis. Now in 8 out of 10 data sets the posterior is maximized for $k_{2,5} = 0$. **i,j**. The KF for various PC and cPCF data sets reduces M_5 to the true data-generating process M_{true} for most data sets by maximizing the posterior for a zero rate into O_5. For current only data 4 out of 7 data sets show this behavior. For cPCF 9 out 10 data sets maximise $\mathbb{P}(k_{2,5})$ for $k_{2,5} \to 0$ and $\mathbb{P}(k_{4,5})$ for $k_{4,5} \to 0$. Across M_4 and M_5 (**g** and **i**) the posteriors $\mathbb{P}(k_{2,5})$ show a similar shape. **j**, Notably, various cases of multimodality occur for PC data. cPCF and PC data share the tendency even though completely different experiments to create a finite peak around $k_{4,5} = 10$. 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 half- normal prior $k_{5,2} \sim \text{halfnormal}(0, 6 \cdot 10^3)$ for one specific rate which we identified to be not confined by the data. We therefore restricted this rate to a reasonable values given the data sampling rate of 10kHz (see Supplementary Sec. 1.2).This prior also reduces the magnitude of the peak in $k_{4,5}$ and thus suggest omitting state O2. There are certainly more informative but still weak priors possible.

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 (Fig. 5 **g** and Fig. 6 **b**) that for similar data sets the true underlying process can still be determined. Remarkably, the KF allows to determine in a macropatch the variance of the open-channel current noise for $\sigma_{\text{op}} = 0.1i$, i.e. when the total noise is dominated by the much larger gating noise $(HP_tH^\top)_{2,2}$ (Fig. 4**e**): For $p_{o,max} = 0.8$, the expected open probability of the true process, the ratio is

$$\frac{(HP_tH^\top)_{2,2}}{\sigma^2_{\text{op}}} \approx \frac{1 - p_{o,max}}{\sigma^2_{\text{op}}} \geq 20 \tag{7}$$

under equilibrium assumption.

Bayesian Model selection via 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 $M_1, M_2, M_{\text{true}}, M_4,\ M_5$ (Fig. 5**a-d**).

The KF enables to identify underfitting, better by plotting the residuals $r(t)$ (Fig. 5**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. 5**f**). For M_{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. Overfitting models 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 continuous model expansion perspective nests all simpler models in the supermodel M_5. 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. 5**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. 5**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. 5**i-j**). For PC data, the posterior reveals multi-modality with some data sets (Fig. 5**l**). Hence, point estimates of the parameters are not reliable while the posterior of M_5 encodes all model uncertainties. A weakly informative half-normal prior which penalizes unrealistic high rates of $k_{5,2}$ suppresses secondary peaks and probability mass in the distribution tails for finite rates $k_{2,5}$ and $k_{4,5}$ further emphasizing to leave out O_5 in the final model (Fig. **i-j** and Supplementary Sec. 1). For ambiguous situations with either multi-modality in the posterior or no maximum in the probability density at $k_{i,j} = 0$, more elaborate prior distributions[54,72,73] or methods are required.

Generally, when selecting a kinetic scheme for a protein, one reduces high-dimensional continuous dynamics to a few discrete states and transition rates. The true data-generating process $\mathbb{F}(y)$ is therefore not included in the set of models from which the best fitting model is chosen[74]. Therefore, one can only attempt to maximize the mean predictive accuracy. 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

$$\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}}) = \log \mathbb{E}_\theta[\mathbb{L}(\tilde{\mathcal{Y}}|\boldsymbol{\theta})] \tag{8}$$

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

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})] = \int \log \mathbb{P}_{\text{pred}}(\tilde{\mathcal{Y}})\mathbb{F}(\tilde{\mathcal{Y}}) \prod_{i=1}^{N_{\text{data}}} \mathrm{d}\tilde{y}_i \tag{9}$$

Maximizing Eq. 9 is equivalent to minimizing the Kullback-Leibler distance which specifies the information loss when approximating the unknown $\mathbb{F}(y)$ by a model[75,76]. In this way one can rank the candidate models by the information loss about $\mathbb{F}(y)$[62]. Because $\mathbb{F}$ is unknown, one has to estimate Eq. 9 by using the available samples from $\mathbb{F}(y)$. Therefore, different estimators of Eq. 9 are compared: AIC, WAIC, maximum *a posteriori* cross-validation and Bayesian cross-validation (see, Methods). As a result, if current time traces are modeled with REs, both information criteria fail to detect the overfitting model (Fig. 6**a**). RE approaches tend to infer more states than present in the true model if judged by the training data. One can exploit this weakness by choosing the kinetic scheme on cross-validated data, since overfitting models derived from RE do not generalise as good to new data as a parsimonious model. As expected from theory, WAIC[63] succeeds while AIC fails to estimate the predictive accuracy (Fig. 6**b**). The failing AIC correlates with the occurrence of a non-normal posterior distribution[60](see, Methods).

