# MINIMALLY AND VAGUELY INFORMATIVE PRIORS TO COMBAT PRACTICAL PARAMETER NON-IDENTIFIABILITY OF HIDDEN MARKOV MODELS EXEMPLIFIED BY ION CHANNEL DATA

**Jan L. Münch Ralf Schmauder Institute of Physiology II**
Jena University Hospital, Friedrich Schiller University, Jena 07743, Germany
jan.muench@med.uni-jena.de

**Fabian Paul**
Department of Biochemistry and Molecular Biology
University of Chicago
Chicago , United States
email@email

**Michael Habeck**
Microscopic Image Analysis Group
Jena University Hospital,
Friedrich Schiller University,
Jena 07743, Germany
michael.habeck@uni-jena.de

## ABSTRACT

Hidden Markov Model (HMM) inference for time-series data from ion channels or other biomolecules is challenging. We argue that inference on partially observed chemical reaction networks (CRNs) suffers from practical parameter non-identifiability (non-PI) that often goes unnoticed in maximum likelihood (ML) inferences. Limitations in the signal bandwidth and a poor signal-to-noise ratio only add to the non-PI problem. We study the role of the prior distribution in the face of non-PI. In particular, we advocate using minimally informative (MI) priors and additional restrictions on the parameter space that can be derived from physical considerations. Using patch clamp (PC) ion-channel measurements as a prototypical time series, we demonstrate Bayesian strategies for alleviating non-PI problems with sharpened prior information.

In Bayesian statistics, the prior can substantially modulate the posterior. We demonstrate that non-PI can be severely harmful when using uniform priors on the rate matrix of HMMs, which are implicitly assumed in ML. We show that MI priors enable meaningful HMM inference with data whose quality can be one to two orders of magnitude worse than required to reach the same accuracy with uniform priors. However, we also demonstrate that non-PI pathologies can persist even with a prior MI. In this case, the MI prior alleviates but does not entirely resolve the problem of improper posteriors. For complex HMMs, stronger prior assumptions are needed to render the posterior proper.

We propose to confine the parameters to a sampling box whose limits are physically reasonable and derived from theory. This fusion of data and physical information allows for meaningful inferences even for the most complex HMM with data of the lowest quality that we tested. However, hard theoretical limits, such as diffusion-limited binding rates, are rarely available. As an alternative, we test a vague prior on the ratios of each pair of binding rates and additionally unbinding rates, thereby softly linking them. This implicitly assumes finite cooperativity and introduces a bias towards non-cooperativity. However, in contrast to the standard practice of choosing equal chemical rates, which supposes strict non-cooperativity, this additional prior still allows for cooperativity. Despite its vagueness, our prior renders the posterior either proper in a strict sense or *sufficiently* proper for all data sets we considered without imposing the assumption of non-cooperativity. Hence, our approach can infer how likely different degrees of cooperativity are. Combining theoretical upper limits and vague finite cooperativity assumptions dramatically improves inferences.

**Keywords** Bayesian statistics · Hidden Markov Model · HMM · minimally informative prior distribution · vaguely informative prior distribution · Parameter non-identifiability · Kinetic scheme · Chemical reaction network · Ion

channels · Patch-clamp · Fokker-Planck expansion · Bayesian filter · Chemical master equation · ligand-gated ion channels

# 1 Introduction

Time series data of highly diverse processes in physics [1, 2], chemistry [3, 4, 5, 6, 7], biology, or epidemics [8, 9], can be explained by chemical reaction networks (CRNs) or, to use an alternative terminology, continuous time Markov Models (MM) [10]. In a biophysical context, the data could probe, e.g., the function of ion channels [11, 12], transporters [13, 14, 15] and molecular motors [16] as well as RNA folding kinetics [17, 18]. Assuming a well-mixed environment, these systems can be approximately modeled by discrete Markov states and stochastic, discrete chemical reactions, thereby defining a CRN. For example, Markov states of a ligand-gated ion channel (LGICs) are classified by the number of bound ligands and their conductance, or more subtle by their statistical behavior [19, 20, 21, 18]. The states of a CRN interconvert via stochastic transitions, characterized by transition rates. The assumption that the system adopts discrete states typically coarse grains and thus simplifies [22] the energy landscape or other aspects of the actual data-generating process (DGP).

Biophysical experiments typically do not observe the states directly but can only measure indirect physical signals such that states with similar properties like brightness or conductance are lumped (aggregated) together. Furthermore, the experimental data are noisy. Therefore, more complex statistical models, called hidden Markov models (HMMs), must be used to describe the data [23, 24]. A prototypical use case for HMMs is single-molecule ion channel data that show discrete stochastic jumps in their conductance [25, 26, 27, 28, 29, 30, 31]. However, HMM analysis is not limited to single-molecule ion channel data [17, 18] and can also be applied to e.g. fast kinetics from fluorescence measurements [32, 33, 34, 35, 36].

Model selection aims to infer from the data the number of states of the CRN and their connectivity. In some cases, model selection for discrete-time HMMs can be automated by a Dirichlet process [37, 38]. However, a correction for data imperfections such as a baseline drift must be included in a complete non-parametric Bayesian analysis [39] to avoid a bias in the estimated number of states. To our knowledge, the infinite HMM [37, 38] only applies to single-molecule data but cannot be extended to ensemble data or continuous time HMMs.

Notably, the hidden variable of a HMM does not have to be discrete but can be used to approximate a discrete CRN [40, 41] A Kalman filter [42, 43] also defines a valid HMM, as demonstrated for ensemble patch-clamp (PC) data [40] using a continuous hidden variable [44, 45, 46, 47]. Here, we build on a generalization of the Kalman filter[41], necessary to employ the ideas of [40] on actual PC data, which uses a continuous hidden variable and continuous time HMM to analyze ensemble PC data of LGICs as well as ensemble binding data [41].

The mean signal of an HMM that typically involves an aggregation of states (partial observations) is a linear projection of the entire dynamics of the CRN to a lower dimensional observable for both single-molecule or ensemble data. This can cause the data to be insensitive to the rates of specific subprocesses within the CRN, which might compromise its biophysical interpretation. Here, we argue that this projection (dimensionality reduction) almost always induces a varying degree of practical parameter non-identifiability (non-PI). Even worse, the HMM might become structurally non-identifiable (non-SI). Non-SI means that even an infinite amount and quality of data do not permit the determination of a unique set of model parameters. At best, the (relevant) system's dynamics can be disentangled by a suitable but not too complex CRN, informative prior information on some rate constants or derived parameters, and time-series data that constrain each rate constant on the associated time scale. In this case, the non-PI problem will still be present but negligible.

## 1.1 Bayesian inference: Priors that alleviate practical parameter non-identifiabilty

Model inference relies on a likelihood function, regardless of whether parameters are estimated by Maximum Likelihood (ML) [48, 49] or Bayesian statistics [50] in which case we also need to specify a prior probability [51]. In Bayesian statistics, parameters or other unknowns are treated similarly to random variables [52, 53] in that probabilities are used to express their uncertainty. The prior encodes their uncertainty before analyzing the data. The result of Bayesian inference, the posterior probability, represents the uncertainty of the unknowns in the light of the combined information encoded in the prior and the data.

Limitations in the amount and quality of data (relative to the complexity of the investigated CRN) severely impair both ML and maximum a posteriori (MAP) inferences. The covariance matrix of the model parameters captures the uncertainty of their ML or MAP estimates only in the asymptotic limit and under some regularity conditions [54, 50, 49]. Using ML or MAP on finite data raises the concern *whether the covariance matrix of the inferred parameters represents their uncertainty accurately.* There is no guarantee that this holds, even in the asymptotic limit, because HMMs do not satisfy the necessary regularity conditions [55]. Hence, the assumed standard asymptotic behavior valid only for ML inference of regular models (convergence of the distribution of the ML estimator to a normal distribution centered at the

true values) is often incorrect and usually associated with overly optimistic uncertainty estimates [56, 41]. Regardless of whether single-molecule or ensemble measurements are used, the reason for invalid regularity conditions is that the signal of an HMM probes hidden variables [55] from, e.g., the internal dynamics of biomolecules. Thus, using Bayesian statistics for HMM inference has advantages [57, 28, 29, 31, 58, 59, 60, 37] over ML concerning both parameter inference and model selection, particularly in biophysical applications.

This work focuses on the prior side of Bayesian inference for continuous-time HMMs. In particular, we study the benefits and limitations of minimally informative (MI) and vague physically motivated priors in the presence of non-PI likelihoods. We argue that non-PI problems are omnipresent for partially observed CRNs due to the projection of the mean signal and the system's intrinsic stochasticity. Experimental noise and signal bandwidth only increase the severeness of non-PI problems.

We benchmark the MI and vague prior assumptions against uniform priors. Uniform priors are a relevant class of priors because they are implicitly assumed in ML and often explicitly employed in Bayesian statistics. This practice originates in the misconception that uniform priors are always MI, which they are only for location parameters like the mean of a normal distribution [61]. As a starting point of the inference, it is essential to make the posterior as sensitive to the data as possible, which is achieved by MI priors. There is no first principle to define MI priors. Instead, several guidelines have been proposed [62, 63, 64, 65], some of which deliver identical priors for specific statistical models. Here, we will study MI priors inspired by Jeffreys's rule [62] and Jaynes's transformation group arguments [61].

Both the uniform and the proposed MI prior have a drawback: they are improper probability distributions, i.e., they cannot be normalized. We demonstrate that this leads to an improper posterior in combination with the non-PI problem of the likelihood of HMMs. We show that using an MI prior alleviates the challenges arising from the non-PI problem but does not fully resolve them. In a finite sampling box, the MI prior increases the accuracy and reduces the uncertainty of the posterior. The MI prior also enables meaningful inferences with *sufficiently* proper posteriors for data up to a hundredfold lower quality than the uniform prior. Furthermore, it drastically improves the convergence of the Hamiltonian Monte Carlo sampler [66, 67, 68, 69, 70] as judged by the Gelman-Rubin statistics [71].

In addition, we present techniques to render the posterior proper or *sufficiently* proper by enforcing theoretically derived diffusion limits for binding rates or by softly coupling parameters through vague priors on their ratios. This coupling biases the CRN towards non-cooperativity but does not enforce it. Adding vague priors is a more flexible approach than setting parameters equal. The prior only defines the scale of plausible positive or negative logarithmic ratios, i.e., cooperativity in homomeric proteins. The most certain and accurate inferences are obtained if the MI prior is combined with both additional assumptions.

### 1.2 Fokker-Planck approximations are used to derive the likelihood

The chemical master equation (CME) [72, 73] is the accepted forward model of the time evolution of a CRN in a well-mixed chemical environment. The solution of the CME is a probability distribution for which the discrete first-order Markov property holds. Given any initial conditions, the solution of the CME can be constructed by the corresponding Markov kernel. The Markov kernel of CRNs, consisting only of mono-molecular, i.e. (pseudo-)first-order, chemical reactions with no birth processes, can be understood as a convolution of $M$ multinomial distributions [41], called by us the generalized-multinomial distribution [41]. Adding birth and death processes requires additional convolution with Poisson distributions [74].

Traditionally, ensemble data on CRNs are analyzed by considering only the mean value of the CME, in the form of deterministic rate equations (REs) [75] thereby losing the Markov property [41]. This simplification of discrete stochastic events by a continuous deterministic process of the mean value inevitably leads to a likelihood misspecification [41], even with the correct CRN topology. This causes a less accurate and over-confident parameter inference [41] independently of the number of molecules, as long as the intrinsic fluctuations in the signal are not vanishing compared to the measurement noise [41].

A better approximation of the CME is the Fokker-Planck approximation (FPa) [76]. The FPa or similar stochastic approximations can correctly model the autocorrelation of the signal fluctuations originating from the first-order Markov property of the DGP. As the FPa models discrete chemical events as a continuous process, some likelihood misspecification persists [76] but vanishes asymptotically (in contrast to approaches based on REs).

Here, we use a Bayesian filter (BF) [41] based on the FPa that extends the ideas of Moffat [40] and Kalman [77, 78] to define a more general and realistic likelihood for PC data. The generalization is necessary to avoid a misspecification of the observation model (as in [40]), which would impede any inference from realistic PC or fluorescence data. This likelihood showed that the first-order Markov property is essential for uncertainty quantification independently of the actual number of molecules [41]. An inference based on an RE approach is too confident about the uncertainties and is less accurate.

## 2 Bayesian inference in a nutshell

Given time series data $\mathcal{Y}_T = \{y_1, \ldots, y_T\}$ of length $T$, and a probabilistic model in the form of a likelihood $p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})$, one can infer the unknown parameters $\boldsymbol{\theta}_{\text{true}}$ with ML [54] by optimization over the parameter space $\boldsymbol{\Theta}$. In contrast to the point estimate obtained by ML, the Bayesian posterior $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ where

$$p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T) = \frac{p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{1}$$

is a probability distribution whose support is $\boldsymbol{\Theta}$. The posterior fuses the information encoded in the modeling assumptions which are reflected in the prior $p(\boldsymbol{\theta})$ and in the functional form of the likelihood with the data $\mathcal{Y}_T$. The denominator needs to satisfy

$$\int p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty \tag{2}$$

such that $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ (Eq. 1) can be normalized; in other words $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ is a proper density. The posterior is a quantitative representation of the uncertainty of $\boldsymbol{\theta}$, given the modeling assumptions. A short comment on the philosophy of Bayesian statistics and Kolmogorov's mathematical foundation of statistics [79, 80] making a frequency interpretation of stochastic variables unnecessary [81] can be found in App. A.
In the following, normalizable posterior densities will be called "proper." We introduce the terminology sufficiently proper to clarify the notion of practical parameter non-identifiability (non-PI) discussed below (Fig 2). Non-PI in combination with MI priors which are often improper, still results in a posterior that is improper in a strict sense. This is illustrated in Fig 2 for the improper uniform prior. For a uniform prior we define sufficiently proper by setting a threshold $\psi$ such that

$$0 < \mathbb{R}(\boldsymbol{\theta}_{\text{flat}} \,|\, \mathcal{Y}_T) = \frac{p(\boldsymbol{\theta}_{\text{flat}} \,|\, \mathcal{Y}_T)}{p(\boldsymbol{\theta}_{\text{MAP}} \,|\, \mathcal{Y}_T)} < \psi < 1\,. \tag{3}$$

The threshold $\psi$ determines by how much the posterior mode at $\boldsymbol{\theta}_{\text{MAP}}$ decays before areas of $\boldsymbol{\Theta}$ are reached where the uniform prior dominates the likelihood. For example, one could subjectively define the posterior to be *sufficiently proper*, if the peak of the posterior/likelihood towers $\psi = 10^{-2}$ or more over the flat parts of the likelihood (in all directions in $\boldsymbol{\Theta}$). For non-uniform priors, one needs to refine definition (Eq. 3): The threshold should now measure, that the defined proportion $\psi$ of the posterior maximum has decayed before the slope of the posterior approaches the slope of the prior. That the slope of posterior approaches the one of the prior can be defined by a relative deviation $1 + \zeta$ of the gradients of posterior and prior:

$$\frac{||\nabla p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)^\top \mathbf{u}||}{||\nabla p(\boldsymbol{\theta})^\top \mathbf{u}||} < 1 + \zeta \tag{4}$$

where $\mathbf{u}$ is a direction in the parameter space with $||\mathbf{u}|| = 1$.
We will also use a much simpler approach to define *sufficiently proper* based on posterior samples, namely if the posterior mode carries most of the probability mass such that samples from the tail region hardly reach the limits of the sampling box. This indicates that probability mass in the tail regions is negligible relative to the probability mass under the posterior mode. Note that the density is only well-defined because we refer to a finite volume in the parameter space; otherwise, the posterior is improper and no longer a density.
We refer to an inference as fully Bayesian if $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T) \propto p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is calculated. See App. B for the analytical posterior of the beta-binomial model obtained with different priors. Unfortunately, there is typically no analytical form of $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ for more complex statistical models. In this case, one resorts to sampling from the posterior:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)\,. \tag{5}$$

Many different algorithms have been designed for that purpose. Bayesian inference then boils down to averaging over the posterior samples to approximate the mean value, $\mathbb{E}[\cdot]$, of some quantity $f$

$$\mathbb{E}_{p(\cdot \,|\, \mathcal{Y}_T)}[f] = \int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)\, d\boldsymbol{\theta} \approx \frac{1}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} f(\boldsymbol{\theta}_i), \tag{6}$$

where $N_{\text{samples}}$ is the number of samples. Interesting quantities are the mean parameter values, their covariances or the mean likelihood of the data. Our algorithm uses automatic differentiation of the statistical model to sample $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ with Hamiltonian Monte Carlo (HMC) [66, 67, 70] as provided by the Stan software [68, 69]. Sometimes, estimates are based on the prior $\mathbb{E}_p[f] = \int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ to express the uncertainty before the data have been seen. In addition to

the covariance matrix of the parameters, $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ allows the calculation of the credibility volume in order to assess parameter uncertainty. The smallest volume $V_P$ that encloses a probability mass $P \in [0, 1]$

$$P = \int_{V_P} p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)\, d\boldsymbol{\theta} \tag{7}$$

is called the Highest Density Credibility Volume/Interval (HDCV/HDCI). Credibility volumes should not be confused with confidence volumes, although they can be equivalent under certain conditions. Given that the model sufficiently captures the DGP, the actual values $\boldsymbol{\theta}_{\text{true}}$ will lie in the credibility volume with a probability $P$ as soon as the likelihood dominates the prior.

Bayesian inference is mathematically consistent, but conditional on the assumed prior and likelihood [82]. Altering $p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})$ as well as adding or removing a small amount of information in $p(\boldsymbol{\theta})$ (see App. B Fig. 10) can change $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ significantly. If little is known *a priori* about the parameter value and the data are suspected to constrain some parameters only vaguely, a MI prior is essential. It attempts to confine $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ within reasonable boundaries, while making it as sensitive to the data as possible. Unfortunately, MI priors are often improper themselves, which might also render $p(\boldsymbol{\theta} \,|\, \mathcal{Y}_T)$ improper if $p(\mathcal{Y}_T \,|\, \boldsymbol{\theta})$ is non-PI or even non-SI. The benefits and limits of MI priors in the presence of non-PI problems in CRNs are discussed below.

## 3 The true Markov kernel of a closed first-order chemical reaction network for identical and independent ion channels

Here, we summarize the assumptions made in [41] to account for the Markov dynamics of single molecules latent in the fluctuations of macroscopic signals. We assume a CRN with only mono-molecular or pseudo-mono-molecular chemical reactions, such as conformation dynamics of a protein or binding/unbinding transitions at excess ligand or protein concentration. If the CRN describes a single object, it can only be in one Markov state $s(t)$ at a time. To define the set of states $S$ of possible outcomes of a stochastic variable, one may use a one-hot encoding

$$S := \{(1, 0, 0, \ldots)^\top,\, (0, 1, 0, \ldots)^\top,\, \ldots,\, (\ldots, 0, 1)^\top\} \subset \{0,\, 1\}^M, \tag{8}$$

where $M$ is the number of states. By $\mathbf{s}(t) = \mathbf{e}_i$, we specify that the channel is in state $i$ at time $t$, where $\mathbf{e}_i$ stands for the $i$-th canonical unit vector. The Markov kernel of discrete state-to-state transitions can be described with a transition matrix $\mathbf{T}$ in discrete time,

$$\mathbf{T}_{ij} := \Pr(\mathbf{s}(t + \Delta t) = \mathbf{e}_i \mid \mathbf{s}(t) = \mathbf{e}_j) = [\exp(\mathbf{K} \cdot \Delta t)]_{ij}, \tag{9}$$

or with a rate matrix $\mathbf{K}$ in continuous time. Instead of parameterizing $\mathbf{K}$ by the chemical rates $k_{ij}$, one can use another parameterization of continuous time MMs

$$\mathbf{K} = \begin{bmatrix} -\sum_{i\neq 1} k_{i1} & k_{12} & \cdots & k_{1M} \\ k_{21} & -\sum_{i\neq 2} k_{i2} & \cdots & k_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ k_{M1} & k_{M2} & \cdots & -\sum_{i\neq M} k_{iM} \end{bmatrix} = \begin{bmatrix} -1/\tau_1 & \epsilon_{12}/\tau_2 & \cdots & \epsilon_{1M}/\tau_M \\ \epsilon_{21}/\tau_1 & -1/\tau_2 & \cdots & \epsilon_{2M}/\tau_M \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{M1}/\tau_1 & \epsilon_{M2}/\tau_2 & \cdots & -1/\tau_M \end{bmatrix}, \tag{10}$$

to facilitate the definition of a MI prior, by Eq. 21. The names of the parameters $\tau$ and $\epsilon$ are chosen such that they match the transformation behavior of the corresponding parameters in the general Eq. 21. By definition, each column of $\mathbf{K}$ sums to zero (Eq. 10), which encodes that the CRN is closed. The dwell times to remain in the $i$-th state are exponentially distributed with mean $\tau_i = (\sum_{j\neq i} k_{ji})^{-1}$. The parameters $\epsilon_{ij}$ denote the probability of transitioning from state $j$ to state $i$ after the random dwell time has passed. Thus, each chemical rate $k_{ji} = \tau_i^{-1}\epsilon_{ji}$ is the product of a probability with the inverse of the mean dwell time. By using the left-hand side of Eq. 10, for each column, we separated the inverse time-like (scale) parameter $\tau_i^{-1}$ from the probabilities $\epsilon_{ji}$, which are shape parameters. The $\epsilon_{ji}$ have no units unless the statistical weight corresponds to a ligand-dependent $k_{ji}$. Note, that since $\sum_j \epsilon_{ji} = 1$, the right-hand side (Eq. 10) has the same number of free parameters per column compared to the left-hand side. Transition probabilities are constrained to $0 \leq \epsilon_{ji} \leq 1$, such that as long as Haldane-like priors $\text{beta}(0, 0)$ are excluded (App. B), and the likelihood remains finite for all $\epsilon_{ji}$, $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ will be proper in these parameters. Then, only the dwell-time parameters $\tau_i^{-1}$ can render the HMM non-PI.