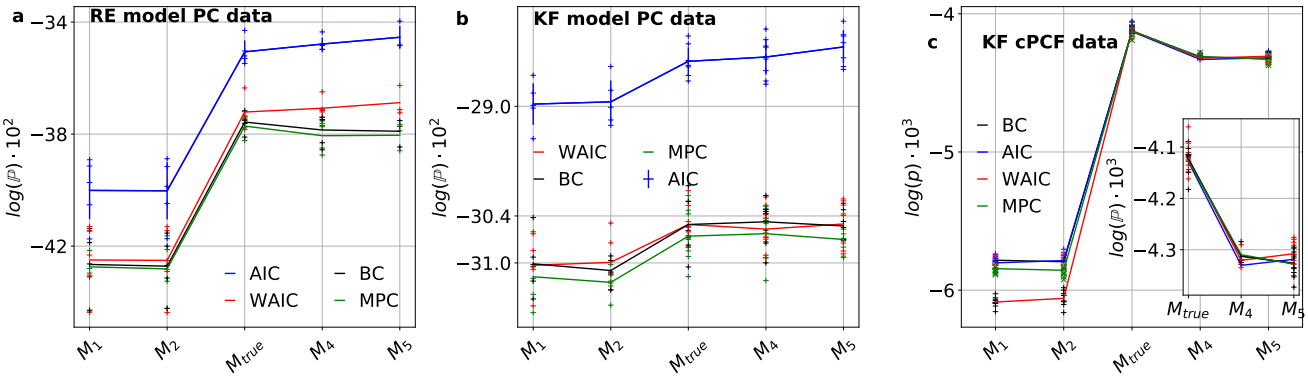

Figure 6: **Bayesian Model selection** Compared is the predictive accuracy of the indicated five models estimated by either the Akaike Information Criterion (AIC), the Widely Applicable Information Criterion (WAIC), Bayesian cross-validation (BC) or Maximum *a posteriori* cross-validation (MPC).We use 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, thereby exploiting the weakness of the RE approach to infer non-existing states on the training data. The additional state does not contribute positively to the predictive accuracy of the model. The kinetic schemes M_4 and M_5 do not gerneralize well to new data if they are fitted with REs. **b**, For current data analyzed with the KF, WAIC estimates the predictive accuracy obtained by Bayesian cross-validation (black) with high accuracy. WAIC predicts the BC value even better than MPC. As expected from theory AIC (blue) fails in the case of a non-normal posterior distribution to approximate the predictive accuracy. It even suggests a better predictive accuracy with complexer models. The mean predictive accuracy (black and red) for M_4 and M_5 is only slightly smaller than that for M_{true} which can be explained by the observation that the KF automatically finds a sharp peak at $\mathbb{P}(k_{2,5} = 0)$ (Fig. 5 g,i,j; see Methods). This can be used as a model selection strategy: If two models have a similar predictive accuracy and are nested one should check whether the posterior of the larger model switches off certain rates. The predictive accuracy not only scores a kinetic scheme. It also evaluates how closely the whole algorithm mimics the true process. Comparing the predictive accuracy of the true kinetic scheme M_{true} in **a** and **b** reveals the higher precision of the KF in modeling the stochastic gating and binding noise compared to the RE approach. **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. 5 **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.

For multidimensional cPCF data all methods yield similar predictive accuracies and select the true data-generating process (Fig. 6**c**). The 2 dimensional data create a unique peak for the true process for all data sets.

Conclusion

We derived the prediction (Methods Eq. 26 and 30), the output statistics (Methods Eq. 48) 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 similarly good as those obtained by deterministic modeling with REs when using 10 times more channels or $\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**). Moreover, we showed that overfitting can be detected by continuous model expansion (Fig. 5**g-i**). Usually the KF maximizes its posterior by a vanishing rate into an non-existing state. This is not the case if the current time traces are analyzed by REs (Fig. 5**h**). The potential of prior distributions is shown by applying a weakly informative prior on one critical rate which increases the accuracy of continuous model expansion approach (Fig. 5**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. 6**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. 6**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. 6**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 $\mathrm{var}(r) = 1$ and overfitting is revealed distinctively by the variance of the fluorescence signal (Fig. 5**f**). Together this demonstrates the potential of a full Bayesian treatment of the state estimation, parameter estimation and model selction.

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 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. Press 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 more detailed description 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 [55,56], 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

$$\mathbb{P}(\boldsymbol{\theta}|\mathcal{Y}_T) = \frac{1}{Z(\mathcal{Y}_T)} \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta}) \prod_j^{N_{\text{par}}} \mathbb{P}(\theta_j) \tag{10}$$

is a probability distribution of a parameter set $\boldsymbol{\theta}$ conditioned on a measured time series $\mathcal{Y}_T$. 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

$$Z(\mathcal{Y}_T) = \int \mathbb{L}(\mathcal{Y}_T|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{11}$$

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. 10 by the expression of Eq. 4c.

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 a dimension 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 $\mathrm{S} := \{(1,0,0,...), (0,1,0,...), ..., (...,0,1)\} \subset \mathbb{N}^M$. In the following, Greek subscripts refer to different states while Latin subscripts refer to different channels. By $\mathbf{s}(t) = \mathbf{e}_\alpha$ we specify that the channel is in state α at time t. Mathematically, $\mathbf{e}_i$ stands for the i-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:

$$\mathbb{P}(\text{path}) = \mathbb{P}(\mathbf{s}(0), \mathbf{s}(1), ..., \mathbf{s}(T)) = \mathbb{P}(\mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(1) \mid \mathbf{s}(0)) \cdot \mathbb{P}(\mathbf{s}(2) \mid \mathbf{s}(1)) \cdots \mathbb{P}(\mathbf{s}(T) \mid \mathbf{s}(T-1)). \tag{12}$$

Markov models (MMs) and rate models are widely used for modeling molecular kinetics (Supplementary Sec. 3). 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 [78]. As 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:

$$\mathbf{T}_{\alpha,\beta} := \mathbb{P}(\mathbf{s}(t+1) = \mathbf{e}_\alpha \mid \mathbf{s}(t) = \mathbf{e}_\beta) = \exp(\mathbf{K} \cdot \Delta t)_{\alpha,\beta}, \tag{13}$$

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

$$\mathbf{p}(t) = \mathbb{E}(\mathbf{s}(t)) \tag{14}$$

The probability vector obeys the discrete-time Master equation

$$\mathbf{p}(t+1) = \mathbf{T}\mathbf{p}(t) \tag{15}$$

$$\mathbb{E}(\mathbf{s}(t+1)) = \mathbf{T}\mathbb{E}(\mathbf{s}(t)) \tag{16}$$

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:

$$\mathbf{n}(t) = \sum_{i=1}^{N_{\text{ch}}} \mathbf{s}_i(t) \tag{17}$$

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 $\mathbf{n} \in \mathbb{R}_{0+}^M$. From the linearity of Eq. 17 in the channel dimension and from the single-channel CME Eq. 16 one can immediately derive the equation for the time evolution of the mean occupancy $\overline{\mathbf{n}}(t) = \mathbb{E}[\mathbf{n}(t)]$:

$$\overline{n}_\alpha(t+1) = \sum_\beta T_{\alpha,\beta} \overline{n}_\beta(t) \tag{18}$$

Symbol	Meaning
$\boldsymbol{\theta}$	set of all unknown model parameters for which the posterior distribution is sampled
$\mathbf{n}(t)$	hidden ensemble occupancy vector of channel states in a specific patch at time t which is a continuous Markov state vector $\mathbf{n}(t) \in \mathbb{R}^M$
$\mathbf{P}(t)$	variance-covariance matrix of a hidden ensemble state $\mathbf{n}(t)$ in a specific patch at time t which contains the dispersion of the ensemble and the lacking knowledge of the algorithm about the true $\mathbf{n}(t)$
$\mathbf{T}$	transition matrix of a single channel
$\mathbf{K}$	rate matrix which is the logarithm of the transition matrix $\mathbf{T} = \exp(\mathbf{K}\Delta t)$
$\mathbf{H}$	observation matrix which projects the hidden ensemble state vector onto its mean signal.
$\mathbf{s}$	single-molecule Markov state vector
$k_{i,j}$	specific transition rate from state j to state i, $[\mathbf{K}]_{i,j} = k_{i,j}$
K_i	ratio of two transition rates i.e. an equilibrium constant
$\mathbf{y}(t)$	data point at time t
$\mathcal{Y}_T$	time series of T data points, $\mathcal{Y}_T = \{\mathbf{y}(t_i)\}_{i=1}^T$
$\mathfrak{N}_T$	time series of T hidden ensemble states, $\mathfrak{N}_T = \{\mathbf{n}(t_i)\}_{i=1}^T$
$N_{\mathrm{ch},j}$	number of channels in patch number j
i	mean electrical current through a single-channel
σ_{m}^2	variance of the current including all noise from the patch and the recording system
σ_{op}^2	variance of the current noise generated by a single open-channel
λ_{b}	mean brightness of a bound ligand
λ_{Fl}	mean brightness of the fluorescence signal from bulk and bound ligands
σ_{bulk}^2	variance of the fluorescence generated by unbound ligands after subtraction of the image obtained for the reference dye
M	number of single-channel states which is the dimension of $\mathbf{n}(t) \in \mathbb{R}_{0+}^M$ in the KF algorithm
N_{obs}	dimensions of the observational space
$\mathbb{F}(\mathcal{Y})$	true probability density of $\mathcal{Y}$, i.e. the true data-generating process
$\mathbb{L}(\mathcal{Y}\|\boldsymbol{\theta})$	likelihood function of the model parameters
$\mathbb{P}(\boldsymbol{\theta}\|\mathcal{Y})$	posterior distribution of the model parameters
$\mathbb{P}_{\mathrm{pred}}(\widetilde{\mathcal{Y}}\|\mathcal{Y})$	predictive distribution of the new data points $\widetilde{\mathcal{Y}}$
$\mathbb{O}(\mathbf{y}\|\mathbf{n})$	distribution of observables for a single time step
$\mathcal{N}(\cdot)$	normal distribution

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 α. Already after one time step, the channels will have spread out over the state space. The channel distribution after one time step is parametrized by the transition probabilities in row number α of the single-channel transition matrix $\mathbf{T}$. According to the theory of Markov models, the final channel distribution of channels originating from state α is the multinomial distribution

$$\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha) = \mathbb{P}(n_1, ..., n_M \mid \mathbf{n}(t) = n_\alpha \mathbf{e}_\alpha) = \frac{n_\alpha!}{n_1! \cdots n_M!} T_{1,\alpha}^{n_1} \cdots T_{M,\alpha}^{n_M} \tag{19}$$

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

$$\mathbf{n}(t+1) = \sum_\alpha \mathbf{n}^{(\alpha)}(t+1) \tag{20}$$

Evidently, the total number of channels is conserved during propagation. The distribution of $\mathbf{n}(t+1)$, defined by Eqs. 19 and 20, is called the *generalized multinomial distribution*:

$$\mathbf{n}(t+1) \sim \text{general-multinomial}(\mathbf{n}(t), \mathbf{T}) \tag{21}$$

While no simple expression exists for the generalized multinomial distribution, closed form expressions for its moments can be readily derived. For large N_{ch} each $\mathbb{P}(\mathbf{n}^{(\alpha)}(t+1) \mid n_\alpha \mathbf{e}_\alpha)$ can be approximated by a multivariate-normal distribution such that also 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 (19) has the following mean and covariance matrix

$$\overline{\mathbf{n}}^{(\alpha)}(t+1) = n_\alpha \mathbf{T}_{:,\alpha} \tag{22}$$

$$\mathbf{\Sigma}^{(\alpha)}(t+1) = n_\alpha \text{diag}(\mathbf{T}_{:,\alpha}) - n_\alpha \mathbf{T}_{:,\alpha,:} \mathbf{T}_{:,\alpha}^\top \tag{23}$$

where $\mathbf{T}_{:,\alpha}$ denotes the column number α of the transition matrix and diag$(\mathbf{T}_{:,\alpha})$ describes the diagonal matrix with $\mathbf{T}_{;,\alpha}$ on its diagonal. Combining Eq. 20 with Eqs. 22 and 23 we deduce the mean and variance of the generalized multinomial distribution:

$$\mathbb{E}\left(\mathbf{n}(t+1) \mid \mathbf{n}(t)\right) = \sum_\alpha n_\alpha(t) \mathbf{T}_{:,\alpha} = \mathbf{T}\mathbf{n}(t) \tag{24}$$

$$\text{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right) = \sum_\alpha n_\alpha(t) \left(\text{diag}(\mathbf{T}_{:,\alpha}) - \mathbf{T}_{:,\alpha}\mathbf{T}_{:,\alpha}^\top\right) = \text{diag}\left(\mathbf{T}\mathbf{n}(t)\right) - \mathbf{T}\text{diag}\left(\mathbf{n}(t)\right)\mathbf{T}^\top \tag{25}$$