The Markov kernel of a single-molecule experiment contains $M$ different multinomial distributions for a CRN of $M$ states (App. C). Given that the system resides in a specific state at time $t$ the corresponding distribution from the

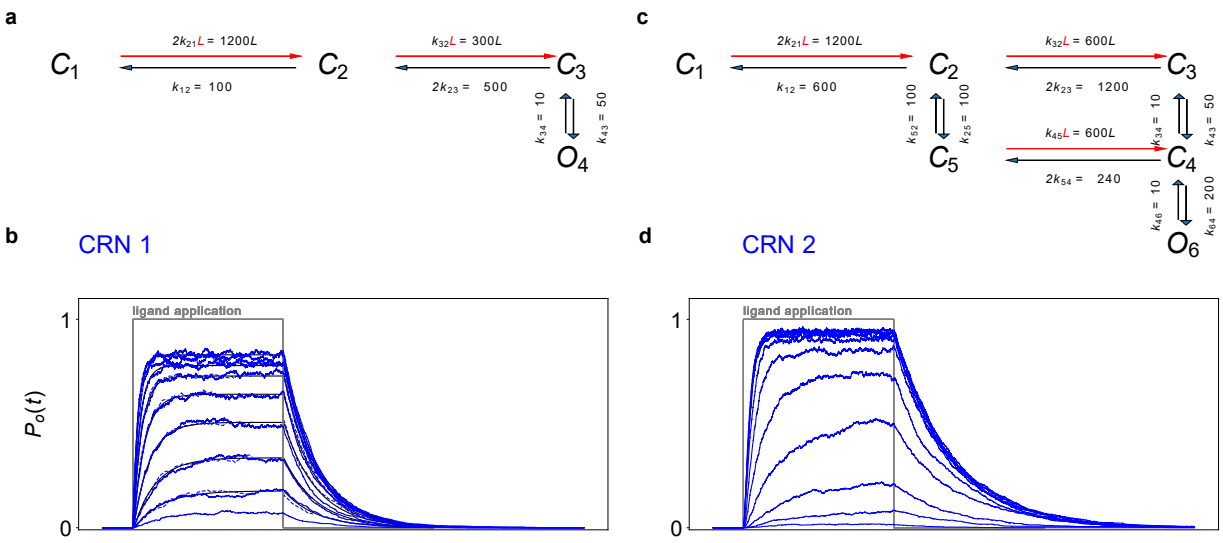


Figure 1: **Chemical reaction networks of LGICs and their simulated patch-clamp data. a,c** The Markov state models (one CRN per column) consist of three or five closed "$C_i$" and one open state "$O_i$". Binding steps (red arrows) have concentration-dependent rates. The CRNs are specified by the absolute rate constants $k_{ij}$, and $L$, the ligand concentration. Rates are given per subunit and stoichiometry factors are added to account for the number of subunits able to under go the respective transitions. The units of the rates are in $\mathrm{a.u.}$. To calculate their SI units $\mathrm{s}^{-1}$ and $\mu\mathrm{M}^{-1}\mathrm{s}^{-1}$, one needs to multiply their value by $6/7$ (Sec. 4.3.4). The open states $O_i$ conduct a mean single-channel current $i = 1$, and the closed states $C_i$ conduct a current of $i = 0$. The synthetic data were simulated with the Gillespie algorithm at a sampling rate of 10 ka.u.. The BF analysis frequency $f_{\mathrm{ana}}$ is 2 to 5 ka.u.. Since the units are in $\mathrm{a.u.}$., the ratios of rates or inverse dwell times determines the CRN. Further their ratio to the sampling frequencies determines, how detailed the kinetics are recorded. Similarly, the relative magnitude of $i$ compared with $\sigma_{\mathrm{op}}$ and $\sigma_{\mathrm{ex}}$ should be used to relate simulations to experimental conditions. **b,d**, Open probability time traces calculated from normalized currents $P_o(t) = \frac{y_I}{N_{\mathrm{ch}} i}$ of simulated relaxation experiments of ligand concentration jumps with $N_{\mathrm{ch}} = 10^3$ channels. For demonstration purposes, no experimental noise is added in this figure such that all fluctuations originate from Markov state transitions. However, when inferring the posteriors, additional experimental noise is added. The black lines are the theoretical open probabilities $P_{\mathrm{o}}(t)$ of the model. Typically, we used the set $\{0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 64\}$ $\mu\mathrm{M}$ of 10 ligand concentrations.

set of $M$ multinomial distribution samples the next state at $t + \Delta t$. However, the Markov kernel of an ensemble of ion channels is a convolution of these $M$ multinomial distributions [74, 41] if at a given time each of the $M$ states is occupied by at least one ion channel (App. C). In [41], we clarified the connection of the time evolution equations of the mean and covariance matrix of [40] to the convoluted multinomial Markov kernel. Fig. 1 shows the time traces of plausible CRNs of two LGICs having two binding pockets obeying the convoluted Markov kernel. These can be simulated with QuB [83] or an inhouse algorithm `https://cloudhsm.it-dlz.de/s/QB2pQQ7ycMXEitE`.

### 3.1 The assumptions of the Bayesian filter

The BF is a continuous state HMM with Gaussian transition probabilities [42]. So the true Markov dynamics on a lattice $\mathbf{n}(t) \in \mathbb{N}^M$ [84, 85], which is governed by the convoluted multinomial distributions, is replaced by an approximate Gaussian Markov process on a space of real-valued vectors, $\mathbf{n}(t) \in \mathbb{R}^M$:

$$\mathbf{n}(t + \Delta t) = \mathbf{T}\mathbf{n}(t) + \boldsymbol{\omega}(t) \sim \mathcal{N}(\cdot\,|\mathbf{T}\mathbf{n}(t), \mathbf{Q}(\mathbf{T}, \mathbf{n}(t))), \tag{11}$$

where $\mathcal{N}(\cdot\,|\mathbf{T}\mathbf{n}(t), \mathbf{Q}(\mathbf{T}, \mathbf{n}(t)))$ is a shorthand for the multivariate normal distribution with mean $\mathbf{T}\mathbf{n}(t)$ and covariance matrix $\mathbf{Q}(\mathbf{T}, \mathbf{n}(t))$. Essentially, the evolution of $\mathbf{n}(t)$ is now modeled by FPa of the CME [86]. Note that the conditional mean of $\mathbf{n}(t)$ evolves according to the equation $\mathbb{E}[\mathbf{n}(t + \Delta t)|\mathbf{n}(t)] = \mathbf{T}\mathbf{n}(t)$. It is perturbed by white noise $\boldsymbol{\omega}$ with the following properties: $\mathbb{E}[\boldsymbol{\omega}(t)] = 0$ and $\mathrm{cov}[\boldsymbol{\omega}(t), \boldsymbol{\omega}(t')] = \delta(t - t')\,\mathbf{Q}(\mathbf{T}, \mathbf{n}(t))$. In short, via the conditioning, Eq. 11 a Gaussian Markov process is defined [87].
From the insight that the true Markov kernel for $\mathbf{n}(t) \in \mathbb{N}_0^M$ is a generalized multinomial distribution, one can derive algebraically the conditional covariance [41] and employ it for the complimentary Gaussian Markov kernel (Eq. 11).

Hence, for $\mathbf{n}(t) \in \mathbb{R}_0^M$, we chose to match the mean and covariance of the true Markov kernel [41] with the mean and covariance of the approximate Markov kernel to define the implicit forward model BF.
The observations $\mathbf{y}(t)$, e.g. measured currents, depend linearly on the hidden state $\mathbf{n}(t)$:

$$\mathbf{y}(t) \sim \mathcal{N}(\,\cdot\,|\mathbf{H}\mathbf{n}(t), \mathbf{\Sigma}(t))\,. \tag{12}$$

The linear function is determined by an observation matrix $\mathbf{H}$, which projects the higher dimensional state $\mathbf{n}(t)$ to the lower dimensional observable. Hence, the CRN is only partially observed. The noise $\boldsymbol{\nu}$ of the measurement setup is modeled as a random perturbation of the mean observation vector. The noise fulfils $\mathbb{E}[\boldsymbol{\nu}] = 0$ and $\mathrm{cov}[\boldsymbol{\nu}_t, \boldsymbol{\nu}_t] = \mathbf{\Sigma}_t$. Equation 12 defines the state-conditioned observation distribution $\mathcal{N}(\,\cdot\,|\mathbf{H}\mathbf{n}(t), \mathbf{\Sigma}(t))$. Based on these assumptions, one can derive a likelihood for PC and confocal patch-clamp fluorometry (cPCF) data [88, 89, 90]. For the details of defining a Bayesian filter, i.e., a likelihood based on the FPa, refer to [40, 41]. For the necessary refinements to capture the statistics of PC and cPCF experiments and a benchmark against RE approaches, see [41].

# 4 Pathologies of the likelihood and how to cure them with minimal and vaguely informative priors

## 4.1 Structural and practical parameter non-identifiability of the likelihood

Asymptotically (assuming regularity conditions), reporting $\boldsymbol{\theta}_{\mathrm{ML}}$ and its corresponding parameter uncertainty prediction $\mathrm{cov}(\boldsymbol{\theta}_{\mathrm{ML}})$ are equivalent to the predictions of $p(\boldsymbol{\theta}\,|\,\mathcal{Y}_T)$ [50], for an infinite amount and quality of data $\mathcal{Y}_T$. The Bernstein-von Mises theorem [91, 50] proves this for regular statistical models. (Note that HMMs or BFs are singular [55] models and not regular.) The reason is that asymptotically $\boldsymbol{\theta}_{\mathrm{ML}}$ and $\mathrm{cov}(\boldsymbol{\theta}_{\mathrm{ML}})$ become the sufficient statistics of a multivariate normal sampling distribution $p(\boldsymbol{\theta}_{\mathrm{ML}})$ [54, 49], which is also the limiting posterior distribution $p(\boldsymbol{\theta}\,|\,\mathcal{Y}_T)$. A necessary assumption that this result holds asymptotically is that the corresponding $p(\mathcal{Y}_T\,|\,\boldsymbol{\theta})$ is structurally identifiable (SI) [92, 93, 94, 95, 96]. This means that $\boldsymbol{\theta}_{\mathrm{ML}}$ converges in distribution to $\boldsymbol{\theta}_{\mathrm{true}}$ with infinitely many data. A heuristic to numerically test for potential non-SI likelihoods is to check for the smallest eigenvalue of the Hessian matrix (curvature) of $p(\mathcal{Y}_T\,|\,\boldsymbol{\theta}_{\mathrm{ML}})$ [97]. If the smallest eigenvalue is effectively zero, then the likelihood is non-SI (or almost). Unfortunately, this is only a local criterion for SI that does not go beyond properties at $\boldsymbol{\theta}_{\mathrm{ML}}$ and its neighborhood. It indicates based on local information the least confined direction in parameter space. However the global shape of the likelihood might be more complex [98, 96]. In contrast, non-SI models are characterized by submanifolds in $\mathbf{\Theta}$ in which the likelihood is constant, even with infinitely many data [92, 99]. An example of a non-SI likelihood is an RE model of the linear birth-death process

$$\mathbb{E}[n_{\mathrm{bakt}}(t)] = n_0 \exp(\xi t) = n_0 \exp((k_{\mathrm{birth}} - k_{\mathrm{death}})t) \tag{13}$$

describing the number of bacteria $n_{\mathrm{bakt}}(t)$ in a well-stirred petri dish. All parameters $\boldsymbol{\theta} = (k_{\mathrm{birth}}, k_{\mathrm{death}})^\top$ with the same difference $\xi(k_{\mathrm{birth}}, k_{\mathrm{death}}) = k_{\mathrm{birth}} - k_{\mathrm{death}}$ result in the same $\mathbb{E}[n_{\mathrm{bakt}}(t)]$. This model is non-SI because we cannot disentangle $k_{\mathrm{birth}}$ and $k_{\mathrm{death}}$ based on $\mathbb{E}[n_{\mathrm{bakt}}(t)]$ alone. In other words, a likelihood derived only from $\mathbb{E}[n_{\mathrm{bakt}}(t)]$ has the symmetry $p(\mathcal{Y}_T\,|\,\boldsymbol{\theta}) = p(\mathcal{Y}_T\,|\,\xi)$ which means that the likelihood is flat, $p(\mathcal{Y}_T\,|\,\boldsymbol{\theta}) = \mathrm{const}$, along straight lines in $\mathbf{\Theta}$ with intercept $\xi$. However, adding the information contained in $\mathrm{var}[n_{\mathrm{bakt}}(t)]$, which can be derived from the CME [100], renders the model SI. Hence, there is independent information in the higher-order statistical moments, rendering the linear birth-death model SI. Thus, non-SI can artificially be caused by likelihoods that ignore higher statistical moments of the DGP. A similar manifestation of this insight is the common misconception that equilibrium data of CRNs do not allow the inference of chemical rates. This is true for deterministic RE-based likelihoods but [101]. However, stochastic descriptions such as Kalman filters [40] or Bayesian filters [41] allow one to infer chemical rates from equilibrium time traces.
Being SI is necessary but not sufficient for a meaningful ML inference. Bayesian inference, on the other hand, only requires that the posterior is proper (App. E) such that SI is not necessary. Nevertheless, a non-SI likelihood still creates inferential limits. Since only finite amounts of data are available in real-world applications, the HMM must also be practically identifiable (PI) for ML inferences; it would be better to say that it must be *sufficient* PI. Let us clarify what we mean by this: The literature offers different definitions of PI vs. non-PI models. Here, we follow the definitions of [95], as they describe a malfunction of $p(\mathcal{Y}_T\,|\,\boldsymbol{\theta})$ that can strongly influence the posterior (Eq. 1) as it induces parts in $\mathbf{\Theta}$ where the posterior is dominated by the prior. Likelihoods of non-PI models do not decay to zero (Fig. 2, blue curve for $\theta > \theta_{\max}$), but stretch out infinitely in regions of $\mathbf{\Theta}$ (in one or multiple dimensions) for any finite amount of data [95]. This can already happen in simple, partially observed CRNs.
Let us assume a fictitious measurement setup that can record the occupation $n_O(t)$ of state $\mathbf{O}$ at any frequency and without any measurement noise such that, for the sake of argument, the additional complications arising from noisy data are avoided. As an example for a non-PI likelihood consider a CRN with a rate $k_{\mathrm{BA}} \cdot L$ that depends on ligand concentration $L$ (or any other stimulus-dependent rate):

$$\mathbf{A} \xrightarrow{k_{\mathrm{BA}} \cdot L} \mathbf{B} \xrightarrow{k_{\mathrm{OB}}} \mathbf{O} \tag{14}$$

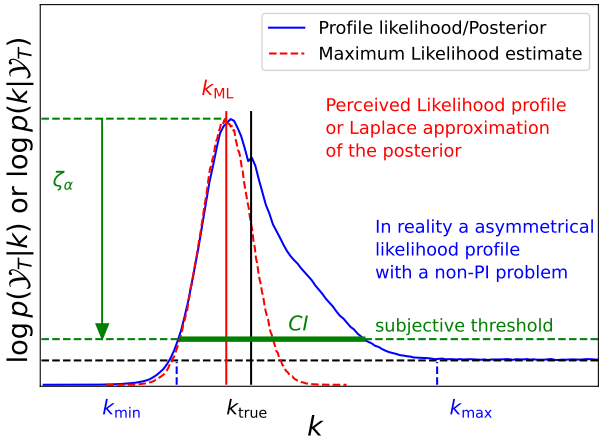


Figure 2: **Sketch illustrating practical parameter non-identifiability (non-PI): The severeness of non-PI depends on the relative height of the peak compared to the non-vanishing tails.** We assume that there is only a single parameter that we want to infer, e.g., a chemical rate $k$. The red dashed line shows a Laplace approximation of the posterior (Gaussian approximation with a mean at the peak location and a precision equal to the observed Fisher information. Note that the Laplace approximation underscores the uncertainty (quicker decay of the Laplace approximation), particularly for $k > k_{ML}$. The likelihood function $p(\mathcal{Y}_T \mid k)$ which is proportional to the posterior $p(k \mid \mathcal{Y}_T)$ (for a uniform prior) indicates the continuous aspect of the non-PI pathology. The maximum value of $p(\mathcal{Y}_T \mid k)$ relative to the constant value in the tails specifies the degree of PI. In likelihood-based inference, we classify a set of parameters as PI, if the value of $p(\mathcal{Y}_T \mid k)$ drops below a subjectively defined threshold (green) at some distance to the location of the maximum such that $p(\mathcal{Y}_T \mid k_{\mathrm{ML}})/p(\mathcal{Y}_T \mid k_{\mathrm{const}}) > p(\mathcal{Y}_T \mid k_{\mathrm{ML}})/|\mathrm{thres}|$, where $|\mathrm{thres}|$ is the threshold. Note that a prediction of the uncertainty based on the curvature at $k_{\mathrm{ML}}$ cannot detect these shortcomings, because any function can be approximated with a second-order Taylor expansion (parabola) around the extreme value. The question of whether ML works well for the given data and model boils down to the question if the quadratic approximation of $\log p(\mathcal{Y}_T \mid k)$ (red dashed line) is still a good approximation at $k \gg k_{\mathrm{ML}}$. Typically, using the curvature at $k_{\mathrm{ML}}$ underestimates the uncertainty and assumes a symmetric distribution.

for a finite number of channels. Assuming the initial condition $\mathbf{n}(t_0) = (n_A, 0, 0)$ and that the experimental signal obeys $y(t) \propto \mathbb{E}[n_O(t)]$, then the part of the general solution of the RE which is probed by the signal is

$$\begin{aligned}\mathbb{E}[n_O(t)] =& \frac{k_{\mathrm{OB}}}{(k_{\mathrm{BA}} \cdot L - k_{\mathrm{OB}})} n_A \exp(-k_{\mathrm{BA}} \cdot L \cdot t) \\ &- \frac{k_{\mathrm{BA}} \cdot L}{(k_{\mathrm{BA}} \cdot L - k_{\mathrm{OB}})} n_A \exp(-k_{\mathrm{OB}} t) + n_A.\end{aligned} \tag{15}$$

Note that the solution of the CME for the initial condition $\mathbf{n}(t = 0) = (n_A, 0, 0)$ is a multinomial distribution. To understand this, consider the case that only one molecule is in state $\mathbf{A}$ at $t = 0$. For all $t > 0$ the molecule has probability $p_O(t)$ to be in $\mathbf{O}$, $p_B(t)$ to be in $\mathbf{B}$ and $p_A(t)$ to remain in $\mathbf{A}$. If, instead, one has $n_A$ independent and identical molecules in state $\mathbf{A}$ at $t = 0$, then each of them individually has the same probability $p_O(t)$, $p_B(t)$ and $p_A(t)$ at $t > 0$. Hence, the distribution over all states is a multinomial distribution which evolves in time (see arguments in App. D). If only state $\mathbf{O}$ is observed, one can reduce the problem to

$$n_O(t) \sim \mathrm{binomial}(p_O(t), n_A) \text{ with } p_O(t) = \frac{\mathbb{E}[n_O(t)]}{n_A}. \tag{16}$$

If none of the rates could be manipulated $k_{\mathrm{BA}}(L) = \mathrm{const}$, one would face a non-SI problem, too. However, for a ligand-dependent rate $k_{BA} \cdot L$, assuming that the process can be measured for at least two values of $L$ such that $k_{BA} \cdot L_1$ and $k_{BA} \cdot L_2$ and $k_{OB}$ are all different but of similar magnitudes, we show the CRN given the experiment is still non-PI. The non-PI problem originates from the following phenomenon: Any combination of values for $k_{\mathrm{BA}} \cdot L_i$ and $k_{\mathrm{OB}}$ which fulfill $k_{\mathrm{BA}} \cdot L_i \gg k_{\mathrm{OB}}$ or $k_{\mathrm{OB}} \gg k_{\mathrm{BA}} \cdot L_i$ will push the amplitude of one of the two exponential decays to zero. If the process is measured at two largely different ligand concentrations, e.g. $L_2 = L_1/10$, then the condition is satisfied for ten times larger $k_{\mathrm{BA}}$. Thus, minor changes in $k_{\mathrm{BA}}$ or $k_{\mathrm{OB}}$ will hardly affect $\mathbb{E}[n_O(t)]$ as long as $k_{\mathrm{BA}} \cdot L \gg k_{\mathrm{OB}}$ or $k_{\mathrm{OB}} \gg k_{\mathrm{BA}} \cdot L$ holds for all $L$. Note that the correct solution of the CME is a multinomial distribution given the initial conditions. Adding information about the entire distribution of $n(t)$ such as variance or skewness, etc., will not resolve

the non-PI problem caused by the vanishing amplitude of one of the exponential decays. However, the multinomial model would improve the accuracy and quality of the uncertainty quantification. This example is reminiscent of the common scenario for coupled CRNs in which only $k_j$ can be inferred if $k_{i,\mathrm{true}} \gg k_{j,\mathrm{true}}$. However, here we are not referring to this scenario in which the DGP's signal is rate-limited. Instead, we assume that at least three different rates are at play, $k_{\mathrm{BA}}(L_1)$, $k_{\mathrm{BA}}(L_2)$, and $k_{\mathrm{OB}}$, which have similar magnitude but non-identical values, such that rate limiting does not exist at least for one combination of $k_{\mathrm{BA}}(L_i)$ and $k_{\mathrm{OB}}$ in the DGP. Nevertheless there are regions in parameter space (far away from the true parameter values) in which one or the other rate is rate-limiting. The CRN and the linear projection governing the signal have the potential to be rate-limited independently of the true parameters of the DGP. Thus, when analyzing data with a model such as $n_O(t) \sim \mathrm{binomial}(p_O(t), n_A(0))$, the rate-limiting effects in regions where $k_i \gg k_j$ holds will make the likelihood essentially a non-zero constant, hence non-PI. See App. F.1 for a discussion on the effects if state $\mathbf{A}$ or $\mathbf{B}$ is observed alone or simultaneously.

The degree by which non-PI affects the parameter inference depends on the intrinsic properties of the model and on the specifics of the experimental data such as the signal-to-noise ratio or the bandwidth. In App. G, we discuss the global sensitivity of the solution of the RE for CRN1 (Fig. 1 **a**) and demonstrate non-PI of the likelihood (App. G.3) even for over-optimistic data and prior knowledge (that only one rate constant is unknown).

SI is a dichotomous feature (a model is either SI or not), whereas PI is gradual (continuous) [102]. To classify models into PI and non-PI, we use the confidence interval CI based on confidence level $\zeta_\alpha$:

$$\mathrm{CI} = \{\boldsymbol{\theta} \mid \log(p(\mathcal{Y}_T \mid \boldsymbol{\theta}_{\mathrm{ML}})/p(\mathcal{Y}_T \mid \boldsymbol{\theta})) < \zeta_\alpha\} \ . \tag{17}$$

We call a model PI if the confidence interval CI is a compact set. The subjectively defined threshold $\zeta_\alpha$ (Fig. 2) [100] should be larger than the asymptotic value of the likelihood $p(\mathcal{Y}_T \mid \boldsymbol{\theta}_{\mathrm{flat}})$ (Fig. 2, black dashed line). In a multiparameter inference, there will be multiple asymptotic values for the different directions in $\boldsymbol{\Theta}$. To assess the degree of PI, $p(\mathcal{Y}_T \mid \boldsymbol{\theta}_{\mathrm{ML}})$ has to exceed a subjectively defined level (green dashed line) which should be larger than the maximum of the asymptotic likelihood values. The interval $[k_{\min}, k_{\max}]$ where the dashed black lines cross the likelihood profile (blue lines) (Fig. 2) is the largest asymmetric confidence interval that can be deduced by using the profile likelihood technique. Above $k_{\max}$, the data contain no information to distinguish values of $k$. However, for values $k < k_{\max}$ and even $k < k_{\min}$ the data contain information. Often, the profile of $p(\mathcal{Y}_T \mid \boldsymbol{\theta})$ might approach its non-zero constant only asymptotically. Defining subjectively a threshold relative to the maximum value of the likelihood [95] is equivalent to subjectively defining a significance level [**?**]. Note that ML does not infer the shape of $p(\mathcal{Y}_T \mid \boldsymbol{\theta})$ globally (Fig. 2 red dashed curve); ML estimates the shape based on the curvature at $\boldsymbol{\theta}_{\mathrm{ML}}$, which is a local property. Therefore, standard ML does not detect PI problems (Fig. 2 red dashed curve).

Due to the challenges indicated by the two toy examples, HMM inference based on ML has several drawbacks compared to sampling from the posterior $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$. First, one must take extra precautions against (non-SI/non-PI) pathologies of $p(\mathcal{Y}_T \mid \boldsymbol{\theta})$ as ML does not reveal them [95]. Second, even if the model is SI and sufficiently PI, the quality and quantity of the data are often insufficient to meet the implicit assumption that $p(\boldsymbol{\theta}_{\mathrm{ML}})$ can be approximated by a normal distribution, which justifies the use of ML.

Strategies such as profile likelihood [95] aim to detect (non-SI/non-PI) pathologies. If $p(\mathcal{Y}_T \mid \boldsymbol{\theta})$ is SI and has only a small degree of non-PI, the profile likelihood method can construct non-symmetric confidence intervals extending the reliability compared to standard ML. Nevertheless, it still relies on the assumption of asymptotically large amount and quality of data [**?**]. In contrast, $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$ displays non-PI/non-SI problems in HMM inferences, if $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$ is scrutinized in detail. Notably, $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$ will be dominated by the prior $p(\mathbf{K})$ in some part of $\boldsymbol{\Theta}$ and will become improper for improper priors. Fortunately, Bayesian statistics provides us with tools to render $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$ proper (or sufficiently proper) even with non-PI likelihoods such as theoretically derived upper limits on parameters or vaguely informative assumptions about cooperativity incorporated in $p(\mathbf{K})$. Incorporating these assumptions renders $p(\boldsymbol{\theta} \mid \mathcal{Y}_T)$ proper for both test systems (Fig. 1) even with very poor data. The benefits and limits of MI priors, vaguely informative assumptions, and theoretically derived upper limits in the presence of non-PI problems in CRNs are discussed below.

### 4.2 Defining and benchmarking a minimally informative (MI) prior by an approximation of the Jeffreys rule

The following section benchmarks the performance of the BF for different combinations of MI priors and physically motivated vaguely informative priors, for cases where the information content of the data is low relative to the number of parameters of the CRN. As long as computational time is not a limiting factor, we argue that Bayesian inference is superior to ML, because it does not rely on the validity of asymptotic theorems about $\boldsymbol{\theta}_{\mathrm{ML}}$ for finite data to express uncertainty of the model parameters. To support our argument, we first compare a uniform prior on the transition rates $p(\mathbf{K})$ with a MI prior defined below (Eq. 23), which promises to be less biased (less misinformed). The reason is that the non-PI of the likelihood (App. G.1) is aggravated with a uniform $p(\mathbf{K})$. In an unfortunate combination, a uniform prior $p(\mathbf{K})$ places the probability mass where the likelihood becomes less and less pronounced and reaches a constant finite value (App. G.1). With a MI prior, however, one can drastically reduce the severeness of this problem. But one should be aware of the limitations of both priors. Also the MI prior cannot completely resolve non-PI of

model parameters. This holds true, particularly, for non-PI arising from large $\tau_i^{-1}$ in any realistic macroscopic HMM likelihood even without experimental noise (App. G.1).

#### 4.2.1 Nonlinear coordinate transformations alter the functional form of a prior.

To demonstrate potential problems of uniform priors in combination with weakly informative data and to contrast them against the benefits of a MI prior, a definition of how to construct a MI prior, $p(\mathbf{K})$, is required. We want the posterior to be as sensitive to the data as possible when little is known about some parameters. At first glance, a uniform prior seems to be the least biased choice. However, uniformity of a distribution is not invariant under non-linear parameter transformations. Therefore, a uniform prior is not in general MI, it often induces biases [62]. For example, one can parameterize a model of a monomolecular chemical reaction in terms of either the rate constant $k$ or the activation energy $E_A$. Their relationship is often approximated by Arrhenius's law $k = A(T) \cdot \exp(\frac{-E_A}{RT})$, where $R$ is the gas constant, $T$ the temperature and $A$ some temperature-dependent proportionality constant. Let's define $k_{\max,\text{box}}$ as the upper limit of the sampling box. What shape has the prior $p_{E_A}(E_A)$ in the parameterization of $E_A$, if we assume $p_k(k) = \text{uni}(0, k_{\max,\text{box}})$ in the $k$ parameterization and $A(T)$ and $T$ to be known? It turns out that $E_A$ would be exponentially distributed. However, if one assumes a log-uniform distribution $p_k(k) \sim \frac{1}{k}$, then $E_A$ would be uniformly distributed.

#### 4.2.2 Definition of minimally informative (MI) priors by approximating Jeffreys's rule

Parameters of many probability distributions can be classified into location, scale and shape parameters based on their behavior under transformation groups. The probability density $p_s$ of a scaling parameter $s$ obeys $p_s(y) = p(y/s)/s$ . Typical examples of scaling parameters are the standard deviation $\sigma$ of a normal distribution or the mean dwell time $\tau$ to remain in a specific Markov state. Note that dwell times $t_{\text{dwell}} \sim 1/\tau \exp(-t_{\text{dwell}}/\tau)$ are exponentially distributed. Various notions of "minimal information" have been proposed [61], which can result in different priors. For example, Jaynes [61] and Jeffreys [63] argued for a log-uniform prior for a scaling parameter:

$$p(\tau) \propto \frac{1}{\tau}. \tag{18}$$

Jeffreys's arguments are mainly based on a specific nonlinear coordinate transformation. He argues that the log-uniform prior is invariant under power transformations, meaning that any $u = \tau^\xi$ with arbitrary exponent $\xi \neq 0$ has the same functional form

$$p(u) \propto \frac{1}{u}\,. \tag{19}$$

For example, a MI statement about $\sigma$ should be also MI for the variance $\sigma^2$. Applied to the parameters of CRNs, the power transformation $k = \tau^{-1}$ also identifies a chemical rate constant as a rate parameter. Note that here the term "rate parameter" should be understood in a statistical sense which defines a "rate parameter" as the inverse of a scaling parameter. Thus, the functional form of a MI prior for $k$ remains a log-uniform distribution. Jaynes's arguments [61] are based on transformation groups and lead to Eq. 18, whereas Jeffreys's original rule yields a different result. In his original formulation [62], Jeffreys defines the prior by

$$p(\boldsymbol{\theta}) \propto \sqrt{\det[\mathbf{F}(\boldsymbol{\theta})]} \tag{20}$$

where $F_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log p(y \,|\, \boldsymbol{\theta})\right]$ is the Fisher information matrix. Jeffreys's prior creates a posterior that transforms covariantly under reparameterizations of the model, meaning that the same probability will be assigned to identical volumes in parameter space independent of the parametrization. Despite its success in single-parameter problems, Jeffreys's rule (Eq. 20) is challenged by multiple inconveniences in multi-parameter inferences [103, 104]. For example, Jeffreys's rule (Eq. 20) applied to a Normal distribution model with unknown mean $\mu$ and variance $\sigma^2$ results in a prior $p(\mu, \sigma) \propto \sigma^{-2}$, contradicting Eq. 18 of [61] but also the revised version of Eq. 20 [63]. If, on the other hand, each parameter is treated independently, $p(\theta_i) \propto \sqrt{|F_{ii}(\theta)|}$, assuming that the others are known, then we obtain the log-uniform prior $p(\mu, \sigma) \propto \sigma^{-1}$. Thus, we will use a revised version Eq. 20 in the following, which treats location, scale and shape parameters independently, as suggested by Jeffreys himself [63] and consistent with [61]:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\epsilon}) \propto \sqrt{\det[\mathbf{F}(\boldsymbol{\epsilon})]} \prod_i \frac{1}{\tau_i}. \tag{21}$$

The location parameters, $\boldsymbol{\mu}$, such as the mean value of the normal distribution, are assigned uniform priors. Each scaling parameter, $\boldsymbol{\tau}$, has a log-uniform prior. Only the shape parameters $\boldsymbol{\epsilon}$ are treated conjointly according to Jeffreys's rule [63]. In addition, we simplify Eq.21 by assuming that $\sqrt{\det[F(\boldsymbol{\epsilon})]}$ can be applied to each column of $\mathbf{K}$ (Eq. 10)

independently. In that way, we obtain closed-form solutions of $\sqrt{\det[F(\boldsymbol{\epsilon})]}$ derived from simpler statistical models for the remaining $\boldsymbol{\epsilon}$ of each column of $\mathbf{K}$. To further substantiate these arguments and study the impact of MI priors, we discuss three different concepts (Laplace principle of insufficient reason, Jeffreys's rule (Eq. 21), and Haldane's prior) to derive MI priors for a binomial likelihood (Eq. 28) in App. B.

#### 4.2.3 Minimally informative prior for the rate matrix inspired by Jeffreys's rule

We use a simplification of Eq. 21 that is common practice [18] for complex multi-parameter MMs. The priors used to infer MMs from MD simulations are constructed by applying Jeffreys's rule to simpler statistical models [18] instead of applying it to the entire model. Bayesian estimation of $\mathbf{T}$ (Eq. 9) from MD simulations often uses one Dirichlet prior per column $\mathbf{T}_{i,:}$ of the transition matrix:

$$\mathrm{dir}(\mathbf{T}_{i,:}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{j=1}^{M} T_{ij}^{\alpha_i - 1} \tag{22}$$

which facilitates the computation [105]. However, the Jeffreys prior for $\mathbf{T}$ is not a product of Dirichlet distributions [106, 107]. Also, in the case of HMMs of single-molecule force spectroscopy experiments [17], products of Dirichlet distributions are used as prior for $\mathbf{T}$.
Here, we do not sample $\mathbf{T}$ but $\mathbf{K}$ (Eq. 9) because, in contrast to sampling $\mathbf{T}$, it is trivial to incorporate information about the scaling of the binding rates at different ligand concentrations in a direct parameterization of $\mathbf{K}$. The same applies to additional prior information on maximal binding rates. Note that with one exception, we use the parameterization of the right-hand side of Eq. 10 to define $p(\mathbf{K})$. The exception will be discussed later when we add the information about theoretical upper diffusion limits on binding rates $k_{21}$ and $k_{32}$ (CRN1 and CRN2) and $k_{45}$ (CRN2). One can mix in the parameterization any $k_{ji}$, $\tau_i$ and $\epsilon_{ji}$ equivalently as long as the pior remains equivalent, i.e., that $p(\hat{\boldsymbol{\theta}})d\hat{\boldsymbol{\theta}} = p(\boldsymbol{\theta})d\boldsymbol{\theta}$ holds, with $\hat{\boldsymbol{\theta}}$ indicating a different parameterization. As the mean dwell times $\tau$ are scaling parameters, we use a log-uniform prior for $\tau_i^{-1}$, following the arguments above. We use the Dirichlet prior for the probabilities $\epsilon_{ji}$, which Markov transition is chosen (Eq. 9), which is the default prior for probability vectors. Applying Eq. 20 to a multinomial likelihood, results in a Dirichlet prior with $\alpha_j = 0.5$ for all $j \in [1, \cdots, W_i]$ parameters. The value of $W_i$ is the number of Markov transitions leaving the $i$-th state. For a physically meaningful CRN there is some sparsity in $\mathbf{K}$, e.g., ligand binding and channel gating does not occur at the exact same instant of time. Thus for the $i$-th Markov state, the number of allowed transitions will usually satisfy $W_i < M$ where $M$ is the total number of states. The set $\Omega_i \subseteq \{1, \ldots, M\} \setminus \{i\}$ contains the indices of all states that can reached by one Markov transition, leaving the $i$-th state. Then, using Eq. 21 and evaluating the $\sqrt{\det[F(\boldsymbol{\epsilon})]}$ factor for each column of $\mathbf{K}$ individually, leads to

$$p(\mathbf{K}) = \prod_{i}^{M} \text{log-uni}(\tau_i^{-1})\, \mathrm{dir}_i(\boldsymbol{\epsilon}_i|\boldsymbol{\alpha}) = \prod_{i}^{M} \frac{1}{\tau_i^{-1} \log[b_i/a_i]} \frac{1}{\mathrm{B}(\boldsymbol{\alpha}_i)} \prod_{j \in \Omega_i} \epsilon_{ji}^{\alpha_{ji}-1}, \tag{23}$$

where $\boldsymbol{\alpha}$ is the vector of concentration parameters and $\mathrm{B}(\boldsymbol{\alpha})$ is the beta function at $\boldsymbol{\alpha}$. The set of $a_i$ and $b_i$ for the log uniform distributions are the upper and lower limits of $\tau_i$. The topology of the CRN is controlled by the Dirichlet distribution. See App. F for an illustration of $p(\mathbf{K})$ for a state with two or three leaving transitions.

### 4.3 Advantages and limits of the minimally informative prior in the presence of non-PI

Using the BF, we study the impact of three different priors (Fig. 3) on the performance of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, focusing on weakly informative PC data.

#### 4.3.1 The scaling of the RSME of the posterior derived from different priors

We define the Euclidean distance or root-mean-square error (RMSE) between $\mathbf{K}_{\mathrm{true}}$ and the posterior mean of all chemical rates in log-space as

$$\mathrm{Error}\,(\log \mathbb{E}[\boldsymbol{k}]) = \sqrt{\sum_{l} (\log(\mathbb{E}[k_l]/k_{l,\mathrm{true}}))^2}. \tag{24}$$

We discuss why we are using the logarithm of the chemical rates in App. H. In Fig. 3**a**, $\mathrm{Error}(\log \mathbb{E}[\boldsymbol{k}])$ of the inferred $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ is plotted vs. the number of ion channels per time trace $N_{\mathrm{ch}}$. We use $N_{ch}$ as a proxy for the data quality or information content in the data. The PC data are sampled from a 4-states-1-conducting-state HMM (CRN1 Fig. 1**a**). One intuitively expects a $1/\sqrt{N_{\mathrm{ch}}}$-scaling for $\mathrm{Error}(\log \mathbb{E}[\boldsymbol{k}])$, but this has only been proven for regular statistical

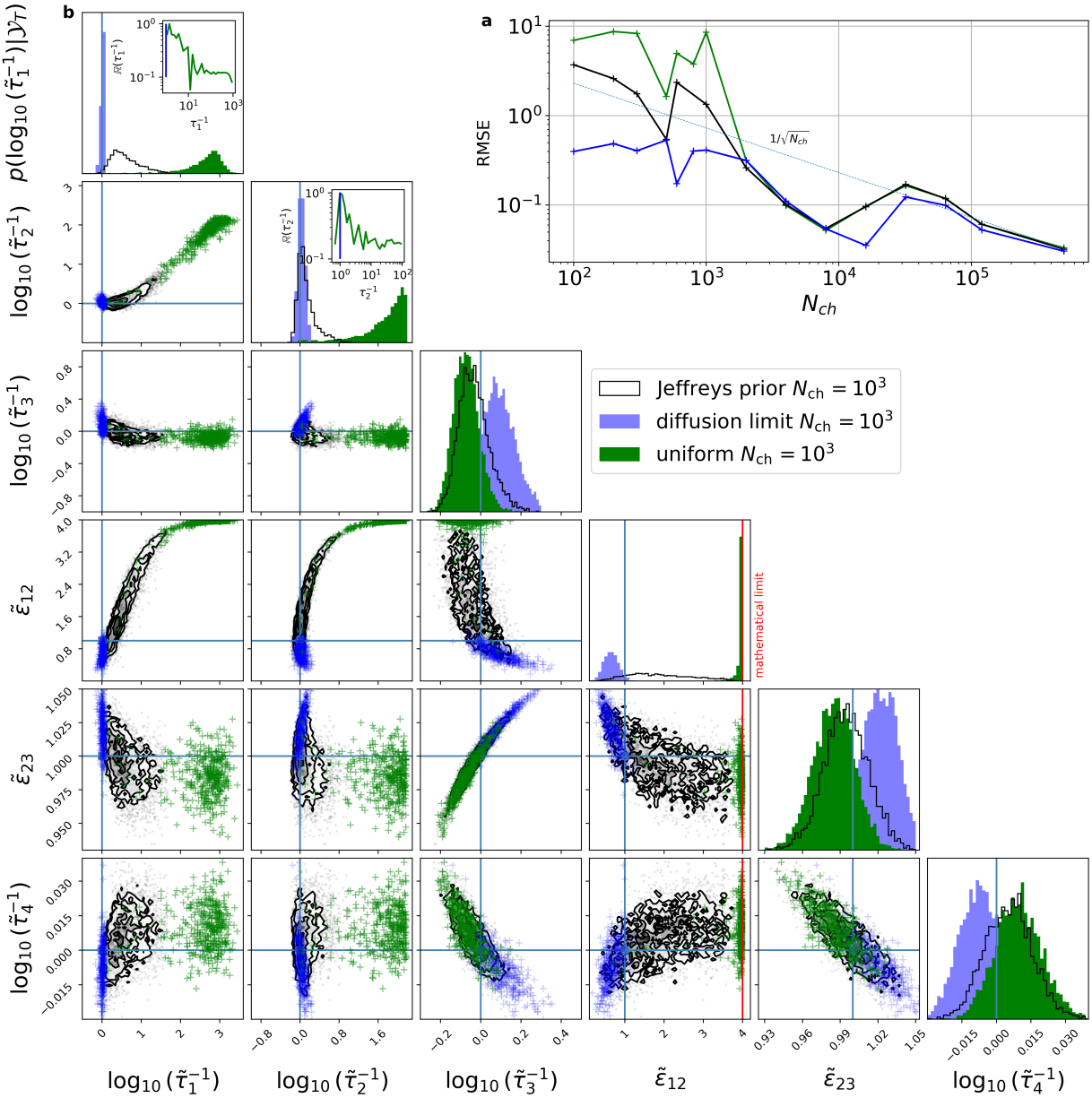


Figure 3: **Prior sensitivity analysis showing that the uniform prior aggravates problems caused by non-PI.** The figure contrasts the posteriors resulting from three different priors: uniform prior (green), MI prior (black), and MI prior with physical limits on the binding rates (blue). The rates are transformed to inverse dwell times and transition probabilities. **a**, RSME of the mean of the marginal distribution vs. $N_{\mathrm{ch}}$. The improper uniform prior (green) induces a significant bias for $N_{\mathrm{ch}} < 2 \cdot 10^3$. The MI prior restricts the RSME to be in a $1/\sqrt{N_{\mathrm{ch}}}$ regime, even though the log-uniform distribution is still improper. Imposing diffusion limits (blue) on the two binding-related rates shrinks the RSME even below $1/\sqrt{N_{\mathrm{ch}}}$ for $N_{\mathrm{ch}} < 2000$. The cyan dashed curve is a fit based on $f(N_{\mathrm{ch}}) = a/\sqrt{N_{\mathrm{ch}}}$. A standard deviation of each data point which follows $\propto \sqrt{N_{\mathrm{ch}}}$ is assumed. All data sets were generated with $\sigma^2 = 1$ and $\sigma_{\mathrm{op}}^2 = 0.2$. **b**, Posterior distributions of the dwell times times $\tau_i$ and transition probabilities $\epsilon_{ji}$ for $N_{\mathrm{ch}} = 10^3$. On the diagonal, 1D marginal posteriors are plotted. The off-diagonal plots show the 2D marginal distributions of the posterior. All samples of the parameters were normalized to their true values $\tilde{\theta} = \theta/\theta_{\mathrm{true}}$. The blue lines indicate the true parameters. The insets on the diagonal show the same histogram only plotted on the logarithmic scale to display the flatness of the posterior if a uniform prior is used. The (red) vertical bar indicates $\epsilon_{12} = 1$, which corresponds to $\tilde{\epsilon}_{12} = 4$ due to the normalization by the true values. The posterior is plotted based on samples with a Rubin-Gelman statistic of $\hat{R} = 1.0$. The number of independent draws from the posterior satisfies $N_{\mathrm{independent,samples}} < N_{\mathrm{samples}}$ due to correlations between the samples.

models, to the best of our knowledge. However, BFs are singular statistical models. Indeed, the error obtained with a uniform prior (green curve) switches below a critical data quality $N_{\mathrm{ch,crit}} \approx 2 \cdot 10^3$ to a regime where the posterior mean deviates visually discernible from the behavior of the other priors (black and green curves). It will become clear in the discussion of Fig. 3**b** that the data are insufficient to overcome the bias of the uniform prior toward shorter time scales. The following discussion will show that $N_{\mathrm{ch,crit}}$ is not a universal data quality/quantity. For a given model and a certain type of experiment, e.g., PC experiments, $N_{\mathrm{ch,crit}}$ depends on the limits of the sampling box and the prior distribution. In the biased regime, $N_{\mathrm{ch}} < N_{\mathrm{ch,crit}}$, the uniform prior causes $\mathrm{Error}(\log \mathbb{E}[\boldsymbol{k}])$ (green curve) to deviate from the $1/\sqrt{N_{\mathrm{ch}}}$-scaling (Fig. 3**a**, black dashed line). In contrast, below $N_{\mathrm{ch,crit}}$ the stronger penalty for higher $\tau_i^{-1}$ from the log-uniform prior compared to the uniform shrinks $\mathrm{Error}(\log \mathbb{E}[\boldsymbol{k}])$ (Fig. 3**a**, black curve) such that it scales with $1/\sqrt{N_{\mathrm{ch}}}$ (black dashed curve) to far lower data qualities. Above $N_{\mathrm{ch,crit}}$, the BF behaves like a regular statistical model, and the MI prior extends that behavior to smaller $N_{\mathrm{ch}}$ by an order of magnitude.
Deviations from $1/\sqrt{N_{\mathrm{ch}}}$-scaling of the RMSE to smaller values occur (blue line) if upper bounds derived from prior knowledge are enforced on some $k_{ij}$. The log transformation eliminates the lower limits (App. H), but the additional upper bounds derived from the diffusion limit (Sec. 4.3.4) are still active. These limits reduce the RMSE for some $N_{\mathrm{ch}}$, even at $N_{\mathrm{ch}}$ that are two orders of magnitude larger than $N_{\mathrm{ch,crit}}$. Note that the RMSE is dominated by the error in $\tau_1^{-1}$ whose uncertainty is reduced most strongly by the constraints (Fig. 16). The additional upper limits are further discussed below. The next section explains why the uniform prior causes $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ to deviate from the $1/\sqrt{N_{\mathrm{ch}}}$-scaling to larger RSMEs compared to the MI prior below $N_{\mathrm{ch,crit}}$.