Note that Eqs. 24 and 25 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. 24, giving

$$\overline{\mathbf{n}}(t+1) = \mathbb{E}\left[\mathbb{E}\left(\mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] = \mathbf{T}\overline{\mathbf{n}}(t), \tag{26}$$

in agreement with the simple derivation of Eq. 18. Similarly, the absolute covariance matrix of $\mathbf{n}(t+1)$ can be found by applying the law of total variance decomposition[68] to Eqs. 25 and 24, giving

$$\mathbf{P}(t+1) := \text{cov}(\mathbf{n}, \mathbf{n})(t+1) = \mathbb{E}\left[\text{cov}\left(\mathbf{n}(t+1), \mathbf{n}(t+1) \mid \mathbf{n}(t)\right)\right] + \text{cov}\left[\mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t)), \mathbb{E}(\mathbf{n}(t+1) \mid \mathbf{n}(t))\right] \tag{27}$$

$$= \text{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\text{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^\top + \text{cov}(\mathbf{T}\mathbf{n}(t), \mathbf{T}\mathbf{n}(t)) \tag{28}$$

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

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

where we have introduced the shorthand $\mathbf{P}(t) = \mathrm{cov}(\mathbf{n}(t), \mathbf{n}(t))$ in the last line. Eqs. 26, 30 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[48,66](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. 26, 30. Comparing Eq. 30 with the KF prediction equation[66] for $\mathbf{P}(t)$ we obtain the state-dependent covariance matrix of Eq. 1 as

$$\mathbf{Q}(\mathbf{T}, \overline{\mathbf{n}}(t)) = \mathrm{diag}\left(\mathbf{T}\overline{\mathbf{n}}(t)\right) - \mathbf{T}\mathrm{diag}\left(\overline{\mathbf{n}}(t)\right)\mathbf{T}^{\top} \tag{31}$$

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.

$$\mathbf{y}(t) = \begin{pmatrix} \text{fluorescence intensity}(t) \\ \text{ion current}(t) \end{pmatrix} = \begin{pmatrix} y_{\mathrm{Fl}}(t) \\ y_{\mathrm{I}}(t) \end{pmatrix} \tag{32}$$

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

$$\mathbb{O}(\mathbf{y}(t)|\mathbf{n}(t)) = \mathcal{N}\big(\mathbf{y}(t)|\mathbf{H}\mathbf{n}(t), \mathbf{\Sigma}(\mathbf{n}(t))\big) \tag{33}$$

$$\Leftrightarrow \mathbf{y}(t) = \mathbf{H}\mathbf{n}(t) + \boldsymbol{\nu}(\boldsymbol{t}). \tag{34}$$

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$(in_4, \sigma_{\mathrm{op}}^2 n_4)$	normal$(0, \sigma_{\mathrm{m}}^2)$	Poisson$(\lambda_b n_i(t))$	scaled Skellam
contribution to $\mathbf{H}$	$H_{2,4} = i$	-	$H_{1,:} = (0, \lambda_b, 2\lambda_b, 2\lambda_b)$	-
contribution to $\mathbf{\Sigma}$	$\Sigma_{2,2} = \sigma_{\mathrm{op}}^2 n_4(t)$	$\Sigma_{2,2} = \sigma_{\mathrm{m}}^2$	$\Sigma_{1,1} = (0, \lambda_b, 2\lambda_b, 2\lambda_b)\mathbf{n}(t)$	$\Sigma_{1,1} = \sigma_{\mathrm{back}}^2$

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:

$$\mathbf{H} = \mathbf{H}_{\mathrm{I}} + \mathbf{H}_{\mathrm{binding}} \tag{35}$$

$$\mathbf{\Sigma}(t) = \mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{meas.}} + \mathbf{\Sigma}_{\mathrm{binding}}(t) + \mathbf{\Sigma}_{\mathrm{back}} \tag{36}$$

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 σ^2_{open}. The instrumental noise of the experimental setup 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

$$\mathbf{H}_{\mathrm{I}} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \end{pmatrix} \tag{37}$$

$$\mathbf{\Sigma}_{\mathrm{open}}(t) + \mathbf{\Sigma}_{\mathrm{meas.}} := \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{\mathrm{op}} n_4(t) + \sigma^2_{\mathrm{m}} \end{pmatrix} \tag{38}$$

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

$$\begin{aligned} \mathrm{var}(\mathbf{y}(t)) &= \mathbb{E}[\mathrm{var}[\mathbf{y}(t)|\mathbf{n}(t)]] + \mathrm{var}[\mathbb{E}[\mathbf{y}(t)|\mathbf{n}(t)]] && \text{(39a)} \\ &= \mathbb{E}[\sigma^2_{\mathrm{op}} n_4(t) + \sigma^2_{\mathrm{m}}] + \mathrm{var}[\mathbf{H}_{\mathrm{I}}\mathbf{n}(t)] && \text{(39b)} \\ &= \sigma^2_{\mathrm{op}} \mathbb{E}[n_4(t)] + \sigma^2_{\mathrm{m}} + (\mathbf{H}_{\mathrm{I}}\mathbf{P}(t)\mathbf{H}_{\mathrm{I}}^{\top})_{2,2} && \text{(39c)} \end{aligned}$$