#### 4.3.2 The likelihood dominates the posterior only in a small region of the parameter space

The impact of the prior for a specific data set with $N_{\mathrm{ch}} = 10^3$ just below $N_{\mathrm{ch,crit}}$ is explored in Fig. 3**b**. The notation $\tilde{\boldsymbol{\theta}}$ indicates that the samples are divided by their true value. For each parameter $\tilde{\tau}_i^{-1}$, the posterior of $\log_{10}(\tilde{\tau}_i^{-1})$ is plotted to display the bias induced by the uniform prior (green). The green samples of $\log_{10}(\tilde{\tau}_1^{-1})$ concentrate between $10^2$ and $10^3$ and would deviate even more strongly from $\log_{10}(\tilde{\tau}_{1,\mathrm{true}}^{-1}) = 0$, if the sampling box was larger. The insets in the diagonal panels (Fig. 3**b**) of the first two columns show the same posterior but without the $\log_{10}$ transformation of the samples. Instead, a log scale of the x-coordinate is used to overview the full sampling box. To indicate the relative probability masses, the posterior ratios $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T) := p(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)/p(\tilde{\tau}_{1,\max}^{-1} \,|\, \mathcal{Y}_T)$ and $\mathbb{R}(\tilde{\tau}_2^{-1} \,|\, \mathcal{Y}_T) := p(\tilde{\tau}_2^{-1} \,|\, \mathcal{Y}_T)/p(\tilde{\tau}_{2,\max}^{-1} \,|\, \mathcal{Y}_T)$ are plotted. The flat right tail of the marginal posterior obtained with the uniform prior (Fig. 3**b**, green posterior) causes the observed deviations of the RSME (Fig. 3**a**, green curve and Fig. 15). Furthermore, consistent with the preceding discussion (Sec. 4.2.1) about the transformation of probability densities, a flat prior-dominated part of $p(\tilde{\tau}_2^{-1} \,|\, \mathcal{Y}_T)$ creates for $p(\log_{10}(\tilde{\tau}_2^{-1}) \,|\, \mathcal{Y}_T)$ an exponentially growing part. $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ and $\mathbb{R}(\tilde{\tau}_2^{-1} \,|\, \mathcal{Y}_T)$ drop from their peak value to their right tail (Fig. 3**b**, insets) only by a fraction in the order of $\approx 0.1$ before the derivative of the marginal posterior becomes zero or effectively zero. It takes $\tau_2^{-1}$ two orders and $\tau_1^{-1}$ one order of magnitude to do so. Hence, the posteriors, $p(\log_{10}(\tilde{\tau}_1) \,|\, \mathcal{Y}_T)$ and $p(\log_{10}(\tilde{\tau}_2) \,|\, \mathcal{Y}_T)$, are dominated by the uniform prior, and it is plausible that the entire $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ is too.
Given that the sampling box covers multiple orders of magnitude for $\tau_1^{-1}$ and $\tau_2^{-1}$, the flat part of the marginal posterior (where the data contain no information to discriminate different values of those parameters) contributes a relevant probability mass to the posterior. Thus, the posterior's statistics, such as mean, median, variance, and derived quantities, such as the ED, become sensitive to the amount of probability mass from the flat part. That should raise one's concern as the results depend on changes in the most often arbitrarily chosen artificial limits of the sampling box. A larger sampling box would diminish the probability mass around the peak even more relative to the probability mass in the flat area. Hence, the nonphysical arbitrary chosen limits of the uniform prior (the sampling box) for $\tau_1^{-1}$ and $\tau_2^{-1}$ are highly informative limits for the posterior in that regime. Note that the HMC sampler of the used Stan programming language internally samples on an unconstrained $\boldsymbol{\theta}$ derived by coordinate transformations. Hence, a malfunctioning sampler near the boundary cannot explain the observation.
Instead, if $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ were to drop till $10^{-4}$ before the marginal posterior has a zero derivative, changes in limits would not affect $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ on a relevant scale unless one increases the sampling box by many orders of magnitude. Without any upper limit for $\tau_i$, one cannot normalize the posterior (Eq. 2); i.e., the posterior is improper. Thus, even though one attains a maximum of the posterior, one cannot define HDCIs to quantify uncertainty if we allow those parameters to exist on an open interval $[0, \infty)$. Inevitably, one needs to recruit more information, which happens most of the time accidentally by setting the sampling box limits. Alternatively, a less informative prior will make $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ more sensitive to the data as demonstrated below.

#### 4.3.3 The MI prior (partially) alleviates the trouble of the non-PI problem of the likelihood for the posterior

Comparing $p(\mathbf{K}\,|\,\mathcal{Y}_T)$ derived from the MI prior (Fig. 3, black) with $p(\mathbf{K}\,|\,\mathcal{Y}_T)$ derived from the uniform prior exemplifies the harm induced by the uniform prior. The MI prior (black) Eq. 23 , i.e., penalizes larger values of all $\tau_i$ (black) and decorrelates $p(\log_{10}(\tau_2^{-1}), \epsilon_{12}\,|\,\mathcal{Y}_T)$. Note that $\tau_2^{-1}$ and $\epsilon_{12}$ belong to the same $\mathbf{K}$ column, describing transitions leaving state $C_2$. This concentrates (Fig. 3**b**) $p(\tau_1\,|\,\mathcal{Y}_T)$, $p(\tau_2^{-1}\,|\,\mathcal{Y}_T)$ and $p(\epsilon_{12}\,|\,\mathcal{Y}_T)$ over the true parameter values in contrast to the posterior (green) based on the uniform prior. Consistently, the prior choice is most important for $\tau_1$, $\tau_2$ and $\epsilon_{12}$ which are much less determined by the data than the parameters $\tau_3^{-1}$, $\epsilon_{23}$ and $\tau_4^{-1}$. Overall, the MI prior concentrates $p(\mathbf{K}\,|\,\mathcal{Y}_T)$ closer to $\mathbf{K}_{\mathrm{true}}$.
Thus, for the RSME vs. $N_{\mathrm{ch}}$, the MI prior reduces the RSME drastically below $N = 2 * 10^3$ (Fig. 3**a**) such that the $1/\sqrt{N_{\mathrm{ch}}}$ regime continues. The $1/\sqrt{N_{\mathrm{ch}}}$ scaling is unfortunate if one tries to increase the inference quality by measuring additional similar data sets. However, the $1/\sqrt{N_{\mathrm{ch}}}$ scaling is a desirable feature in the reverse direction of lower $N_{\mathrm{ch}}$, i.e., data quality or quantity, as it prevents inferences from becoming meaningless. Despite the heuristic definition, the prolonged $1/\sqrt{N_{\mathrm{ch}}}$-regime indicates that with this MI prior, one is much closer to an unbiased and minimally informative prior than the uniform prior is. $p(\mathbf{K}\,|\,\mathcal{Y}_T)$ is more sensitive to the data. This can be seen (Fig. 3**a**) by $p(\log_{10}(\tilde{\tau}_1^{-1})\,|\,\mathcal{Y}_T)$ and $p(\log_{10}(\tilde{\tau}_2^{-1})\,|\,\mathcal{Y}_T)$. The probability mass grows exponentially (green posterior) with the distance to the true values given the uniform prior. This exponential growth is the uniform prior transformed to the log-parameter space. We only see the prior and hardly any effect of the data. The area around the true values has essentially no probability mass. However, the minimal informative prior concentrates the probability mass much closer to $\mathbf{K}_{\mathrm{true}}$. Nevertheless, if the likelihood is flat in the directions of $\tau_1^{-1}$ and $\tau_2^{-1}$ in some distance to the peak, it follows that $p(\tau_1^{-1}\,|\,\mathcal{Y}_T)$ and $p(\tau_2^{-1}\,|\,\mathcal{Y}_T)$ are improper. This does not change by equipping $p(\mathbf{K}\,|\,\mathcal{Y}_T)$ with the MI prior because the log-uniform prior cannot be normalized if used on the entire positive real axis. It diverges to infinity at $\tau_i \to 0$ and decays too slowly to zero for $\tau_i \to \infty$. The article returns to this question with this prior setup in sec. 4.4.
Cutting $\boldsymbol{\theta}$ with the sampling box into sampled and non-sampled parts is not always a problem. Without knowing about non-PI problems, we showed in [41] that the HDCVs and, thus, the peak have a frequentist interpretation as long as the peak of the posterior is towering highly enough over the flat parts. Thus, one finds $\mathbf{K}_{\mathrm{true}}$ inside a volume with a certain probability mass with a frequency approximately equal to the probability [41].

#### 4.3.4 Adding information from physically motivated upper limits for binding rates

So far, we have shown that using a MI prior robustifies the inference below $N_{\mathrm{ch,crit}}$. To alleviate the persisting problem that the posterior potentially is improper or is vague for some parameters, one can add more physically motivated information, such as the diffusion limit for the binding rates (Fig. 1, red arrows). Typically the random collision rate is used as upper limits for binding rates [108] here we use a more realistic estimate for small ligands binding to binding-pockets [109]. Notably, taking full advantage of upper limits of parameters is to the best of our knowledge unique to Bayesian statistics because adding some upper limits to a sampling distribution of $\boldsymbol{\theta}_{\mathrm{ML}}$, which might without the constraint be approximated by a normal distribution, reduces the validity of the normal approximation. To implement the diffusion limit consistently in a MI way (besides the new information about the diffusion limit itself), we needed to change the definition of the prior Eq. 23. The rate holds $k_{32} = \epsilon_{32}/\tau_2$ such that we can also draw $k_{32} \sim \text{log-uniform}$. So, the log-uniform prior is still used to set the time scale for all transitions leaving state $C_2$. Then, we draw the statistical weights from the Dirichlet distribution for the second column of $\mathbf{K}$ such that $\tau_2^{-1} = k_{32}/\epsilon_{32}$. The other rates can be defined by $k_{j2} = \tau_2^{-1} \cdot \epsilon_{j2}$. It is irrelevant for which parameter $\tau_2^{-1}$ or some $k_{j2}$ the log-uniform distribution is used as a prior as long as it is evoked once per column of $\mathbf{K}$.
We simulated the first binding rate (Model 1) with $k_{21} = 2 \cdot 600\ \mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ which is the diffusion limit derived in [109] for ligand binding. The 2 is the stoichiometric factor, incorporating the structural information that two free binding pockets are available. Of course the binding rate of a real process will be slower than the upper limit from [109]. To incorporate this aspect we take advantage of the fact that in our simulation of data and inference $k_{21} = 2 \cdot 600\ \mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ is an arbitrary definition. In fact, we simulated $k_{21} = 2 \cdot 600$ a. u. and defined this value to represent the value in SI units. We can also apply a different mapping of the arbitrary units to SI units. We may decide that 700 a. u. is supposed to represent the diffusion limit [109] in SI units. Thus we defined $700 a.u. = 600\ \mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$. In that way we avoid the unrealistic and extreme scenario that $k_{21,\mathrm{true}}$ is identical to the sampling box limit (which would still be for Bayesian inference a valid use case). All other "time-like" parameters such as sampling rates, dwell times and chemical rates need to rescaled by $6/7 \frac{\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}}{a.u.}$.
The consequences of the additional information induced by the upper limits are shown in Fig. 3**a-b** and Fig. 4 by the blue error curve and posterior. The upper limits of the sampling box for the parameters $k_{21,\max}$ and $k_{32,\max}$ now reduced to much smaller maximal values than the arbitrarily set sampling box limits used previously. This immediately solves the pressing non-normalizability problem in the crucial $\tau_1^{-1} = k_{21}$ and reduces for $\tau_2^{-1} = k_{32}/(1-\epsilon_{32})$ the possibilities to be non-normalizable. Still, $\tau_2^{-1}$ could diverge if the posterior for $\epsilon_{32} \to 1$ diverges. The parameters $\tau_1^{-1}$

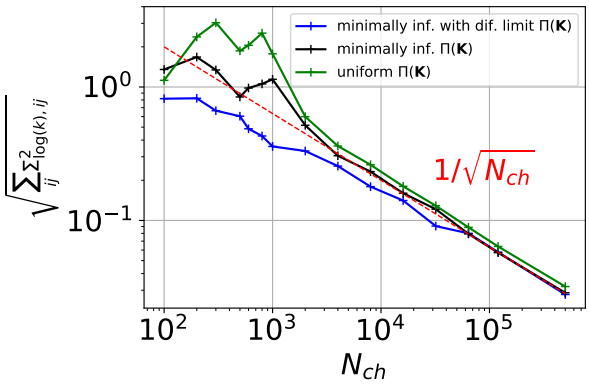


Figure 4: The uniform prior results in posteriors with a higher uncertainty. We define a multivariate analog of the standard deviation (Frobenius norm of the covariance matrix of the posterior) vs. $N_{\text{ch}}$. The colors encode the prior assumptions. The larger the Frobenius norm, the more uncertainty remains after the inference. Plot is based on the same data used in Fig. 3.

and $k_{32}$ are then bounded from above and below by mathematical and physically motivated limits.
The defined diffusion limit $k_{21} < 2 \cdot 700$ a. u. and $k_{32} < 700$ a. u. constrains the RSME (Fig. 3**a**, blue lines) to smaller values such that the RSME drops below the $1/\sqrt{N_{\text{ch}}}$ regime for $N_{\text{ch}} < N_{\text{ch,crit}}$. However, even for data sets with $N_{\text{ch}} \in [10^4, 10^5]$, the constraint adds information and decreases the ED. Constraining the two binding rates $\tau_1^{-1}$ and $k_{32}$ also influences via the likelihood the marginal posteriors of other parameters, particularly $p(\epsilon_{21} \,|\, \mathcal{Y}_T)$ (Fig. 3**b**). The constraints shape $p(\epsilon_{21} \,|\, \mathcal{Y}_T)$ into a more peaking distribution which covers the true value.

#### 4.3.5 The uniform prior increases the uncertainty expressed by the posterior

Considering only RSME is too limited to demonstrate how long the constraint still adds relevant information to $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ (i.e., that $\mathcal{Y}_T$ are not informative enough to make the information of the prior's upper limit irrelevant). Notably, $\text{Error}(\log \mathbb{E}[\boldsymbol{k}])$ is not any quantity possible to report after an inference of real experimental data either. Ultimately, one judges the inference by its credibility intervals/regions (or confidence intervals/regions in an ML context) of the inferred parameters. Thus, only the posterior's shape in general (the median/mean/peak, covariance, and higher order statistical moments describing the tails) are at the modeler's disposal to assess the quality of the inference. Therefore, we define a quantitative measure of the spread of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, which is the Frobenius norm

$$||\boldsymbol{\Sigma}_{\log(k)}|| = \sqrt{\sum_{i,j}^{N_{\text{params}}} \boldsymbol{\Sigma}^2_{\log(k),ij}}, \tag{25}$$

of the covariance matrix of the samples of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ on the log space of the chemical rates. The Frobenius norm is a generalization of the Euclidean norm for matrices. The smaller the variance and of each parameter posterior or the covariance between each pair of parameters gets, the smaller becomes Eq. 25. The observed transition around $N_{\text{ch,crit}}$ (Fig. 3**a**), that the location posterior derived from the uniform prior becomes equivalent to the posterior derived from Jeffreys prior without diffusion limit at $N_{\text{ch}} = 2 \cdot 10^3$, as juged by the RSME, is only to some degree present in $||\boldsymbol{\Sigma}_{\log(k)}||$ (Fig. 4, black and green lines), which as noted measures the posterior's spread not location. The spread of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ based on the uniform prior (green curve) does not converge to the spread of the MI prior (black curve). It shrinks instead parallel given the $N_{\text{ch}}$ range that is investigated. This effect becomes more prominent for the more complex CRN. Further, the larger the size of the sampling box the larger $||\boldsymbol{\Sigma}_{\log(k)}||$ due to the non-PI problem. The Frobenius norm of the covariance of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ based on the Jeffreys prior, including the diffusion limit (blue curve), has prolonged smaller values than without the diffusion limit but finally converges towards the Frobenius norm of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ derived from the prior without upper limits (black curve). Hence, the information from the added diffusion limit is made use of, even up to $N_{\text{ch}} = 3 \cdot 10^4$.
Above $N_{\text{ch,crit}}$ the RSME (Fig. 3**a**) is still a much more fluctuating parameter to benchmark the behavior of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ than the rather non-fluctuating $||\boldsymbol{\Sigma}_{\log(k)}||$. $||\boldsymbol{\Sigma}_{\log(k)}||$ follows a straight $1/\sqrt{N_{\text{ch}}}$-scaling (Fig. 4). In other words, repeating the experiment under identical conditions will leave $||\boldsymbol{\Sigma}_{\log(k)}||$ almost unchanged, but the posterior's location will be randomly translocated from data set to data set within the soft constraints of $\boldsymbol{\Sigma}_{\log(k)}$. Larger simulation boxes

increase the value of the green curve the most, followed by a minor effect on the black curve. This is an elegant result because, ultimately, with real experimental data, the inference quality is judged by uncertainty quantification and not by the RSME.

### 4.4 The MI prior still generates an improper posterior.

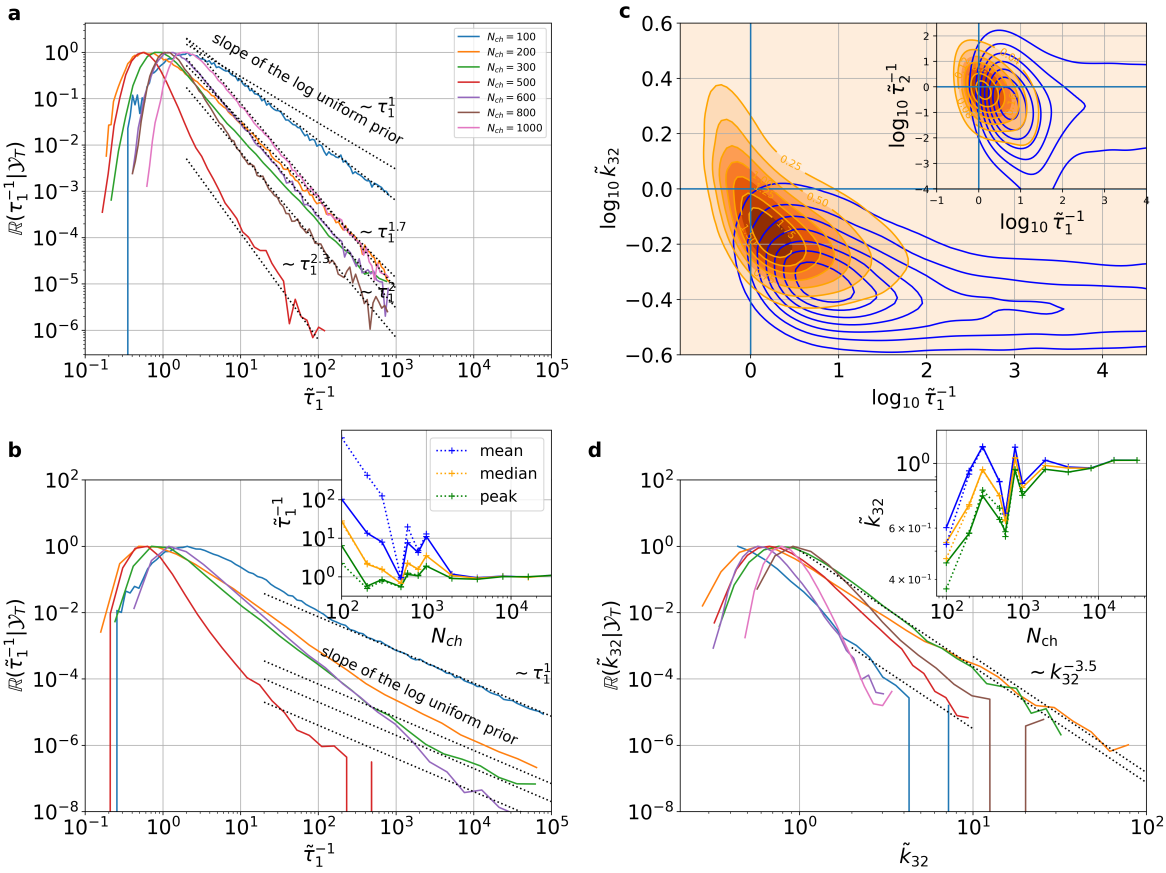


Figure 5: **For the weakly identified parameters of the inference of CRN1 (Fig. 3 and 4), the MI prior without diffusion limit generates still improper posteriors.** The relative marginal posteriors $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ and $\mathbb{R}(\tilde{k}_{32} \,|\, \mathcal{Y}_T)$ are plotted for different $N_{\rm ch}$ and different sampling boxes. The tilde over the parameters indicates that each parameter is normalized to its true value. **a** $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ displays power law-like behavior, but appears proper given the limits of the sampling box. The dashed lines indicate the limiting negative slope of the log uniform prior and of steeper power laws as a guide to the eye. **b,** However, the same $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ sampled from a two magnitudes larger sampling box displays an area for large $\tilde{\tau}_1^{-1}$ where the prior entirely dominates $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$. The likelihood is locally flat (as the likelihood is not altering the marginal posterior); in other words, the model has some degree of a non-PI problem. The inset shows the mean, median, and peak for $\tau_1^{-1}$ vs. $N_{\rm ch}$ for the smaller sampling box (solid line) and larger sampling box (dotted) line. The higher the data quality, the more robust the estimates from all sampling boxes become. **c,** $p(\tilde{\tau}_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$ for $N_{\rm ch} \in \{100, 200\}$ (blue, red densities) displays a non-local correlation structure leading to a bias of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$. The inset is based on the same data $N_{\rm ch} \in \{100, 200\}$ but $k_{32}$ is transformed to $\tilde{\tau}_2 = k_{32}/(\epsilon_{32})$. While the peak might even roughly be at the same position the overall bias highly reduced for $\tilde{\tau}_2$. **d,** For $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ the heavy tails are much less a concern. The power law exponent for the smallest $N_{\rm ch}$ is $n \approx 3.5$. Notably, the mode and the highest density area of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ are biased towards too small values, indicating that the coverage frequency of small HDCVs [41] needs higher $N_{\rm ch}$ to be established. The insets of the panels display the mean (blue), the median (orange), and the peak (green) of the marginal posterior.

To confirm that $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ is still improper (Eq. 2) and where in the parameter space $p(\mathbf{K})$ dominates $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, the most worrying $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ and $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ are plotted for $N_{\text{ch}} < 2 \cdot 10^3$ and sampled in larger sampling boxes (Fig. 5). The PC data originates from the 4-states-1-open-state CRN (Fig. 1 **a**). Note that $\tau_1^{-1} = k_{21}$ because only one Markov transition leaves state $C_1$.
As a guide to the eye, functions $f(\tau^{-1}) \propto (\tau^{-1})^{-\phi}$ (dashed lines, Fig. 5**a-b,d**) that have a power law scaling are plotted, including the log-uniform prior. The right tails of $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ are well approximated by a power law until the limit, $10^3\tau_{1,\text{true}}$, of the sampling box (Fig. 5**a**). The slopes of the tails are heavily influenced by the log uniform prior $f(\tau^{-1}) \propto (\tau_1^{-1})^{-1}$ but also the likelihood contributes some additional slope (in other words information) till approximately $\tau_{1,\text{true}}^{-1}10^3$. For the smallest sampling box, for all values of $N_{\text{ch}}$ the $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ seem to be proper but only vaguely identified. The exponent $\phi$ gradually increases with $N_{\text{ch}}$ indicating the increase of information about $\tau_1^{-1}$. However, using a two-orders of-magnitude larger sampling box (Fig. 5**b**) demonstrates that for even larger $\tau_1^{-1}$, $\phi = 1$ holds. I.e. the prior complete dominates $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ in this region of $\boldsymbol{\Theta}$ in the direction of $\tau_1^{-1}$. Hence, $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ indicates a non-PI problem of the HMM as onyl the prior and not the data or likelihood contributes to the posterior in this region. For at least $N_{\text{ch}} \leq 600$ the likelihood is essentially flat for $\tilde{\tau}_1^{-1} > 10^3$. The data do not contribute significant new information in this region of $\boldsymbol{\theta}$. The limited influence of some chemical rates or dwell times in some parts of $\boldsymbol{\theta}$ on the signal is a general feature of partially observed CRNs (Sec. 4.1 and App. G). In the inset of Fig. 5**b**, the mean (blue), the median (orange), and the peak (green) of $p(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ vs. $N_{\text{ch}}$ for the smaller sampling box (solid) and larger simulation boxes (dashed) are plotted. Mean values or higher statistical moments do not exist for distributions with power law tails with exponents $\phi < 2$, thus the peak should be reported for the barely identified $\tau_1^{-1}$. Consistently, the peak of $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ (green curve inset Fig. 5 **b**) has maximally a relative error of a factor of 2 and seems unbiased and not diverging with increasing sampling box size, given the tested data and MI prior. For the same data, the residuals of the mean (blue curve inset Fig. 5 **b**) of $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$ are heavily biased, sensitive to the sampling box limits and orders of magnitude away from $\tilde{\tau}_1^{-1,\text{true}}$. The median (orange curve inset Fig. 5 **b**) is more robust against different sampling box sizes. However, it is still biased towards too large values of $\tau_1^{-1}$, which is unsurprising as it is not defined on the full support $(0, \infty)$ if the shape of the log uniform prior [110] eventually dominates $p(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ for large $\tau_1^{-1}$. This potentially indicates for parameters with strong non-PI degree (the magnitude of the posterior peak does not decay by multiple decades before reaching the area where the MI prior dominates the posterior) that it is better to report peak values. In Fig 5**c** we plot $p(\tilde{\tau}_1, \tilde{k}_{23} \,|\, \mathcal{Y}_T)$ and in the inset $p(\tilde{\tau}_1^{-1}, \tilde{\tau}_2^{-1} \,|\, \mathcal{Y}_T)$ to demonstrate that the non-PI $\tau_1^{-1}$ likely causes bias $k_{32}$ however the bias is much reduced if not unpresent for $\tau_2^{-1}$. The bias of the corresponding $p(k_{32} \,|\, \mathcal{Y}_T)$ is also present for different data sets (inset of Fig. 5**d**).
The tails of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ given the sampling box, seem to be less heavy tailed (Fig. 5**d**) than $\mathbb{R}(\tilde{\tau}_1^{-1} \,|\, \mathcal{Y}_T)$. For $N_{\text{ch}} =\in \{100, 200, 300\}$, $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ appears to follow a power law, hence, $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T) \sim k_{32}^{-\phi}$ with $\phi = 3.5$. The quicker decaying power-law tails of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ still create a skewed distribution. The skewed tail of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ compensates partially for the bias of the peak region of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$ (Fig. 5**d**). Hence, $\mathbb{E}[k_{32}]$ (blue curve, inset, Fig. 5**d**) is the least biased point estimator towards too small $k_{32}$ values until the data are strong enough. In contrast to $\tau_1^{-1}$, the standard Bayesian point estimate, $\mathbb{E}[k_{32}]$ performs best for $k_{32}$. Besides our observation, mathematics dictates that reporting $\theta_{\text{MAP}}$ for parameters whose posterior has a powerlaw tail with small $\phi \leq 2$ is more robust because mean values and eventually medians for $\phi \leq 1$ do not exist. However, it is unclear if reporting the mean value for parameters with large $\phi$ because of the smallest bias (inset of Fig. 5**d**). generalizes to other CRNs. Note that from the bias of the peak of $\mathbb{R}(k_{32} \,|\, \mathcal{Y}_T)$, it is clear that small HDCIs around the peak will not include the $\mathbf{K}_{\text{true}}$ with the frequency equal to the probability mass they claim to have. Thus, posteriors which fullfill decently the asymptotic behaviour of the Berstein-von-Mises theorem need larger $N_{\text{ch}}$ or a second observable [41].
In summary, the MI prior, particularly, its convex log uniform distribution for $\tau_i$'s or $k_{ij}$'s has the desirable feature of concentrating $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ much closer to $\mathbf{K}_{\text{true}}$, but it still produces improper $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ if no further upper limits can be justified. The MI prior alleviates the improperness of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ by making the posterior less sensitive to the often nonphysical and arbitrary limits of the sampling box, but the non-PI problem will become relevant when increasing the sampling boxes at some point for all data sets. Hence, the degree of the non-PI problem has to be judged based on how much the peak of $\mathbb{R}(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ decays before the slope of $\mathbb{R}(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ is essentially the slope of the log uniform prior.

### 4.5 Solving the improperness and non-PI problem with vague additional information on the ratios of rates as e.g. assumption on cooperativity

We exemplify how to robustify $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, such that the mean and the median of parameters with a non-PI problem have a meaning. For CRN1, we enforce a soft physical constraint on the one hand on the binding rates $k_{21}$ and $k_{32}$ and on the other hand on the unbinding rates $k_{12}$ and $k_{23}$ by a regularizing prior, plausible for homomeric proteins.

Physical common sense dictates that one should be skeptical *a priori*, if binding rates or unbinding rates of a HMM have by orders of magnitude differing values. One modeling assumption encoding this skepticism could be

$$p_{\text{cop}}\left(\log_{10}\left(k_{21}/k_{32}\right)\right) = \text{normal}\left(\mathbb{E}[\log_{10}\left(k_{21}/k_{32}\right)] = 0, \sigma_{\text{cop}}\right) \tag{26}$$

for binding and for unbinding

$$p_{\text{cop}}\left(\log_{10}\left(k_{23}/k_{12}\right)\right) = \text{normal}\left(\mathbb{E}[\log_{10}\left(k_{23}/k_{12}\right)] = 0, \sigma_{\text{cop}}\right) \tag{27}$$

with some finite standard deviation $\sigma_{\text{cop}}$ (Fig. 6). Note that this corresponds to the classical definition of cooperativity: The ration of $K_{32}$ to $K_{21}$: If the affinity $K_{ij} = k_{ij}/k_{ji}$ increases with the number of occupied binding sites, it is called positive cooperativity. I.e. if the microscopic binding rates(the binding rates per binding site) are constant, e.g. diffusion limited, the ratio $r = K_{32}/K_{21} = k_{12}/k_{23}$ can be used as measure of cooperativity with $= 1$ equals no cooperativity, $> 1$ is positive, $< 1$ negative cooperativity. The vague prior is a much less radical prior (assumption) than assuming identical microscopic rates (which is the non-cooperative assumption, $\sigma_{\text{cop}} = 0$, for the binding and unbinding), as frequently done for ligand gated [111, 112] and more excessively for voltage gated ion channels [113, 114, 115, 116] to alleviate non-SI or non-PI problems in HMM inferences. The CRN of the shaker channel [114] is a good example of how the structural prior information of four subunits within the shaker channel implies a necessary complexity of the CRN to allow it to explain the data and to be physically interpretable. This CRN has that many voltage depended rates that setting certain subsets of rates equal is necessary to avoid non-SI/PI problems. However, assuming that cooperativity cannot change the chemical rates beyond some reasonable scale is physically plausible. One might debate the prior's variance and the prior's shape. Note that adding this additional regularization prior solves the improperness problem of $p(\tau_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$ originating from a flat likelihood for high values of $\tau_1^{-1}, k_{32}$ no matter what finite $\sigma_{\text{cop}}$ is used because of the quickly decaying tails of normal distributions. Notably, any additional prior that adds at least a tiny amount of decay till infinity to the otherwise log uniform dominated posterior would also render $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ proper. The normal distribution on the logspace has, by definition, desirable properties: It is symmetrical on the order of magnitude, and defines an area, within one standard deviation, with little impact on $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ and an area of increasing penalty for values of multiple standard deviations. In that way extreme conclusions from the data, expressing strong cooperativity effects in the DPG, have to be more and more supported by the data to be represented with a relevant magnitude in $p(\mathbf{K} \,|\, \mathcal{Y}_T)$. In Fig. 6 panel a,b, we demonstrate how applying the additional prior renders $p(\tau_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$ immediately proper even for $N_{\text{ch}} = 10^2$, even for $\sigma_{\text{cop}} = 1$ i.e. $\pm\sigma_{\text{cop}}$ covers 2 oders of magitude. The bias (inset, Fig, 6**a**) of $p(k_{32} \,|\, \mathcal{Y}_T)$ is drastically reduced with a finite $\sigma$ compared to the pure MI prior, and decreasing $\sigma$ further improves the inference (Fig. 6**b**) in terms of bias and variance. Notably, the prior forces the posterior statistically to concentrate its mass between $\tau_{1,\text{true}}^{-1}$ and $k_{32,true}$. Thus, at some point, the smaller $\sigma_{\text{cop}}$ of Eq. 26 and 27, the more is the variance of the posterior decreased but also the bias of $p(\log_{10}\left(k_{23}/k_{12}\right))$ starts to dominate the posterior more and more. The traditional non-cooperativity assumption trades variance of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ in for a maximum of bias and thus should only be applied if there is strong *a priori* evidence that non-cooperativity is true in the DGP.
The reason why the bias is reduced is that the global correlation of $p(\tau_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$ is changed with the additional information from the prior (Fig. 6**c**). The ridge-like structure of $p(\tau_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$ towards $\tau_1^{-1} \to \infty$ does not exist anymore.
One should keep in mind that decreasing $N_{\text{ch}}$ not only lowers the information content of the data but also makes the normal approximation of the generalized multinomial dynamics less accurate.

### 4.6 Increasing the complexity of the HMM. The minimal informative prior battles the curse of complexity and the curse of dimensionality which also helps the sampler to converge

Next, we investigate the challenges arising when the complexity of the CRN increases, hence, a more complex $\mathbf{K}$ to infer. CRN2 additionally contains two flip states (Fig. 1 **d**). We enforce microscopic reversibility, which reduces the number of to-be-inferred parameters by one. For a comment on how the MI prior improves the convergence of the sampler to the true posterior (Ap. I.0.1) and alleviates *the curse of dimensionality*, see App. J. In Fig. 7 **a** the pathological dimensions of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ derived from the uniform $p(\mathbf{K})$ (green) against the same dimensions of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ (black) derived from the MI $p(\mathbf{K})$ are compared. The less pathological dimensions, in the sense that they deliver roughly Gaussian marginal posteriors, are plotted in Fig. 7 **b1-6**. For the uniform prior, the chemical rates are sampled, and then the samples are mapped to the parameters of the MI prior. Even for $N_{\text{ch}} = 10^5$, the posterior based on the uniform prior clearly demonstrates the non-PI feature of the likelihood in $\tau_2^{-1}$, $\tau_3^{-1}$. The ridge-like correlation of $p(\tau_2^{-1}, \tau_3^{-1})$ (panel **a3**) with a slow decay along the ridge demonstrates the non-PI problem. Notably, the ridge is strongly constraint for values smaller than $\tau_{2,true}$ and $\tau_{2,true}$ (panel **a1-2**) but above it seems to extend to infinity. The strong correlation between $\tau_2^{-1}$, $\tau_3^{-1}$ produces an interesting challenge of the MI prior which could turn out problematic for smaller $N_{ch}$. Due to the strong correlation, the sample value for $\tau_2^{-1}$ can be predicted by knowing the value of $\tau_3^{-1}$. The same applies the other way around. To large extend we can predict $\tau_2^{-1}$ with a affine linear function $\tau_2^{-1} = m\tau_3^{-1} + n$ ((panel **a3**

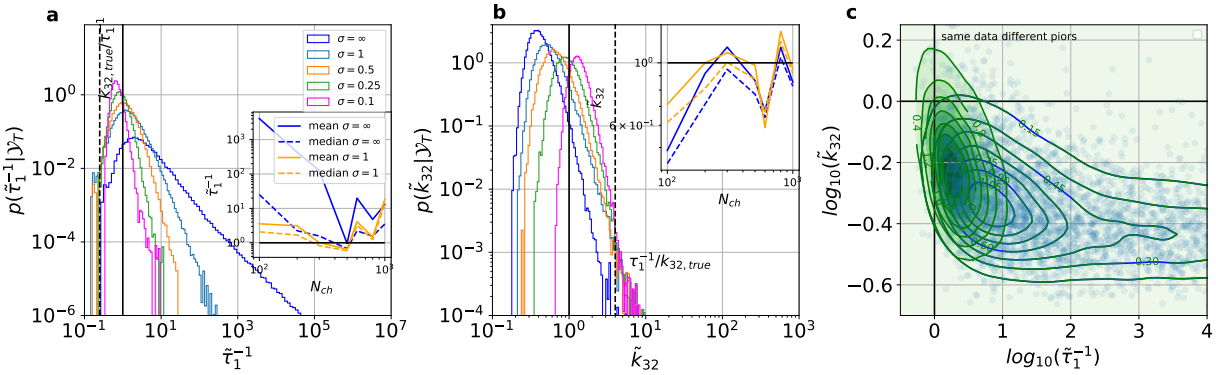


Figure 6: **A vaguely regularizing prior on the cooperativity factor renders the posterior proper even for the lowest quality data.** We demonstrate for the previously used data sets of CRN1 (Fig. 1**a**) with $N_{\text{ch}} = 10^2$ the effects of the value of $\sigma_{\text{cop}}$ using the larger sampling box. This makes the regularization more urgent. Note that $\tau_1^{-1} = k_{21}$ holds. The black continuous lines indicate the value of $\tau_{i,\text{true}}^{-1}$ and $k_{32,\text{true}}$ on the x-axis. The black dashed lines indicate the corresponding other $\tau_{i,\text{true}}^{-1}$ and $k_{32,\text{true}}$ in units of the true value of x-coordinate to visualize the bias of the regularizing prior, constraining the posterior more and more between the dashed and solid black line with decreasing $\sigma_{\text{cop}}$. **a,** $p(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ for either no or a series of increasingly strong regularization. The inset compares the effect of the vague regularization (orange) on the mean (solid curve) and median (dashed curve) vs. $N_{\text{ch}}$ with no regularization (blue). **b,** $p(k_{32} \,|\, \mathcal{Y}_T)$ for the same series of increasing regularization. The inset compares the effect of the vague regularization (orange) on the mean (solid curve) and median (dashed curve) vs. $N_{\text{ch}}$ with no regularization (blue). **c,** The effect of the regularization on $p(\tau_1^{-1}, k_{32} \,|\, \mathcal{Y}_T)$, with no regularization (blue) and the most vague regularization prior ($\sigma$= 1, green).

dashed line)) with $m = 0.5$ being the correlation coefficient and $n = 0.4$. Ignoring the affine part $n$ of the linear map originating from the CRN, the slope of loguniform prior on $\tau_3^{-1}$ will trivially be mapped by the linear function to $p(\tau_2^{-1} \,|\, \mathcal{Y}_T)$. Additionally, $\tau_2^{-1}$ has its own log uniform prior. Hence, both one dimensional priors together with the non-PI ridge, contribute to the posterior a scaling $\tau_i^{-2}$ for both parameters, potentially creating bias to too small values. Based on the corresponding 2d marginal distributions (Fig. 7 **a**), the Markov transition probabilities $\epsilon_{12}$, $\epsilon_{32}$ and $\epsilon_{23}$ appear to have a more and more diverging posterior the larger the sampling box for $\tau_2^{-1}$, $\tau_3^{-1}$ gets. These parameters should be considered unidentified, based on the herein-employed heuristic and visual criterion (see, discussion around Eq. 4) to asses the degree of PI $p(\mathbf{K} \,|\, \mathcal{Y}_T)$. Surprisingly, $\tau_1^{-1}$ is for CRN2 not as unidentified as for CRN1 (Fig. 3 **b**). We employed a smaller sampling box for the uniform prior (because of an often not converging sampler, likely caused by the non-PI problem and the curse of dimensionality which is discussed in App. J). Using a smaller sampling box for the uniform prior also disadvantages the MI prior because less of the part of the parameter space where the likelihood is flat is possible for the sampler to reach. In the App. Fig. 17 we demonstrate that *sufficiently proper* posteriors are achieved, for the MI prior at the possible, but certainly challenging data set of $N_{\text{ch}} = 5 \cdot 10^3$. However, with the uniform prior one needs an impossible data quality of $N_{ch} = 2.5 \cdot 10^5$. Hence, the MI $p(\mathbf{K})$ increases the range of acceptable data for this CRN at least 20-fold, likely more for this CRN. In Fig. 8 **a** we display $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ for the smallest tested data quality $N_{\text{ch}} = 10^2$. Using only the MI prior (black posterior) produces for $k_{32}$ and $k_{45}$ improper marginal posteriors for $N_{\text{ch}} < 5 \cdot 10^3$ (see above). From the discussion before, it is clear that the likelihood for $N_{\text{ch}} = 5 \cdot 10^3$ is only *sufficiently* PI that the improperness of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ is not detected visually even though there is no reason to assume that it is not there if one would sample from much larger simulation boxes.

### 4.7 Effects of the objective (theoretical diffusion limit) or common sense prior information (vague non-cooperativity) for the more complex CRN.