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 λ_{Fl}

$$y_{\mathrm{Fl}}(t) \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}(t)). \tag{40}$$

The total emission rate λ_{Fl} can be modeled as a weighted sum of the specific emission rates λ_b of each ligand class $\{0, 1, 2\}$. The weights are given by the stoichiometric factors of each ligand class. In order to cast the Poisson distribution into the functional form of the observation model (Eq. 33), we invoke the central limit theorem to approximate

$$y_{\mathrm{Fl}} \sim \mathrm{pois}(\lambda_{\mathrm{Fl}}) \approx \mathcal{N}(\lambda_{\mathrm{Fl}}(t), \lambda_{\mathrm{Fl}}(t)) \tag{41}$$

The larger λ_{Fl} the better is the approximation. We assume, that the confocal volume is equally illuminated. For our model of ligand fluorescence, we assume for a moment that there is no signal coming from ligands in the bulk.

We will drop this assumption in the next section. With these assumptions, we arrive at the following observation matrix

$$\mathbf{H}_{\text{binding}} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{42}$$

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. 41. We find

$$\mathbf{\Sigma}_{\text{binding}}(t) := \begin{pmatrix} (\mathbf{Hn}(t))_1 & 0 \\ 0 & 0 \end{pmatrix} \tag{43}$$

Under these assumptions the observation matrix can be written as follows

$$\mathbf{H} := \begin{pmatrix} 0 & \lambda_b & 2\lambda_b & 2\lambda_b \\ 0 & 0 & 0 & i \end{pmatrix} \tag{44}$$

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

$$\mathbf{y}(t) = \mathbf{Hn}(t) + \boldsymbol{\nu}_{\text{m}}(t) + \begin{pmatrix} \nu_{\text{pois}}(t) \\ \nu_{\text{op}}(t) \end{pmatrix} \Leftrightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{Hn}(t), \mathbf{\Sigma}(t)). \tag{45}$$

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

$$\begin{aligned} \text{var}[y_{\text{fluo}}(t)] &= \mathbb{E}[\text{var}(y(t)|\mathbf{n}(t))] + \text{var}[\mathbb{E}(y(t)|\mathbf{n}(t)] && \text{(46a)} \\ &= \mathbb{E}[\sigma^2_{\text{back}} + (\mathbf{Hn}(t))_1] + \text{var}(\mathbf{Hn}(t)) && \text{(46b)} \\ &= \sigma^2_{\text{back}} + (\mathbf{Hn}(t))_1 + (\mathbf{HP}(t))\mathbf{H}^\top)_{1,1} && \text{(46c)} \end{aligned}$$

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

$$\begin{aligned} \text{cov}(y_{\text{patch}}, y_{\text{fluo}}) &= \mathbb{E}[\text{cov}(y_{\text{patch}}, y_{\text{fluo}}|\mathbf{n})] + \text{cov}(\mathbb{E}(y_{\text{patch}}|\mathbf{n}), \mathbb{E}(y_{\text{fluo}}|\mathbf{n})) && \text{(47a)} \\ &= 0 + \text{cov}(\mathbf{H}_{2,:}\mathbf{n}, \mathbf{H}_{1,:}\mathbf{n}) && \text{(47b)} \\ &= \mathbf{H}_{2,:}\,\text{cov}(\mathbf{n}, \mathbf{n})\mathbf{H}^\top_{1,:} = \mathbf{H}_{2,:}\mathbf{P}(t)\mathbf{H}^\top_{1,:} && \text{(47c)} \end{aligned}$$

yielding the matrix

$$\text{cov}(\mathbf{y}, \mathbf{y}) = \mathbf{HP}(t)\mathbf{H}^\top + \mathbf{\Sigma}(t) \tag{48}$$

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. 24) and for the covariance matrix (Eq. 30) as well as with the expressions for the signal variance we possess all parameters that are needed in the correction equation of the KF [49,66].

Bayesian Model selection via predictive accuracy

There are several statistical approaches to identify the best fitting model [77,79–81]. Those approaches balance accuracy of the model's predictions with its simplicity. Some of them, such as Maximum evidence or BIC, [79] rely on the assumption that the true data-generating process is included in the set of models to be tested. That is called the

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[41]. 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[41]. 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.

M−closed situation[79]. These approaches perform well in simulation studies[82] if one is in fact in an M−closed setting. However, it remains an open question how they perform when applied on real data, where all models can only be approximations of the unknown real data-generating process $\mathbb{F}(y)$. When we model continuous, high-dimensional dynamics by models with a few discrete states we obviously do not include $\mathbb{F}(y)$ in the set of candidate models. Moreover, the "simplicity" of a model is ultimately subjective because it depends on the choice of the (formal) language in which a model is described.[81] 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]. The minimum of the Kullback-Leibler divergence can be found asymptotically by maximizing the predictive accuracy[62,83] The predictive accuracy is defined as

$$\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T) := \log \mathbb{E}_{\theta,\mathrm{post}}[\mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})] = \log \int \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta})\mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta}|\mathcal{Y}_T)\,\mathrm{d}\boldsymbol{\theta} \tag{49}$$

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. 49 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. 49 to $\log \mathbb{L}(\tilde{\mathcal{Y}}_T|\boldsymbol{\theta}_{\mathrm{MLE}})$. 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

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] = \int \cdots \int \log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)\mathbb{F}(\tilde{\mathcal{Y}}_T) \prod_i^T \mathrm{d}\tilde{\mathbf{y}}_i \tag{50}$$

Unfortunately, $\mathbb{F}(\tilde{\mathcal{Y}}_T)$ is unknown. In practice the expectation value in Eq. 50 is approximated summing over independent realizations generated by the experiment.

$$\mathbb{E}_{\mathbb{F}}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T|\mathcal{Y}_T)] \approx \sum_{k=1}^{K}[\log \mathbb{P}_{\mathrm{pred}}(\tilde{\mathcal{Y}}_T^{(k)}|\mathcal{Y}_T)] \tag{51}$$

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

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

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 predictive accuracy estimate[62]. To compensate for that optimistic bias one needs to penalize the model complexity[62]. Recently, Watanabe[63] showed for singular models that asymptotically for large data sets WAIC is equivalent to Bayesian cross-validation and therefore equivalent by definition to the predictive accuracy. Thus it is an estimator of the predictive accuracy. The predictive accuracy of the model Eq. 50 can be estimated by

$$\mathrm{WAIC} := \log \mathbb{P}_{\mathrm{post}}(\mathcal{Y}_T) - p_{\mathrm{WAIC}} \tag{52}$$

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 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. 6 a, b. Since WAIC exploits the full posterior our approach needs inevitably to be Bayesian.
The famous non-Bayesian predecessor of WAIC was found by Akaike[84]. 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

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

Only then is AIC asymptotically equivalent to cross-validation[85]. These very restrictive model conditions are not satisfied by models with hidden structures[60]. We show (Fig. 6 b), that AIC fails to predict the true data-generating process unless A second way to estimate Eq. 50 by Eq. 49 is with hold out data. The KF modeling reveals the true model when using WAIC or cross-validation. The small decrease in the predictive accuracy when using a KF can be explained by the observation that the KF assigns most probability mass to vanishing rates into non-existing states (Fig. 5**g-j**). The KF switches off the rates into non-existing states of our model system. This means M_4 effectively becomes M_{true} after the training. In order to decide for 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] J. Oyrer, S. Maljevic, I. E. Scheffer, S. F. Berkovic, S. Petrou, and C. A. Reid, "Ion channels in genetic epilepsy: from genes and mechanisms to disease-targeted therapies," *Pharmacological reviews*, vol. 70, no. 1, pp. 142–173, 2018.

[2] M. P. Goldschen-Ohm, D. A. Wagner, S. Petrou, and M. V. Jones, "An epilepsy-related region in the gabaa receptor mediates long-distance effects on gaba and benzodiazepine binding sites," *Molecular pharmacology*, vol. 77, no. 1, pp. 35–45, 2010.

[3] C. E. Clancy and Y. Rudy, "Cellular consequences of herg mutations in the long qt syndrome: precursors to sudden cardiac death," *Cardiovascular research*, vol. 50, no. 2, pp. 301–313, 2001.

[4] A. O. Verkerk and R. Wilders, "Pacemaker activity of the human sinoatrial node: effects of hcn4 mutations on the hyperpolarization-activated current," *Europace*, vol. 16, no. 3, pp. 384–395, 2014.

[5] H. A. Colquhoun D., *The Principles of the Stochastic Interpretation of Ion-Channel Mechanisms. In: Sakmann B., Neher E. (eds) Single-Channel Recording.* Springer, Boston, MA, (1995).

[6] H. A. Colquhoun D., *A Q-Matrix Cookbook. In: Sakmann B., Neher E. (eds) Single-Channel Recording.* Springer, Boston, MA, (1995).

[7] N. Erwin and S. Bert, "Single-channel currents recorded from membrane of denervated frog muscle fibres," *Nature*, vol. 260, no. 5554, pp. 799–802, 1976.

[8] C. David, H. A. G., and K. Bernard, "On the stochastic properties of single ion channels," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 211, no. 1183, pp. 205–235, 1981. doi: 10.1098/rspb.1981.0003.

[9] R. Horn and K. Lange, "Estimating kinetic constants from single channel data," *Biophysical Journal*, vol. 43, no. 2, pp. 207–223, 1983.

[10] Epstein Michael, Calderhead Ben, Girolami Mark A., and Sivilotti Lucia G., "Bayesian Statistical Inference in Ion-Channel Models with Exact Missed Event Correction," *Biophysical Journal*, vol. 111, no. 2, pp. 333–348, 2016. doi: 10.1016/j.bpj.2016.04.053.

[11] I. Siekmann, M. Fackrell, E. J. Crampin, and P. Taylor, "Modelling modal gating of ion channels with hierarchical markov models," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 472, no. 2192, p. 20160122, 2016.

[12] S.-H. Chung, J. B. Moore, L. Xia, L. Premkumar, and P. W. Gage, "Characterization of single channel currents using digital signal processing techniques based on hidden markov models," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 329, no. 1254, pp. 265–285, 1990.

[13] D. R. Fredkin and J. A. Rice, "Maximum likelihood estimation and identification directly from single-channel recordings," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 249, no. 1325, pp. 125–132, 1992.

[14] F. Qin, A. Auerbach, and F. Sachs, "Hidden Markov Modeling for Single Channel Kinetics with Filtering and Correlated Noise," *Biophysical Journal*, vol. 79, no. 4, pp. 1928–1944, 2000.

[15] L. Venkataramanan and F. Sigworth, "Applying hidden Markov models to the analysis of single ion channel activity," *Biophysical journal*, vol. 82, no. 4, pp. 1930–1942, 2002.

[16] C. Y. K. J. B. Ball F. G. and O. A., "Bayesian inference for ion–channel gating mechanisms directly from single–channel recordings, using Markov chain Monte Carlo," *455Proc. R. Soc. Lond. A*, 1999.