In Fig. 9 **a** we display $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ for the lowest tested data quality tested for the MI prior and for additional (vague and hard) constrains discussed below. Using only the MI prior (black posterior) produces for $k_{32}$ and $k_{45}$ improper marginal posteriors. The first *visually* proper $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ appears with $N_{\text{ch}} = 5 \cdot 10^3$ (Fig. 17 **a**). From the discussion before, it is clear that the likelihood for $N_{\text{ch}} = 5 \cdot 10^3$ is only *sufficiently* PI that the improperness of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ is not

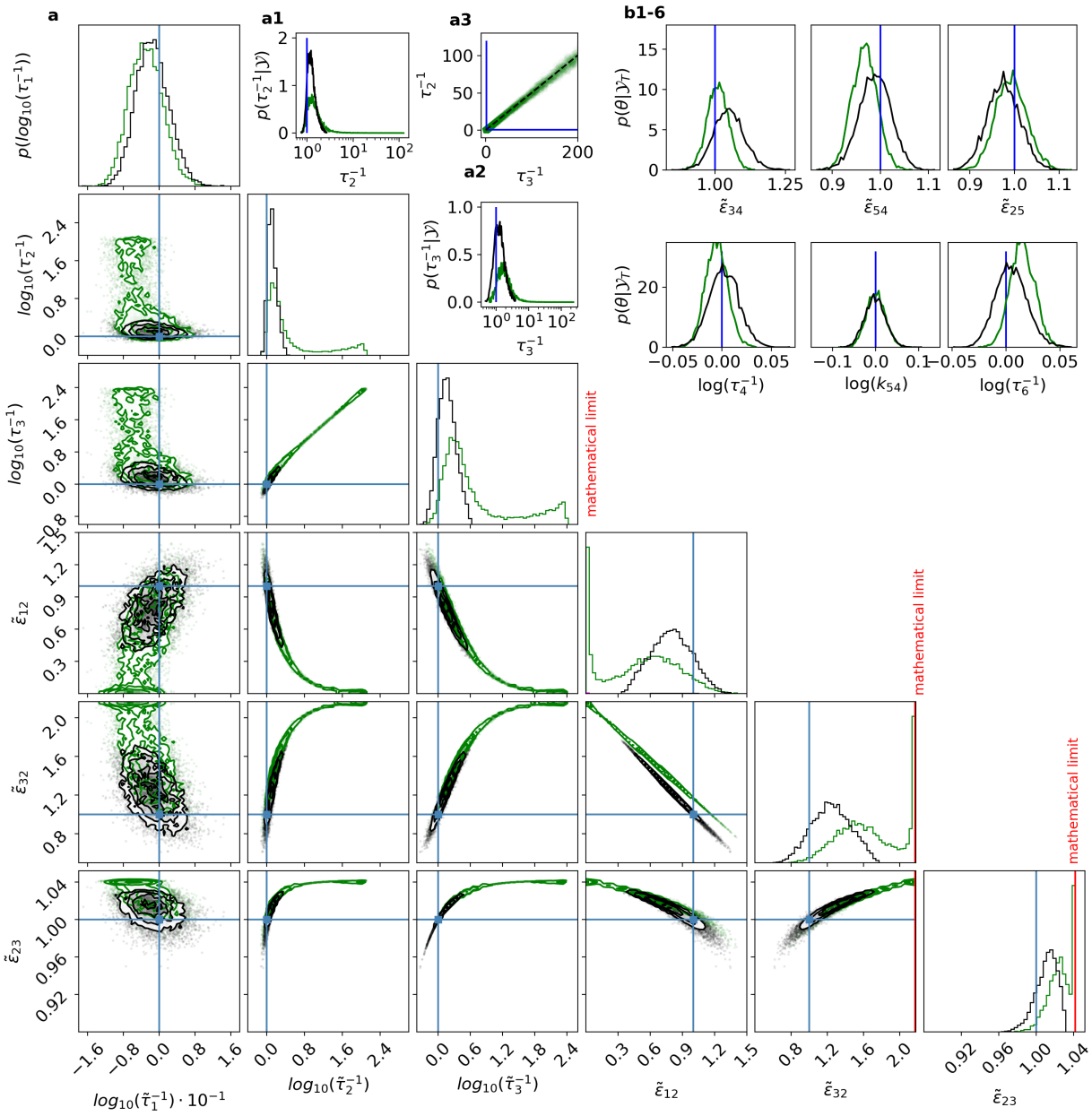


Figure 7: **For the more complex CRN2 the MI prior is necessary even for data sets of of unrealistic high quality such as $N_{\text{ch}} = 10^5$. a,** Posteriors based on uniform (green) and MI priors (black) are compared. The clearly non-Gaussian-shaped marginal posteriors + plus those concerning $\tau_1^{-1}$ are plotted in **a**, all other, rather Gaussian marginal posteriors, are shown in **b**. The insets **a1, a2** visualize the posterior without the log transformation. The black posterior is equipped with the MI $p(\mathbf{K})$. The green posterior is based on the uniform $p(\mathbf{K})$. The flat (green) posterior for $\tau_2^{-1}$ and $\tau_3^{-1}$, creates an exponential increase for the posterior of $p(\tau_2^{-1} \,|\, \mathcal{Y}_T)$ and $p(\tau_3^{-1} \,|\, \mathcal{Y}_T)$. **a3**, demonstrates the positive linear correlation contained in $p(\tau_2^{-1}, \tau_3^{-1} \,|\, \mathcal{Y}_T)$. Deviations of both parameters from the corresponding true value can be compensated to some extend by the other parameter. Note that the sampling box for the posterior samples of $\tau_2^{-1} \in [10^{-2}, 10^{6.5}]$ updated from the minimal information is more than an order of magnitude larger than the range of the simulation box used when the uniform $p(\mathbf{K})$ is used. This disadvantages the posterior based on the MI prior but demonstrates the larger robustness of the MI prior. The posteriors derived from the uniform $p(\mathbf{K})$ are sampled by $k_{ij}$ and then mapped to the $(\boldsymbol{\tau}, \boldsymbol{\epsilon})-$space.

detected visually even though there is no reason to assume that it is not there if one would sample from much larger simulation boxes. therefor employ the addional assumptions.

#### 4.7.1 Validation of the vague no cooperativity prior part 1: The vague no-cooperativity assumption increases the accuracy and decreases the uncertainty of the inference.

If one applies strict upper diffusion limits for $k_{21}$, $k_{32}$ and $k_{45}$ (Fig. 8, blue posterior and curves) one gains a proper $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ for $N_{ch} = 10^2$ in the corresponding dimensions and also the other dimensions become *sufficiently proper*. Adding further vague prior assumptions $\log_{10}(k_{21}/k_{32}) \sim \text{normal}(0, 0.5)$, $\log_{10}(k_{12}/k_{23}) \sim \text{normal}(0, 0.5)$, $\log_{10}(k_{45}/k_{32}) \sim \text{cauchy}(0, 1)$ and $\log_{10}(k_{54}/k_{23}) \sim \text{cauchy}(0, 1)$, which regularizes $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ gently towards a non-cooperative CRN (Fig 8 red posterior), reduces the mean-distance between $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ and $\mathbf{K}_{\text{true}}$ and reduces the uncertainty (Fig. 8 **c**). Note, that a less-vague no-cooperativity assumption $\log_{10}(k_{21}/k_{32}) \sim \text{normal}(0, 0.1)$, $\log_{10}(k_{12}/k_{23}) \sim \text{normal}(0, 0.1)$ reduces the Frobenius norm and RSME further (posterior suppressed for clarity in Fig. 8 **a**). But we show the consequences of the less vague prior on the posterior's RSME and Frobenius norm (magenta curves in Fig. 8 **b-c**). The RSME (Fig. 8 **b**) shows that at $N_{\text{ch}} \approx 5 \cdot 10^4$ the RSME transitions to $\propto 1/\sqrt{N_{\text{ch}}}$ while for $N_{\text{ch}} < 5 \cdot 10^4$ the non-PI problem combined with prior assumptions influence the posterior. Above, $N_{\text{ch}} = 5 \cdot 10^4$, the posteriors equipped with priors with diffusion limit produce similar RSMEs, unless the less vague cooperativity assumptions (Fig. 8 **b** magenta curves and posterior) are used. For the less vague prior the RSME converges onto the curves of the other two prior assumptions (blue and red curve) around $N_{\text{ch}} = 10^6$. In contrast, the pure MI prior has different RSMEs (Fig. 8 **b** black curve) for each data set. The vague no-cooperativity assumptions lost their influence on the RSME, while the diffusion still influences the RSME.
The Frobenius norm of the covariance matrix of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ shows (Fig. 8 **c**) that enforced upper diffusion limits (blue, red, and magenta curve) still add information and reduce the uncertainty of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$. Hence, even for data qualities of $N_{\text{ch}} > 10^6$, an ML inference would ignore relevant information to reduce the uncertainty of the inference. The Frobenius norm of the posteriors based on the pure MI prior without additional assumptions transitions at $N_{\text{ch}} = 10^5$ to the $N_{\text{ch}}^{-0.5}$-scaling.
To summarize, with MI prior with diffusion limits (Fig. 8), one can make inferences with more than $10^3$ times smaller $N_{\text{ch}}$ per time trace compared to Bayesian inferences with the uniform prior (Fig. 7) or ML/MAP inferences. Note that theoretical data qualities of $N_{\text{ch}} > 10^4$ are beyond possible to achieve data qualities. The added vague non-cooperativity assumptions contribute information to the posterior approximately $N_{\text{ch}} = 10^5$ as judged by the RSME and Frobenius norm. For the less vague prior assumptions the gain of relevant information lasts for values even higher than $N_{\text{ch}} = 10^6$ as juged by the Frobenius norm, but the RSME are roughly the same.

#### 4.7.2 Validation of the vague no-cooperativity prior to part 2: To what extent can channel properties be assessed against the bias of the no-cooperativity prior?

We test for the different priors, what are typical data requirements that $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ also supports the positive cooperativity defined as $r = k_{12,\text{true}}/(k_{54,\text{true}}) > 1$ (Sec. 4.5). One could also ask at what point becomes the bias of the additional soft coupling by the prior assumptions towards no-cooperativity harmful for the inference because the DGP (Fig. 1) holds $k_{12,\text{true}}/k_{54,\text{true}} = 5/2$. There are essentially two categories (*negative* $k_{12}/k_{54} < 1$ and *positive* $k_{12}/k_{54} > 1$ *cooperativity*) and the infinitesimal thin (green) line in between with no$-$*cooperativity* $k_{12}/k_{54} = 1$. Let $r = k_{12}/k_{54}$ be our measure of coorperativity. If $\int_{r=1}^{\infty} p(r \,|\, \mathcal{Y}_T) dr > 1$ the data supports positive cooperativity Notably, if $\int_{r=1}^{\infty} p(r \,|\, \mathcal{Y}_T) dr \approx 0.5$, also a negative cooperativity model is plausible just as positive cooperativity. As such, we are not looking for any significant test result. The minimal requirement that the data supports positive cooperativity is that the posterior median (solid lines) hold $\text{med}(r)_{post} > 1$. Further, the less probabilty mass is placed below $r = 1$ the higher the significance level would be in frequentist testing setting. We contrast the median with HDCIs that tell *what is the smallest most probable interval*. For skewed posteriors (because of uninformative data), HDCIs might indicate a different cooperativity model than the median. Note that for small $N_{ch}$ we work in a regime where frequentist testing would likely not produce significant results.
For the pure MI prior (Fig. 9 **a**), and $N_{\text{ch}} \in [5 \cdot 10^3, 2 \cdot 10^4]$ the posterior flucuates between a weak indication based on the median (solid line) that there is positive cooperativity and that both models are equally plausible. In particular, the median is biased towards too small values of $r$. The 0.5-HDCI (lower limit dashed black lines) is almost entirely smaller than $r = 1$. To have a unbiased median and a 0.5-HDCI that additionally supports qualitatively positive cooperativity one needs at least $N_{\text{ch}} > 5 \cdot 10^4$. In contrast, working with the uniform requires $N_{ch} = 2.5 \cdot 10^5$.
One may ask where the bias towards too small $r$ values for $N_{ch} \in [10^3 - 4 \cdot 10^4]$ comes from. The normal approximations used to derive the likelihood [41] is for $N_{ch} \in [10^3 - 4 \cdot 10^4]$ justified. Thus, we suspect the bias to originate from correlations and a strong non-PI problem in likelihood/posterior. Imagine a strong positive correlation after the inference between, e.g., $\tau_1$ an $\tau_2$ such that one can predict very accuratly from $\tau_1$ the scale of $\tau_2$. The limiting extreme case

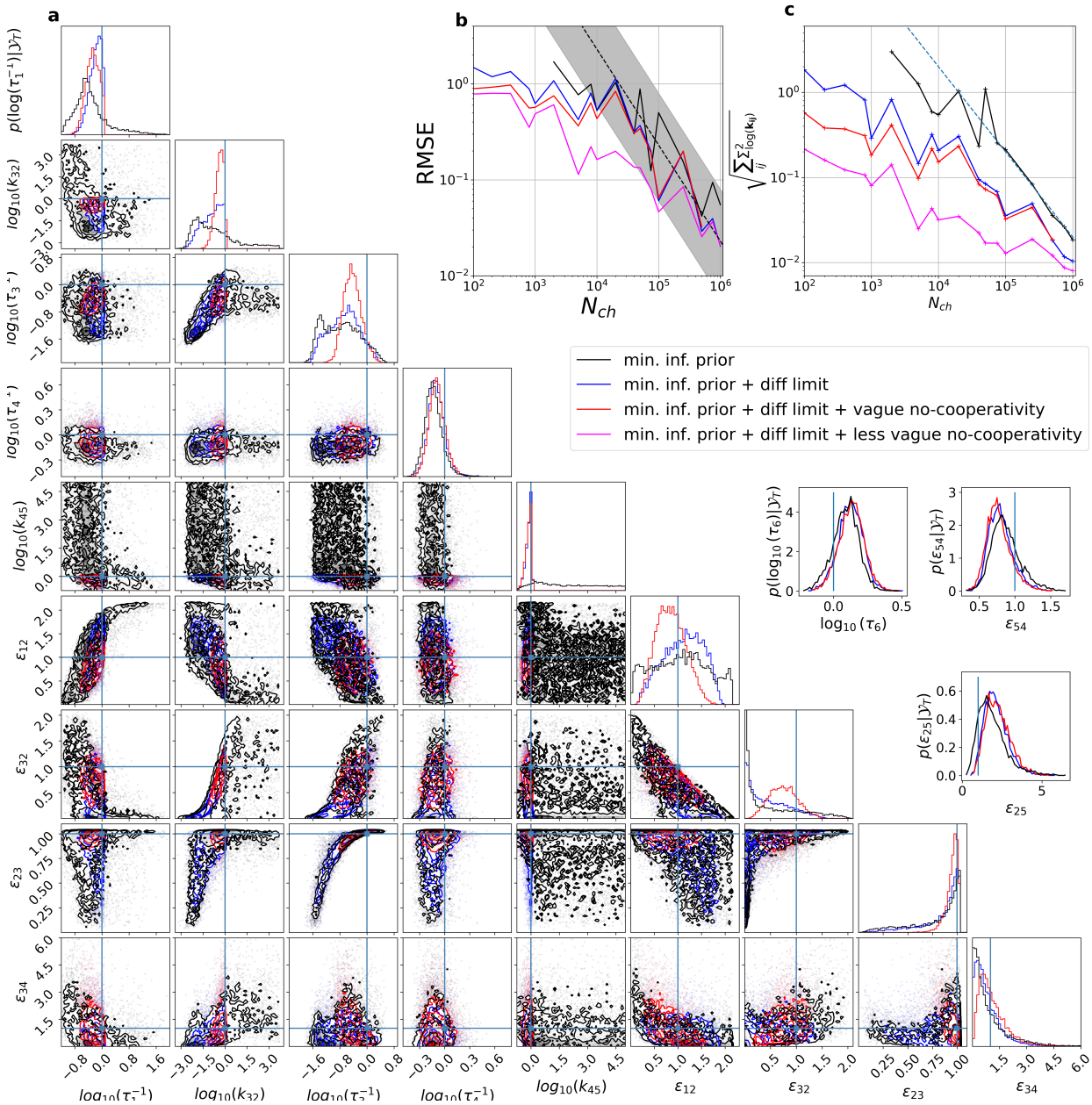


Figure 8: **For more complex CRNs, the MI prior enables inference for about two orders of magnitude lower data quality or quantity. Adding information by diffusion limits and vague bias towards non-cooperativity allows us to work with three orders of magnitude lower data quality.** The data sets simulated from CRN2 (Fig. 1) are analyzed The color black refers in all plots to $p(\mathbf{K} \mid \mathcal{Y}_T)$ based on the minimal informative prior. Blue corresponds to assuming diffusion-limited binding. Red an additional assumed vague prior on the cooperativity of the binding and unbinding rates. Magenta less vague no-cooperativity assumption. **a,** The true values of $\mathbf{K}$ are indicated by the blue lines. Posterior for $N_{\text{ch}} = 10^2$ for the MI prior, MI with upper limits and with an added vague no-cooperativity assumption. For visual clarity, we suppress $\tau_6^{-1}$, $\epsilon_{54}$, $\epsilon_{25}$ in the main plot but add sup panels which display the corresponding posteriors. Note that these parameters are only slightly influenced by the priors, and even without the priors, the posterior is peaking Gaussian-like with some skewness. **b,** The RSME of the log space of the chemical rates is plotted vs. $N_{\text{ch}}$ for the median (solid curve) and the marginal peak (dashed curve) for different prior assumptions. **c,** The trace (sum of all variances of the samples of all $k_{ij}$s) of the covariance matrix of the samples of $\mathbf{K}$

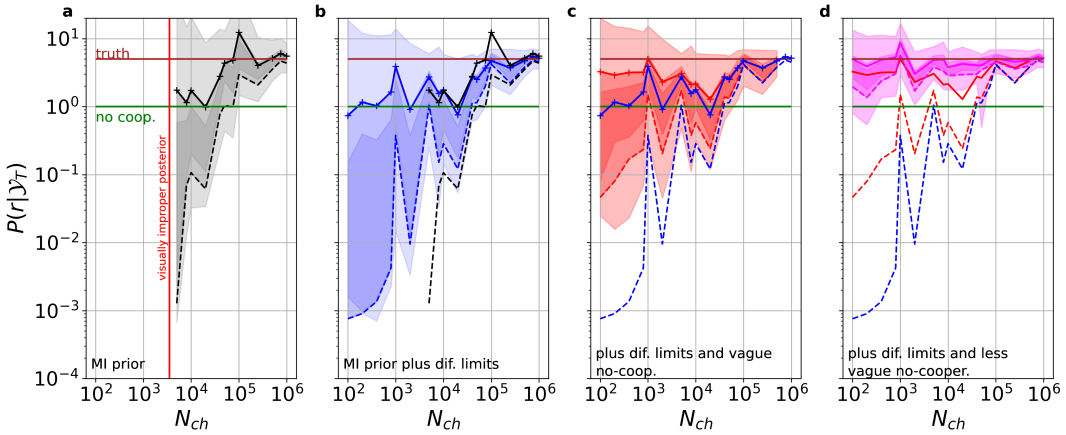


Figure 9: **The prediction of positive *cooperativity* (accelerated unbinding) of the unbinding of the second ligand gains certainty with vague informative prior assumptions with a bias towards no-cooperativity.** On the x-axis we plot $N_{\mathrm{ch}}$ as a proxy for the information content in the data originating from CRN2. The dashed lines indicate the lower limit of the $0.5$-HDCI, and the solid lines are the medians. The color corresponds to the prior assumptions with more information added from left to right. The set of three [0.99,0.5,0.3]-HDCIs is plotted in each panel. As a guide to the eye to discern the gain of certainty that there is negative cooperativity and the reduction of bias, we replot the median and lower limit of the $0.5$-HDCI in each panel. For visual clearity we supress the black lines in panel **d**. **a,** MI prior. **b** With an additional diffusion limit assumption. **c** With additional vague no-cooperativity assumption $\log_{10}(k_{21}/k_{32}) \sim \mathrm{normal}(0, 0.5)$, $\log_{10}(k_{12}/k_{23}) \sim \mathrm{normal}(0, 0.5)$, $\log_{10}(k_{45}/k_{32}) \sim \mathrm{cauchy}(0, 1)$ and $\log_{10}(k_{54}/k_{23}) \sim \mathrm{cauchy}(0, 1)$. **d** With an additional less vague no-cooperativity assumption $\log_{10}(k_{21}/k_{32}) \sim \mathrm{normal}(0, 0.1)$, $\log_{10}(k_{12}/k_{23}) \sim \mathrm{normal}(0, 0.1)$. A standard deviation of these priors corresponds approximately to 3.1 relative deviation of the corresponding parameters. Further, $\log_{10}(k_{45}/k_{32}) \sim \mathrm{cauchy}(0, 1)$ and $\log_{10}(k_{54}/k_{23}) \sim \mathrm{cauchy}(0, 1)$. Hence, we increase the bias towards non-cooperatively of the binding and unbinding rates in the CRN2.

of such a correlation between to parameters is realized in the non-SI problem the linear birth-death model (Eq. 13). Giving $k_{\mathrm{birth}}$ a log-uniform prior would result in $k_{\mathrm{death}}$ having a log uniform prior. Supplying to both parameters log uniform priors results after the inference in a effect of the combined prior on the posterior as $k_{\mathrm{death}}^{-2}$ and $k_{\mathrm{birth}}^{-2}$. An anticorrelation between the parameters would eliminate the effect of the prior.

Adding the diffusion limit to the posterior (Fig. 9**b**) extends, the region of a visually *sufficiently proper* posterior (Fig. 8) at least to $N_{\mathrm{ch}} = 10^2$ and decreases uncertainty (difference between black and blue dashed line). But an effect of the capabilities of the median to predict qualitatively positive cooperativity is small or not existent. While not indicating necessarily indicating a wrong model the median is typically undecided and thus biased to too small $r$ values. Hence, adding the diffusion limit reduces uncertainty of $p(r \,|\, \mathcal{Y}_T)$ but does not help to answer the fine grained question of cooperativity, unless one works with unrealistic data qualities (see uncertainty Fig. 9**b** above $N_{ch} > 10^5$).

This changes when adding the vague no-cooperativity prior (Fig. 9 **c**), which by its definition biases $p(r \,|\, \mathcal{Y}_T)$ more around the green line. The median and the HDCIs (see, Fig. 9 **c** the difference between red and blue lines) are shifted towards $r_{\mathrm{true}}$ against the bias of the Cauchy prior acting on $r$. The median now always indicates positive cooperativity. The $0.5$-HDCI is, roughly speaking, undecided but much less biased without the vague regularization.

In Fig. 9**d**, increase the on the hand between $k_{12}$ and $k_{23}$ and on the other hand between $k_{21}$ and $k_{32}$, while the cooperativity priors for on the one hand $k_{45}$ and $k_{32}$ and on the other hand $k_{54}$ and $k_{23}$, remain the same. This structure vaguely incorporates the prior knowledge that the vertical Markov transitions in the CRN represent changes in the protein, which might alter the binding and unbinding rates to some amount. A further shift of the lower limit of the 0.5-HDCI (dashed lines) and the median (dotted lines) can be observed. For realistic PC data quality assumptions $N_{\mathrm{ch}} < 10^4$ the prior combining diffusion limits with less vague non-cooperativity assumptions performs the strongest but also supposes the most. Only for high and unrealistic data quality $N_{\mathrm{ch}} > 10^5$ the posteriors without the vague non-cooperativity assumptions (black) seem to have a smaller bias to smaller negative cooperativity.

## 5 Conclusion

In Bayesian inference, an essential step is to specify a prior (distribution) that describes the plausibility of parameter values before the data are analyzed. If little about the actual values of some parameters is known *a priori* and if the data are not informative *enough* about these parameters, as e.g. often occurring in HMM inferences of biomolecule data, we show that MI priors are crucial to expand the range of acceptable data quality. They attempt to make posteriors as sensitive to the data as possible, thereby also alleviating non-PI pathologies of the dwell times in HMM models. Unfortunately, we show that for typical data qualities and quantities and realistic CRNs further *objective* or subjective assumptions are necessary to obtain interpretable HMM inferences (a proper or at least *sufficiently proper* posterior) to overcome the challenges from the non-PI problem of the likelihood. As a likelihood, we use a BF based on a FPa [41] to infer HMMs from macroscopic data from simulated PC data. Sufficient data for uniform priors or even ML can not realistically be obtained to determine the more complex CRN The data quality for uniform priors or extrapolated for ML inferences needs to be roughly above an impossible $10^5$ ion channels per patch for the more complex CRN, given 10 time traces stimulated with a different ligand concentration each.

In a first step, we perform a prior sensitivity analysis. We demonstrate the consequences for the posterior given two different priors: The default uniform prior of ML and Bayesian inference, which is often mistakenly seen as MI, and a prior inspired by the Jeffreys rule [62] and transformation group arguments of Jaynes [61]. The latter can be seen as MI or, at least, as much less informative and biased than the uniform prior.

PC data is highly aggregated, in other words, a partially observed CRN. For single- channel data, typically, there are only two types of signal classes *conducting* and *not conducting* in which all of the Markov states are aggregated. A statistical method to find the minimal number of Markov states within a signal class is called a dwell time analysis [21, 117, 118] This method is essentially an inference of an exponential mixture model (a sum of weighted exponential distributions), which will have the problem that some exponential distributions have small statistical weights. This poses a challenge for the inference.

Comparably, ensemble data of relaxation experiments, the signal, the sum of the population in all conducting states plus noise, will contain some of the eigenvalues of the CRN only with small or even zero spectral amplitudes. This makes it challenging or impossible to infer the complete rate matrix to a satisfying degree of certainty. Further, combinations of the chemical rates can lead to the same or almost the same eigenvalues, creating scenarios where the likelihood is practically non-identifiable (non-PI) or even structurally non-identifiable (non-SI). We show, analytically, for a simple linear CRN of two transitions and three states that both rates have a degree of non-PI problem. It is plausible that this is an inherent structure of partially observed CRNs due to the finite amount of molecules observed in CRNs. The numerical investigation of a more realistic CRN of an LGIC (App G) makes no exception.

In aid of clarity, in both CRNs discussed herein, one *infers* the ligand binding dynamics *filtered* by the kinetics of the channel gating (which needs also to be inferred). For most ligand concentrations in both used CRNs, the gating rates are the rate-determining kinetic steps. That means, e.g., values of the ligand binding rates which are, relatively speaking, instantaneous and have additionally little influence on the equilibrium value for all ligand concentrations, compared to the channel gating, cannot be distinguished anymore from values that would make ligand binding even faster. By little influence on the equilibrium we mean,e.g, a scenario when the binding rate for all used ligand concentrations is much larger than the unbinding rate then has the value of the binding rate little influence on the open probability. Interestingly this issue cannot be resolved by knowing the exact value of the unbinding rate or fixing it around the true value by a prior. This makes the likelihood insensitive at some parameter value toward faster binding rates (Fig. 14).

A non-PI likelihood means that even though one can find a unique and global maximum of the likelihood, one cannot predict the corresponding parameter uncertainty based on curvature at the optimum as done in ML approaches, as the likelihood becomes asymptotically a non-zero constant at some distance from the optimum in some directions in the parameter space. Naturally, any prior dominates the posterior in the regions of a constant likelihood value. The bias of the uniform prior to larger inverse dwell times or chemical rates (ascribing each decade ten times more probability mass than the previous smaller decade) combines in an unfortunate way with the non-PI problem of the likelihood itself. In contrast, the log-uniform part of the MI prior puts equal statistical weight on each decade and thus alleviates this problem. While both uniform- and MI-prior based posteriors are improper posteriors on the entire parameter space (from zero to infinity) in any finite sampling box (necessary for a sampler), the results of the posterior, given the MI prior, are therefore much less sensitive to the limits of the sampling box.

At first glance, for the simpler CRN used herein, one can expand the range of sufficient data by 10-fold when using the MI prior. However, the adequate data range can be broadened at least a 20-fold, likely more, for the more complex CRN. The posterior based on the MI prior reports a smaller uncertainty (which contains the true value) even for the highest-quality data sets. Importantly, for more complex CRNs, it is much simpler for the adaptive HMC sampler to produce well-mixing (converging) parameter chains i.e. the samples indicate that the typical set [119] of the true posterior was sampled.

Note that the non-PI problem is not limited to Bayesian statistics; it also limits the uncertainty quantification of ML inferences. If a non-Bayesian approach is used, we consider the profile likelihood [95, 120] as advantageous.

Nevertheless, the profile likelihood technique refers similarly to ML inference to an asymptotic large data set [121], which Bayesian statistics does not.

At a second glance, one can detect the posteriors's improperness using larger simulation boxes (Fig. 5). Hence, the usually arbitrarily chosen simulation box limits determine the posterior on a relevant scale as soon as the simulation box is large enough. The MI prior desensitizes the posterior concerning the sampling box limits. That is true in principle for any amount and quality of data and all nontrivial CRNs. Under good conditions (simple CRNs and informative data), this problem is only a technical note because the posterior has a peak close to the true values, which is multiple orders of magnitude higher than the purely prior-dominated parts. This makes it possible to ignore the strictly prior-dominated parts. However, often, the peak will be less dominant.

An *objective* solution to the non-PI/improper posterior problem is if theoretically derived, physical meaningful limits of the sampling box can be applied, even to a parameter subset. It shrinks the sampling box and the posterior to the physically possible size, reducing uncertainty of the posterior and making it *proper* at least in the dimensions where the limits are applied. Herein, it is shown that this information fusion from data and prior knowledge creates meaningful inferences with the lowest tested data quality, even for the most complex tested CRN. Nevertheless, derived theoretical limits might only sometimes be at hand, or even after their application, non-PI problems might remain in parameter dimensions where the upper limits do not apply. Herein, an additional vague prior on the ratio of some rates -a hyperparamter corresponding to the cooperativity of ligand binding and respective unbinding - is applied. Combing the objective and the common sense prior assumptions delivers the best inferences in terms of RSME and uncertainty estimation of the posterior. This additional prior biases the CRN gently towards CRNs where ligand binding and unbinding of the different channel subunits occur independently but still allow for positive and negative cooperativity over orders of magnitude (depending on the choice of hyperparameters). Hence it is a much less radical assumption compered to the commonly used non-cooperativty assumption [111, 112]. Not using such a prior would mean that one is willing to accept *a priori* any order of magnitude of cooperativity effects to occur, which is against common physical sense or biochemical experience - i.e. experience based priors. Thus, extreme effects should only be considered if the data is very certain about them. Using this prior, even without the physically motivated upper sampling box limits, renders the posterior proper in the relevant dimensions of the parameter space since the prior itself is proper for all data sets. With these prior assumptions, one can learn from the HMM inference about negative cooperativity within the CRN with $10^3$ times smaller data sets than with plain uniform prior assumptions or ML inferences. One could also apply this technique to heteromeric proteins or across homomeric proteins containing mutated binding sites [122, 123, 124] if the scale of the differences between binding pockets can be coarsely estimated *a priori*. The more coarse the *a priori* estimate is the heavier should the tails of the regularization prior be. A Cauchy prior on the logspace provides a heavier tail but is still a proper prior. A detailed study of the different possibilities is out of the scope of this paper.

On the meta-level, the last two paragraphs implicitly raise the question of when data is actually informative enough to use ML inference or uniform priors. The way the scientific community benchmarks HMM inference algorithms might be misleading regarding this question. In benchmark studies of inference algorithms (including this publication), results are obtained from synthetic data simulated from CRNs using the Gillespie algorithm [125, 126], thereby enforcing that the optimal description of the data is a fixed discrete state HMM because the DGP is an HMM (of a fixed number of states). Thus, the number of to-be-inferred chemical rate constants is fixed, too. Hence, with increasing data quality and quantity, the ML estimator may become sufficient given that the model is SI/PI and the likelihood is correctly specified. The DGP is included in the parametric family of CRNs. However, this is not the case in HMM inferences of actual experimental data. Analyzing actual experimental CRN data of a protein faces potentially an unlimited *matryoshka* problem: A more realistic description of some protein dynamics would be a continuous-state MM in the free energy landscape [127] of the protein-ligand system. This indicates that increasing the experimental data's amount or quality could require more intermediate discrete Markov states for a CRN that optimally [128, 129, 130] describes the process, given the experimental data at hand. This means that even more rates need to be inferred, increasing the risks for SI or PI problems in the inference. It is *a priori* unclear if ML inferences or uniform prior for Bayesian inference of HMMs for protein data has a reasonable uncertainty quantification for the optimal HMM ever, as long as the synthetic data that is used in the benchmark studies are not simulated from free energy landscapes by MD simulations.

In support of this skeptical thought, for MMs of MD simulations of proteins, it is shown in [131] that including transient states is essential to derive MMs that only violate the Markov assumption with neglectable errors, in particular, to correctly model diffusive processes that can't be correctly described by "jumping" over a few dominant and high barriers that separate regions of conformation space with no discernible internal structure. The same concerns apply to coupled binding and conformational change [132]. Currently, simulating long enough PC data time series from MD simulations poses unresolved challenges for the computational resources; however, recent progress might allow us to answer this question in the future. Similarly, HMMs of single-molecule FRET measurements [133] perform better when adding the transient states. Further, an algorithm recently proposed by [127] for single-molecule FRET directly samples the posterior of the low-dimensional model free energy landscape of a reaction coordinate by a Gaussian process without any discrete state HMM. His article shows that HMMs, using only the metastable states and the Arrhenius law, are biased in the inferred free energy landscape. Presumably, having more than two metastable states used in [127] would make his

HMM more appropriate. Can their findings be extrapolated to ensemble PC or cPCF data? How does the optimal complexity of a CRN scale, with increasing quality of PC or cPCF data? Would the optimal CRN require more and more states without a limit if the data quality is pushed to infinity? If the optimal complexity of CRNs for protein data scales with the data quality, ML inference might always produce an insufficient uncertainty report. These considerations are less of a concern for Bayesian inference because Bayesian statistics do not require that the asymptotic theorems of ML inference are met with sufficient accuracy for finite data. However, in Bayesian statistics, it is essential that priors are reasonable and balance sensitivity to the data and identifiability of parameters in a tailored way to make the most of the data.

## Acknowledgments

The authors are grateful to E. Schulz for designing a software to simulate channel activity in ensemble patches and to Th. Eick for performing simulations. This project was funded by the German Research Foundation (DFG) within Research Group FOR 2518 DynIon (project P2). F. Paul acknowledges funding from the Yen PostDoctoral Fellowship in Interdisciplinary Research and from the National Cancer Institute of the National Institutes of Health (NIH) through Grant CAO93577. M. Habeck acknowledges funding by the Carl Zeiss Foundation within the program "CZS Stiftungsprofessuren". We like to thank to M. Bücker for help with the computer cluster at the Friedrich Schiller University Jena and to K. Benndorf for comments on the manuscript.

## References

[1] Harry Bateman. The solution of a system of differential equations occurring in the theory of radioactive transformations. In Proc. Cambridge Philos. Soc., volume 15, pages 423–427, 1910.

[2] Zachary R Adam, Albert C Fahrenbach, Sofia M Jacobson, Betul Kacar, and Dmitry Yu Zubarev. Radiolysis generates a complex organosynthetic chemical network. Scientific Reports, 11(1):1743, 2021.

[3] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, and John ZH Zhang. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. Nature communications, 11(1):5713, 2020.

[4] Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T Margraf. Machine learning in chemical reaction space. Nature communications, 11(1):5505, 2020.

[5] Daniel Barter, Evan Walter Clark Spotte-Smith, Nikita S Redkar, Aniruddh Khanwale, Shyam Dwaraknath, Kristin A Persson, and Samuel M Blau. Predictive stochastic analysis of massive filter-based electrochemical reaction networks. Digital Discovery, 2(1):123–137, 2023.

[6] Evan Walter Clark Spotte-Smith, Samuel M Blau, Daniel Barter, Noel J Leon, Nathan T Hahn, Nikita S Redkar, Kevin R Zavadil, Chen Liao, and Kristin A Persson. Chemical reaction networks explain gas evolution mechanisms in mg-ion batteries. Journal of the American Chemical Society, 2023.

[7] Mingjian Wen, Evan Walter Clark Spotte-Smith, Samuel M. Blau, Matthew J. McDermott, Aditi S. Krishnapriyan, and Kristin Aslaug Persson. Chemical reaction networks and opportunities for machine learning. Nature Computational Science, 3:12 – 24, 2023.

[8] Lili Zhuang and Noel Cressie. Bayesian hierarchical statistical sirs models. Statistical Methods & Applications, 23:601–646, 2014.

[9] Shaobo He, Yuexi Peng, and Kehui Sun. Seir modeling of the covid-19 and its dynamics. Nonlinear dynamics, 101:1667–1680, 2020.

[10] David F Anderson and Thomas G Kurtz. Continuous time markov chain models for chemical reaction networks. In Design and analysis of biomolecular circuits: engineering approaches to systems and synthetic biology, pages 3–42. Springer, 2011.

[11] Bertil Hille. Ionic channels in excitable membranes. current problems and biophysical approaches. Biophysical journal, 22(2):283–294, 1978.

[12] David Colquhoun, G. A. Hawkes, and Katz Bernard. On the stochastic properties of single ion channels. P. of the Roy. Soc. of London. Series B. Biological Sciences, 211(1183):205–235, 1981.

[13] DD Loo, Akihiro Hazama, Stephane Supplisson, ERIc TuRK, and Ernest M Wright. Relaxation kinetics of the na+/glucose cotransporter. Proceedings of the National Academy of Sciences, 90(12):5767–5771, 1993.

[14] S Eskandari, EM Wright, and DDF Loo. Kinetics of the reverse mode of the na+/glucose cotransporter. The Journal of membrane biology, 204:23–32, 2005.

[15] August George and Daniel M Zuckerman. From average transient transporter currents to microscopic mechanism–a bayesian analysis. bioRxiv, pages 2023–10, 2023.

[16] Sheyum Syed, Fiona E Müllner, Paul R Selvin, and Fred J Sigworth. Improved hidden markov models for molecular motors, part 2: extensions and application to experimental data. Biophysical journal, 99(11):3696–3703, 2010.

[17] John D. Chodera, Phillip Elms, Frank Noe, Bettina Keller, Christian M. Kaiser, Aaron Ewall-Wice, Susan Marqusee, Carlos Bustamante, and Nina Singhal Hinrichs. Bayesian hidden markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty, 2011.

[18] Bettina G Keller, Andrei Kobitski, Andres Ja?schke, G Ulrich Nienhaus, and Frank Noe? Complex rna folding kinetics revealed by single-molecule fret and hidden markov models. Journal of the American Chemical Society, 136(12):4534–4543, 2014.

[19] Rafael A Rosales, William J Fitzgerald, and Stephen B Hladky. Kernel estimates for one-and two-dimensional ion channel dwell-time densities. Biophysical journal, 82(1):29–35, 2002.

[20] Feng Qin and Ling Li. Model-based fitting of single-channel dwell-time distributions. Biophysical journal, 87(3):1657–1671, 2004.

[21] Christopher Shelley and Karl L Magleby. Linking exponential components to kinetic states in markov models for single-channel gating. The Journal of general physiology, 132(2):295–312, 2008.

[22] Frank Noé, Hao Wu, Jan-Hendrik Prinz, and Nuria Plattner. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. J. Chem. Phys., 139(18):184114, 2013.

[23] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics, 37(6):1554–1563, 1966.

[24] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 77(2):257–286, Feb 1989.

[25] S-H Chung, John B Moore, Lige Xia, LS Premkumar, and Peter W Gage. Characterization of single channel currents using digital signal processing techniques based on hidden markov models. Philos. T. of the Roy. Soc. of Lond. Series B Bio. Sci., 329(1254):265–285, 1990.

[26] Arne Albertsen and Ulf-Peter Hansen. Estimation of kinetic rate constants from multi-channel recordings by a direct fit of the time series. Biophysical journal, 67(4):1393–1403, 1994.

[27] Feng Qin, Anthony Auerbach, and Frederick Sachs. Hidden Markov Modeling for Single Channel Kinetics with Filtering and Correlated Noise. Biophys J., 79(4):1928–1944, 2000.

[28] MC M de Gunst, HR Künsch, and JG Schouten. Statistical analysis of ion channel data using hidden markov models with correlated state-dependent noise and filtering. Journal of the American Statistical Association, 96(455):805–815, 2001.

[29] Rafael Rosales, J Alex Stark, William J Fitzgerald, and Stephen B Hladky. Bayesian restoration of ion channel records using hidden markov models. Biophysical journal, 80(3):1088–1103, 2001.

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

[31] Rafael A. Rosales. MCMC for hidden Markov models incorporating aggregation of states and filtering. Bull. Math. Biol., 66(5):1173–1199, Sep 2004.

[32] Colin D Kinz-Thompson and Ruben L Gonzalez Jr. Increasing the time resolution of single-molecule experiments with bayesian inference. Biophysical journal, 114(2):289–300, 2018.

[33] Zeliha Kilic, Ioannis Sgouralis, and Steve Pressé. Generalizing hmms to continuous time for fast kinetics: hidden markov jump processes. Biophysical journal, 120(3):409–423, 2021.

[34] Zeliha Kilic, Ioannis Sgouralis, Wooseok Heo, Kunihiko Ishii, Tahei Tahara, and Steve Pressé. Extraction of rapid kinetics from smfret measurements using integrative detectors. Cell Reports Physical Science, 2(5), 2021.

[35] Ayush Saurabh, Mohamadreza Fazel, Matthew Safar, Ioannis Sgouralis, and Steve Pressé. Single-photon smfret. i: Theory and conceptual basis. Biophysical Reports, 3(1):100089, 2023.

[36] Korak Kumar Ray, Colin D Kinz-Thompson, Jingyi Fei, Bin Wang, Qiao Lin, and Ruben L Gonzalez. Entropic control of the free-energy landscape of an archetypal biomolecular machine. Proceedings of the National Academy of Sciences, 120(21):e2220591120, 2023.

[37] Keegan E. Hines, John R. Bankston, and Richard W. Aldrich. Analyzing Single-Molecule Time Series via Nonparametric Bayesian Inference. Biophys J., 108(3):540–556, 2015.

[38] Ioannis Sgouralis and Steve Pressé. An introduction to infinite hmms for single-molecule data analysis. Biophysical journal, 112(10):2021–2029, 2017.

[39] Ioannis Sgouralis and Steve Pressé. Icon: an adaptation of infinite hmms for time traces with drift. Biophysical journal, 112(10):2117–2126, 2017.

[40] Luciano Moffatt. Estimation of Ion Channel Kinetics from Fluctuations of Macroscopic Currents. Biophys J., 93(1):74–91, 2007.

[41] Jan L Münch, Fabian Paul, Ralf Schmauder, and Klaus Benndorf. Bayesian inference of kinetic schemes for ion channels by kalman filtering. Elife, 11:e62714, 2022.

[42] Zoubin Ghahramani. Learning dynamic bayesian networks. In International School on Neural Networks, Initiated by IIASS and EMFCSC, pages 168–197. Springer, 1997.

[43] John Bechhoefer. Hidden markov models for stochastic thermodynamics. New Journal of Physics, 17(7):075003, 2015.

[44] Michal Komorowski, Bärbel Finkenstädt, Claire V. Harper, and David A. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinformatics, 10:343–343, 2009.

[45] Martin Fink and Denis Noble. Markov models for ion channels: versatility versus identifiability and speed. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1896):2161–2179, 2009.

[46] Paul Fearnhead, Vasilieos Giagos, and Chris Sherlock. Inference for reaction networks using the linear noise approximation. Biometrics, 70(2):457–466, 2014.

[47] Maria Myrto Folia and Magnus Rattray. Trajectory inference and parameter estimation in stochastic models with temporally aggregated data. Statistics and Computing, 28(5):1053–1072, Sep 2018.

[48] In Jae Myung. Tutorial on maximum likelihood estimation. Journal of mathematical Psychology, 47(1):90–100, 2003.

[49] George Casella and Roger L Berger. Statistical inference. Cengage Learning, 2021.

[50] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis. Chapman and Hall/CRC, 1995.

[51] T Toss. Bayes' theorem. nature methods, 12(4):277, 2015.

[52] Keegan E. Hines. A Primer on Bayesian Inference for Biophysical Systems. Biophys J., 108(9):2103–2113, 2015.

[53] Frank Ball. MCMC for Ion-Channel Sojourn-Time Data: A Good Proposal. Biophys J., 111(2):267–268, 2016.

[54] R. A. Fisher. Theory of statistical estimation. Mathematical Proceedings of the Cambridge Philosophical Society, 22(5):700–725, 1925.

[55] Sumio Watanabe. Almost all learning machines are singular. In 2007 IEEE Symposium on Foundations of Computational Intelligence, pages 383–388. IEEE, 2007.

[56] Milind Joshi, Andreas Seidel-Morgenstern, and Andreas Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. Metabolic engineering, 8(5):447–455, 2006.

[57] FG Ball, Yuzhi Cai, JB Kadane, and A O'hagan. Bayesian inference for ion–channel gating mechanisms directly from single–channel recordings, using markov chain monte carlo. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 455(1988):2879–2932, 1999.

[58] Elan Gin, Martin Falcke, Larry E. Wagner, David I. Yule, and James Sneyd. Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors. J. of Theoretical Biology, 257(3):460–474, 2009.

[59] Ivo Siekmann, Larry E. Wagner, David Yule, Colin Fox, David Bryant, Edmund J. Crampin, and James Sneyd. MCMC Estimation of Markov Models for Ion Channels. Biophys J., 100(8):1919–1929, 2011.

[60] Ivo Siekmann, James Sneyd, and Edmund J. Crampin. MCMC Can Detect Nonidentifiable Models. Biophys J., 103(11):2275–2286, 2012.

[61] Edwin T Jaynes. Prior probabilities. IEEE Transactions on systems science and cybernetics, 4(3):227–241, 1968.