[17] R. Rosales, J. A. Stark, W. J. Fitzgerald, and S. B. Hladky, "Bayesian Restoration of Ion Channel Records using Hidden Markov Models," *Biophysical Journal*, vol. 80, no. 3, pp. 1088–1103, 2001.

[18] R. A. Rosales, "MCMC for hidden Markov models incorporating aggregation of states and filtering," *Bulletin of Mathematical Biology*, vol. 66, pp. 1173–1199, Sep 2004.

[19] E. Gin, M. Falcke, L. E. Wagner, D. I. Yule, and J. Sneyd, "Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors," *Journal of Theoretical Biology*, vol. 257, no. 3, pp. 460–474, 2009.

[20] I. Siekmann, L. E. Wagner, D. Yule, C. Fox, D. Bryant, E. J. Crampin, and J. Sneyd, "MCMC Estimation of Markov Models for Ion Channels," *Biophysical Journal*, vol. 100, no. 8, pp. 1919–1929, 2011.

[21] I. Siekmann, J. Sneyd, and E. J. Crampin, "MCMC Can Detect Nonidentifiable Models," *Biophysical Journal*, vol. 103, no. 11, pp. 2275–2286, 2012.

[22] K. E. Hines, J. R. Bankston, and R. W. Aldrich, "Analyzing Single-Molecule Time Series via Nonparametric Bayesian Inference," *Biophysical Journal*, vol. 108, no. 3, pp. 540–556, 2015.

[23] J. Zheng and W. N. Zagotta, "Gating rearrangements in cyclic nucleotide-gated channels revealed by patch-clamp fluorometry," *Neuron*, vol. 28, no. 2, pp. 369–374, 2000.

[24] J. W. Taraska and W. N. Zagotta, "Structural dynamics in the gating ring of cyclic nucleotide–gated ion channels," *Nature structural & molecular biology*, vol. 14, no. 9, p. 854, 2007.

[25] J. W. Taraska, M. C. Puljung, N. B. Olivier, G. E. Flynn, and W. N. Zagotta, "Mapping the structure and conformational movements of proteins with transition metal ion fret," *Nature methods*, vol. 6, no. 7, p. 532, 2009.

[26] A. Bruening-Wright, F. Elinder, and H. P. Larsson, "Kinetic Relationship between the Voltage Sensor and the Activation Gate in spHCN Channels," *The Journal of General Physiology*, vol. 130, no. 1, pp. 71–81, 2007.

[27] T. Kalstrup and R. Blunck, "Dynamics of internal pore opening in kv channels probed by a fluorescent unnatural amino acid," *Proceedings of the National Academy of Sciences*, vol. 110, no. 20, pp. 8272–8277, 2013.

[28] T. Kalstrup and R. Blunck, "S4–s5 linker movement during activation and inactivation in voltage-gated k+ channels," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6751–E6759, 2018.

[29] M. Wulf and S. A. Pless, "High-sensitivity fluorometry to resolve ion channel conformational dynamics," *Cell reports*, vol. 22, no. 6, pp. 1615–1626, 2018.

[30] Biskup Christoph, Kusch Jana, Schulz Eckhard, Nache Vasilica, Schwede Frank, Lehmann Frank, Hagen Volker, and Benndorf Klaus, "Relating ligand binding to activation gating in CNGA2 channels," *Nature*, vol. 446, p. 440, feb 2007.

[31] J. Kusch, C. Biskup, S. Thon, E. Schulz, V. Nache, T. Zimmer, F. Schwede, and K. Benndorf, "Interdependence of Receptor Activation and Ligand Binding in HCN2 Pacemaker Channels," *Neuron*, vol. 67, no. 1, pp. 75–85, 2010.

[32] Kusch Jana, Thon Susanne, Schulz Eckhard, Biskup Christoph, Nache Vasilica, Zimmer Thomas, Seifert Reinhard, Schwede Frank, and Benndorf Klaus, "How subunits cooperate in cAMP-induced activation of homotetrameric HCN2 channels," *Nature Chemical Biology*, vol. 8, p. 162, dec 2011.

[33] Biskup Christoph, Kusch Jana, Schulz Eckhard, Nache Vasilica, Schwede Frank, Lehmann Frank, Hagen Volker, and Benndorf Klaus, "Relating ligand binding to activation gating in CNGA2 channels," *Nature*, vol. 446, p. 440, feb 2007.

[34] S. Wu, Z. V. Vysotskaya, X. Xu, C. Xie, Q. Liu, and L. Zhou, "State-Dependent cAMP Binding to Functioning HCN Channels Studied by Patch-Clamp Fluorometry," *Biophysical Journal*, vol. 100, no. 5, pp. 1226–1232, 2011.

[35] H. A. G. Colquhoun David and K. Bernard, "Relaxation and fluctuations of membrane currents that flow through drug-operated channels," *Proc. R. Soc. Lond. B*, 1997.

[36] P. Alcantara, L. M. Cardenas, S. Swillens, and R. S. Scroggs, "Reduced Transition between Open and Inactivated Channel States Underlies 5HT Increased INa+ in Rat Nociceptors," *Biophysical Journal*, vol. 83, no. 1, pp. 5–21, 2002.

[37] W. Wang, X. Feng, Z. Xuhui, Y. Jing, Y. Ming, and D. Jiuping, "Optimal Estimation of Ion-Channel Kinetics from Macroscopic Currents," *PLOS ONE*, vol. 7, pp. 1–12, 04 2012.

[38] T. G. Kurtz, "The relationship between stochastic and deterministic models for chemical reactions," *The Journal of Chemical Physics*, vol. 57, no. 7, pp. 2976–2978, 1972.

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

[40] T. Jahnke and W. Huisinga, "Solving the chemical master equation for monomolecular reaction systems analytically," *Journal of Mathematical Biology*, vol. 54, pp. 1–26, Jan 2007.

[41] L. S. Milescu, G. Akk, and F. Sachs, "Maximum Likelihood Estimation of Ion Channel Kinetics from Macroscopic Currents," *Biophysical Journal*, vol. 88, no. 4, pp. 2494–2515, 2005.

[42] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Molecular Systems Biology*, vol. 5, no. 1, p. 318, 2009.

[43] J. J. Celentano and A. G. Hawkes, "Use of the Covariance Matrix in Directly Fitting Kinetic Parameters: Application to GABAA Receptors," *Biophysical Journal*, vol. 87, no. 1, pp. 276–294, 2004.

[44] A. R. Stepanyuk, B. A. L., and B. P. V., "Efficient Maximum Likelihood Estimation of Kinetic Rate Constants from Macroscopic Currents," *PLOS ONE*, vol. 6, pp. 1–18, 12 2011.

[45] A. Stepanyuk, A. Borisyuk, and P. Belan, "Maximum likelihood estimation of biophysical parameters of synaptic receptors from macroscopic currents," *Frontiers in Cellular Neuroscience*, vol. 8, p. 303, 2014.

[46] T. Jahnke and W. Huisinga, "Solving the chemical master equation for monomolecular reaction systems analytically," *Journal of mathematical biology*, vol. 54, no. 1, pp. 1–26, 2007.

[47] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb 1989.

[48] L. Moffatt, "Estimation of Ion Channel Kinetics from Fluctuations of Macroscopic Currents," *Biophysical Journal*, vol. 93, no. 1, pp. 74–91, 2007.

[49] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[50] M. Komorowski, B. Finkenstädt, C. V. Harper, and D. A. Rand, "Bayesian inference of biochemical kinetic parameters using the linear noise approximation," *BMC Bioinformatics*, vol. 10, pp. 343–343, 2009.

[51] B. Finkenstädt, D. J. Woodcock, M. Komorowski, C. V. Harper, J. R. Davis, M. R. White, D. A. Rand, *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*, vol. 7, no. 4, pp. 1960–1982, 2013.

[52] P. Fearnhead, V. Giagos, and C. Sherlock, "Inference for reaction networks using the linear noise approximation," *Biometrics*, vol. 70, no. 2, pp. 457–466, 2014.

[53] M. M. Folia and M. Rattray, "Trajectory inference and parameter estimation in stochastic models with temporally aggregated data," *Statistics and Computing*, vol. 28, pp. 1053–1072, Sep 2018.

[54] C. S. Gillespie and A. Golightly, "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, p. 23, 2012.

[55] K. E. Hines, "A Primer on Bayesian Inference for Biophysical Systems," *Biophysical Journal*, vol. 108, no. 9, pp. 2103–2113, 2015.

[56] F. Ball, "MCMC for Ion-Channel Sojourn-Time Data: A Good Proposal," *Biophysical Journal*, vol. 111, no. 2, pp. 267–268, 2016.

[57] Auger-Methe Marie, Field Chris, Albertsen Christoffer M., Derocher Andrew E., Lewis Mark A., Jonsen Ian D., and Mills Flemming Joanna, "State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems," *Scientific Reports*, vol. 6, p. 26677, may 2016.

[58] K. E. Hines, T. R. Middendorf, and R. W. Aldrich, "Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach," *The Journal of General Physiology*, vol. 143, no. 3, pp. 401–416, 2014.

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

[60] S. Watanabe, "Almost All Learning Machines are Singular," in *2007 IEEE Symposium on Foundations of Computational Intelligence*, pp. 383–388, April 2007.

[61] P. Kienker, "Equivalence of aggregated markov models of ion-channel gating," *Proceedings of the Royal Society of London. B. Biological Sciences*, vol. 236, no. 1284, pp. 269–309, 1989.

[62] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, vol. 24, pp. 997–1016, Nov 2014.

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

[64] Gillespie Daniel T., "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977. doi: 10.1021/j100540a008.

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

[66] B. D. Anderson and J. B. Moore, *Optimal filtering.* Courier Corporation, 2012.

[67] Z. Chen *et al.*, "Bayesian filtering: From kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

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

[69] M. D. Hoffman and A. Gelman, "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.

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

[71] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for bayesian inference and optimization," *Journal of Educational and Behavioral Statistics*, vol. 40, no. 5, pp. 530–543, 2015.

[72] D. J. Zwickl and M. T. Holder, "Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics," *Systematic Biology*, vol. 53, no. 6, pp. 877–888, 2004.

[73] A. Gelman, D. Simpson, and M. Betancourt, "The prior can often only be understood in the context of the likelihood," *Entropy*, vol. 19, no. 10, p. 555, 2017.

[74] A. Vehtari and J. Ojanen, "A survey of Bayesian predictive methods for model assessment, selection and comparison," *Statist. Surv.*, vol. 6, pp. 142–228, 2012.

[75] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[76] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding aic and bic in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

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

[78] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, "Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules," *The Journal of Chemical Physics*, vol. 139, no. 18, p. 184114, 2013.

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

[80] J. Piironen and A. Vehtari, "Comparison of bayesian predictive methods for model selection," *Statistics and Computing*, vol. 27, no. 3, pp. 711–735, 2017.

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

[82] J. E. Bronson, J. Fei, J. M. Hofman, R. L. Gonzalez Jr, and C. H. Wiggins, "Learning rates and states from biophysical time series: a bayesian approach to model selection and single-molecule fret data," *Biophysical journal*, vol. 97, no. 12, pp. 3196–3205, 2009.

[83] K. Burham and D. Anderson, "Model selection and multinomial inference," 1998.

[84] H. Akaike, *A New Look at the Statistical Model Identification*, pp. 215–222. New York, NY: Springer New York, 1998.

[85] M. Stone, "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.