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

[63] Robert E Kass and Larry Wasserman. Formal rules for selecting prior distributions: A review and annotated bibliography. Journal of the American Statistical Association, 435:1343–1370, 1996.

[64] Ruoyong Yang and James O Berger. A catalog of noninformative priors, volume 2. Institute of Statistics and Decision Sciences, Duke University Durham, NC, USA, 1996.

[65] Guido Consonni, Dimitris Fouskakis, Brunero Liseo, and Ioannis Ntzoufras. Prior distributions for objective bayesian analysis. Bayesian Analysis, 13(2):627–679, 2018.

[66] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. Physics letters B, 195(2):216–222, 1987.

[67] Radford M Neal and Radford M Neal. Monte carlo implementation. Bayesian learning for neural networks, pages 55–98, 1996.

[68] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. of Machine Learning Research, 15(1):1593–1623, 2014.

[69] Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for bayesian inference and optimization. J. of Educational and Behavioral Statistics, 40(5):530–543, 2015.

[70] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018.

[71] Andrew Gelman and Donald B Rubin. A single series from the gibbs sampler provides a false sense of security. Bayesian statistics, 4:625–631, 1992.

[72] Max Delbrück. Statistical fluctuations in autocatalytic reactions. The Journal of Chemical Physics, 8(1):120–124, 1940.

[73] Donald A McQuarrie. Kinetics of small systems. i. The journal of chemical physics, 38(2):433–436, 1963.

[74] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. J. Math. Biol., 54(1):1–26, Jan 2007.

[75] Petr Ptáček, Tomáš Opravil, and František Šoukal. A brief introduction to the history of chemical kinetics. Introducing the Effective Mass of Activated Complex and the Discussion on the Wave Function of this Instanton, 1:1–25, 2018.

[76] Daniel T Gillespie. The chemical langevin equation. The Journal of Chemical Physics, 113(1):297–306, 2000.

[77] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. J. of basic Engineering, 82(1):35–45, 1960.

[78] Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. Journal of BAsic Enginheering, 83(1):95–108, 1961.

[79] Andrey Kolmogoroff. Grundbegriffe der wahrscheinlichkeitsrechnung. 1933.

[80] Glenn Shafer and Vladimir Vovk. The sources of kolmogorov's grundbegriffe. 2006.

[81] Richard T Cox and ET Jaynes. The algebra of probable inference. American Journal of Physics, 31(1):66–67, 1963.

[82] Andrew Gelman and Yuling Yao. Holes in bayesian statistics. Journal of Physics G: Nuclear and Particle Physics, 48(1):014002, 2020.

[83] Christopher Nicolai and Frederick Sachs. Solving ion channel kinetics with the qub software. Biophysical Reviews and Letters, 8(03n04):191–211, 2013.

[84] Nicolaas Godfried Van Kampen. Stochastic processes in physics and chemistry, volume 1. Elsevier, 1992.

[85] Darren J Wilkinson. Stochastic modelling for systems biology. CRC press, 2018.

[86] Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. Physical Review E, 92(4):042124, 2015.

[87] Brian DO Anderson and John B Moore. Optimal filtering. Courier Corporation, 2012.

[88] Christoph Biskup, Jana Kusch, Eckhard Schulz, Vasilica Nache, Frank Schwede, Frank Lehmann, Volker Hagen, and Klaus Benndorf. Relating ligand binding to activation gating in cnga2 channels. Nature, 446:440, feb 2007.

[89] Jana Kusch, Christoph Biskup, Susanne Thon, Eckhard Schulz, Vasilica Nache, Thomas Zimmer, Frank Schwede, and Klaus Benndorf. Interdependence of Receptor Activation and Ligand Binding in HCN2 Pacemaker Channels. Neuron, 67(1):75–85, 2010.

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

[91] Van der Vaart and AW Asymptotic Statistics. Cambridge university press: New york. NY, USA, 1998.

[92] Paul Kienker. Equivalence of aggregated markov models of ion-channel gating. P. of the Roy. Soc. of London. B. Biological Sciences, 236(1284):269–309, 1989.

[93] Sandor Vajda, Keith R Godfrey, and Herschel Rabitz. Similarity transformation approach to identifiability analysis of nonlinear compartmental models. Mathematical biosciences, 93(2):217–248, 1989.

[94] Stefania Audoly, Giuseppina Bellu, Leontina D'Angio, Maria Pia Saccomani, and Claudio Cobelli. Global identifiability of nonlinear models of biological systems. IEEE Transactions on biomedical engineering, 48(1):55–65, 2001.

[95] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25(15):1923–1929, 2009.

[96] Thomas R Middendorf and Richard W Aldrich. Structural identifiability of equilibrium ligand-binding parameters. Journal of General Physiology, 149(1):105–119, 2017.

[97] Aslak Tveito, Glenn T Lines, Andrew G Edwards, and Andrew McCulloch. Computing rates of markov models of voltage-gated ion channels by inverting partial differential equations governing the probability density functions of the conducting and non-conducting states. Mathematical Biosciences, 277:126–135, 2016.

[98] Keegan E. Hines, Thomas R. Middendorf, and Richard W. Aldrich. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. The J. of General Physiology, 143(3):401–416, 2014.

[99] Franz-Georg Wieland, Adrian L Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. Current Opinion in Systems Biology, 25:60–69, 2021.

[100] Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. Identifiability analysis for stochastic differential equation models in systems biology. Journal of the Royal Society Interface, 17(173):20200652, 2020.

[101] Lorin S. Milescu, Gustav Akk, and Frederick Sachs. Maximum Likelihood Estimation of Ion Channel Kinetics from Macroscopic Currents. Biophys J., 88(4):2494–2515, 2005.

[102] Nicholas N. Lam, Paul D. Docherty, and Rua Murray. Practical identifiability of parametrised models: A review of benefits and limitations of various approaches. Mathematics and Computers in Simulation, 199:202–216, 2022.

[103] Gauri Sankar Datta and Malay Ghosh. On the invariance of noninformative priors. The annals of Statistics, 24(1):141–159, 1996.

[104] James O Berger, Jose M Bernardo, and Dongchu Sun. Overall objective priors. Bayesian Analysis, 10(1):189–221, 2015.

[105] Benjamin Trendelkamp-Schroer and Frank Noé. Efficient bayesian estimation of markov model transition matrices with given stationary distribution. The Journal of chemical physics, 138(16):04B612, 2013.

[106] Souad Assoudou and Belkheir Essebbar. A bayesian model for binary markov chains. International Journal of Mathematics and Mathematical Sciences, 2004(8):421–429, 2004.

[107] S Assoudou and B Essebbar. A bayesian model for markov chains via jeffrey's prior. Communications in Statistics - Theory and Methods, 32(11):2163–2184, 2003.

[108] M v Smoluchowski. Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. Zeitschrift für physikalische Chemie, 92(1):129–168, 1918.

[109] KE van Holde. A hypothesis concerning diffusion-limited protein–ligand interactions. Biophys. Chem, 101:249–254, 2002.

[110] Mark EJ Newman. Power laws, pareto distributions and zipf's law. Contemporary physics, 46(5):323–351, 2005.

[111] Remigijus Lape, David Colquhoun, and Lucia G Sivilotti. On the nature of partial agonism in the nicotinic receptor superfamily. Nature, 454(7205):722–727, 2008.

[112] Anmar Khadra, Zonghe Yan, Claudio Coddou, Melanija Tomić, Arthur Sherman, and Stanko S Stojilkovic. Gating properties of the p2x2a and p2x2b receptor channels: experiments and mathematical modeling. Journal of General Physiology, 139(5):333–348, 2012.

[113] Gideon Koren, Emily R Liman, Diomedes E Logothetis, Bernardo Nadal-Ginard, and Peter Hess. Gating mechanism of a cloned potassium channel expressed in frog oocytes and mammalian cells. Neuron, 4(1):39–51, 1990.

[114] William N Zagotta, Toshinori Hoshi, and Richard W Aldrich. Shaker potassium channel gating. iii: Evaluation of kinetic models for activation. The Journal of general physiology, 103(2):321–362, 1994.

[115] Jonathan Silva and Yoram Rudy. Subunit interaction determines i ks participation in cardiac repolarization and repolarization reserve. Circulation, 112(10):1384–1391, 2005.

[116] Kylie A Beattie, Adam P Hill, Rémi Bardenet, Yi Cui, Jamie I Vandenberg, David J Gavaghan, Teun P de Boer, and Gary R Mirams. Sinusoidal voltage protocols for rapid characterisation of ion channel kinetics. The Journal of physiology, 596(10):1813–1828, 2018.

[117] David Landowne, Bin Yuan, and Karl L Magleby. Exponential sum-fitting of dwell-time distributions without specifying starting parameters. Biophysical journal, 104(11):2383–2391, 2013.

[118] Michael Epstein, Ben Calderhead, Mark A. Girolami, and Lucia G. Sivilotti. Bayesian Statistical Inference in Ion-Channel Models with Exact Missed Event Correction. Biophys J., 111(2):333–348, 2016.

[119] Thomas M Cover. Elements of information theory. John Wiley & Sons, 1999.

[120] Andreas Raue, Clemens Kreutz, Fabian Joachim Theis, and Jens Timmer. Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984):20110544, 2013.

[121] Michael C Neale and Michael B Miller. The use of likelihood-based confidence intervals in genetic models. Behavior genetics, 27:113–120, 1997.

[122] Vasilica Nache, Nisa Wongsamitkul, Jana Kusch, Thomas Zimmer, Frank Schwede, and Klaus Benndorf. Deciphering the function of the cngb1b subunit in olfactory cng channels. Scientific reports, 6(1):29378, 2016.

[123] Nisa Wongsamitkul, Vasilica Nache, Thomas Eick, Sabine Hummert, Eckhard Schulz, Ralf Schmauder, Jana Schirmeyer, Thomas Zimmer, and Klaus Benndorf. Quantifying the cooperative subunit action in a multimeric membrane receptor. Scientific Reports, 6(1):20974, 2016.

[124] Jana Schirmeyer, Sabine Hummert, Thomas Eick, Eckhard Schulz, Tina Schwabe, Gunter Ehrlich, Taulant Kukaj, Melanie Wiegand, Christian Sattler, Ralf Schmauder, et al. Thermodynamic profile of mutual subunit control in a heteromeric receptor. Proceedings of the National Academy of Sciences, 118(30):e2100469118, 2021.

[125] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of computational physics, 22(4):403–434, 1976.

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

[127] J Shepard Bryan and Steve Pressé. Learning continuous potentials from smfret. Biophysical journal, 122(2):433–441, 2023.

[128] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statist. Surv., 6:142–228, 2012.

[129] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24(6):997–1016, Nov 2014.

[130] Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. Statistics and Computing, 27(3):711–735, 2017.

[131] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schuette, and Frank Noé. Markov models of molecular kinetics: Generation and validation. The J. of Chem. Phys., 134(17):174105, 2011.

[132] Hao Wu, Fabian Paul, Christoph Wehmeyer, and Frank Noé. Multiensemble markov models of molecular thermodynamics and kinetics. Proceedings of the National Academy of Sciences, 113(23):E3221–E3230, 2016.

[133] Flurin Sturzenegger, Franziska Zosel, Erik D Holmstrom, Karin J Buholzer, Dmitrii E Makarov, Daniel Nettels, and Benjamin Schuler. Transition path times of coupled folding and binding reveal the formation of an encounter complex. Nature communications, 9(1):4708, 2018.

[134] Frank Tuyl, Richard Gerlach, and Kerrie Mengersen. A comparison of bayes–laplace, jeffreys, and other priors: the case of zero events. The American Statistician, 62(1):40–44, 2008.

[135] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry, 81(25):2340–2361, 1977.

[136] Thomas R Middendorf and Richard W Aldrich. The structure of binding curves and practical identifiability of equilibrium ligand-binding parameters. J. of General Physiology, 149(1):121–147, 2017.

[137] Clemens Kreutz, Andreas Raue, Daniel Kaschek, and Jens Timmer. Profile likelihood in systems biology. The FEBS journal, 280(11):2564–2571, 2013.

[138] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. Statistical science, 7(4):457–472, 1992.

[139] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7(4):434–455, 1998.

[140] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, Paul-Christian Bürkner, et al. Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc. Bayesian Analysis, 16(2):667–718, 2021.

[141] Dootika Vats and Christina Knudson. Revisiting the gelman–rubin diagnostic. Statistical Science, 36(4):518–529, 2021.

# A A comment on Bayesian statistics as a generalization of propositional logic, frequency interpretation and Kolmogorov's axioms

Propositional logic defines a calculus to draw a conclusion, based on two premises which can be true are false. The conclusion can only be drawn, if the premises allows to draw the conclusion. As such the propositional calculus is unable to quantify information with some amount of uncertainty. However, for scientific reasoning and decision making it is desirable to assign a statement (conclusion) a value between 0 (false) and 1 (true) corresponding to how uncertain it is. One can understand Bayesian statistics as a generalization of propositional logic [81], which allows assigning the plausibility of a statement that is unknown to be true or not, such as "the single channel current holds $i = 1.6 \pm 0.6\,\mathrm{pA}$" or "tomorrow there will be an earthquake in Bavaria" a probability. Notably, the statements do not necessarily have a frequency interpretation. The outcome of a coin flip experiment *tails* or *heads* has a frequency interpretation because one can repeat the experiment under identical *enough* conditions and observe the proportion of the outcomes approaching the frequentist *definition* of probability with increasing trials. However, the statements "tomorrow it will rain with a probability of X" or "the value of a parameter lies within a specified interval with a probability of Y" have no frequency interpretation. Heuristically, because there is only one trial for the statement "tomorrow it will rain... ". One cannot reproduce the exact weather today to repeat the experiment of tomorrow's weather prediction. This legitimate because the modern mathematically definition of probabilities, given by Kolmogorov [79, 80], are indifferent if the event that is attributed a probability either has or does not have frequency interpretation.

# B Bayesian update exemplified on the beta-binomial model: The influence of MI priors.

Assume that the data $n$ is the outcome of a binomial experiment with and $N_{\mathrm{trials}}$ trials and that the modeler chooses the parametric family of binomial likelihoods. Note that this family includes the DGP. The following provides some intuition for the influence of MI beta priors on the posterior. The posterior has an analytical expression if the prior is the conjugate of the likelihood. In the case of a binomial likelihood

$$p(n \,|\, N_{\mathrm{trials}}, p_{\mathrm{suc}}) = \mathrm{binomial}(n, N_{\mathrm{trials}}, p_{\mathrm{suc}}) = \binom{N_{\mathrm{trials}}}{n} p_{\mathrm{suc}}^{n} \cdot (1 - p_{\mathrm{suc}})^{N_{\mathrm{trials}} - n}, \tag{28}$$

for $n$ success, the conjugate prior is the beta distribution

$$p(p_{\mathrm{suc}}, \boldsymbol{\alpha}) = \mathrm{beta}(p_{\mathrm{suc}}, \alpha_1, \alpha_2) = \frac{1}{\mathrm{B}(\alpha_1, \alpha_2)} p_{\mathrm{suc}}^{\alpha_1 - 1} (1 - p_{\mathrm{suc}})^{\alpha_2 - 1} \tag{29}$$

$\mathrm{B}(\boldsymbol{\alpha})$ is the beta function at the position $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. Thus, the posterior (Eq. 1) is

$$\begin{aligned} p(p_{\mathrm{suc}} \,|\, n) &\propto p_{\mathrm{suc}}^{k} \cdot (1 - p_{\mathrm{suc}})^{N_{\mathrm{trials}} - n} p_{\mathrm{suc}}^{a-1} (1 - p_{\mathrm{suc}})^{b-1} \\ &\propto \mathrm{beta}(p_{\mathrm{suc}}, n + \alpha_1, N_{\mathrm{trials}} - n + \alpha_2), \end{aligned} \tag{30}$$

again a beta distribution [98]. The parameters $\alpha_1$ and $\alpha_2$ of the beta prior can be interpreted mathematically as fictitious observations. The uniform prior, which can be parameterized as $\mathrm{beta}(1, 1)$, adds to each possible outcome one fictitious observation (Eq. 30). The beta distribution samples probability vectors $\mathbf{P} = (p_{\mathrm{succ}}, 1 - p_{\mathrm{succ}})$ with

$$\sum_i p_i = 1 \text{ and } 0 \leq p_{\mathrm{succ}} \leq 1. \tag{31}$$

Due to the constraints, $\mathbf{P}$ forms a straight line in a 2D space called simplex. In general, the probability simplex lies on a hyperplane defined by Eq. 31.

In Eq. 30, an intuition for the information contained in the prior (even if it is considered MI) and its influence on $p(p_{\mathrm{suc}} \,|\, \mathcal{Y}_N)$ is demonstrated: In Jeffreys's prior case, $\mathrm{beta}(0.5, 0.5)$ adds half a fictitious observation to each outcome. Correspondingly, uniform prior, parameterized as $\mathrm{beta}(1, 1)$, adds to each possible outcome one fictitious observation. Considering the information in units of fictitious observations, the improper Haldane's prior [134] $\mathrm{beta}(0, 0)$ would add the least information to $p(p_{\mathrm{suc}} \,|\, \mathcal{Y}_N)$. The Haldane prior can be derived as MI from transformation group arguments [61].

In each panel $p(p_{\mathrm{suc}} \,|\, \mathcal{Y}_N)$ is plotted (Fig. 10) derived from the Haldane prior $\mathrm{beta}(0, 0)$ (blue), the Jeffreys prior $\mathrm{beta}(0,5, 0.5)$ (red), and the uniform prior $\mathrm{beta}(1, 1)$ (black). These priors are contrasted by a prior $\mathrm{beta}(1.1, 1.1)$ (green) with a more obvious regularization tendency than the uniform. Note that the peak of the black posterior equals always $p_{\mathrm{suc,ML}}$. With only one observation (Fig. 10 **a1**) $p(p_{\mathrm{suc}} \,|\, \mathcal{Y}_N)$ derived from the Haldane prior is improper and thus not displayed. The posterior inherits the improperness from the Haldane prior. The likelihood values, identical to the black posterior, cannot compensate for the singular points of the Haldane prior. Using the Haldane prior, $p(p_{\mathrm{suc}} \,|\, \mathcal{Y}_N)$

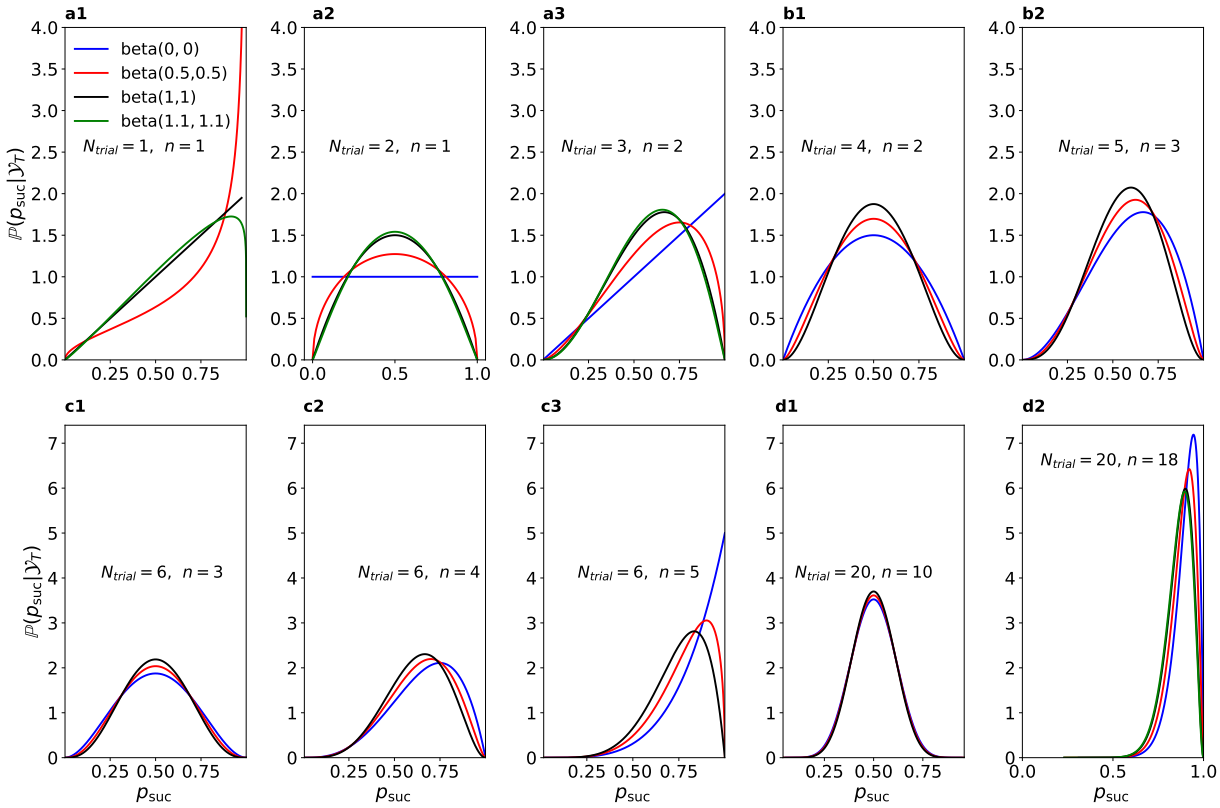


Figure 10: **The Jeffreys prior for the beta-binomial model is a compromise between the regularizing tendency of the uniform and the extreme uncertainty of the Haldane prior.** In each panel are the posteriors displayed derived from the Haldane (blue), Jeffreys (red), and uniform prior (black). The green posterior indicates the regularization towards the center of the possible values if a small amount of more information is applied in the prior $\text{beta}(1.1, 1.1)$ compared to the uniform prior. All of the priors and the resulting posteriors, too, can be parametrized by the beta prior. From panel to panel, the quantity of data is increased and shown for symmetrical and asymmetrical outcomes.

is always improper for binomial data sets in which one of the two outcomes did not occur in any trials. The second panel displays $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$ derived from a data set with two trials and a symmetrical result of 1 to 1. The Haldane prior appeals to emphasize too strongly the limiting values $p_{\text{suc}} = 0$ and $p_{\text{suc}} = 1$. Note that $p_{\text{suc}} = 0$ and $p_{\text{suc}} = 1$ are themselves not included in the support of the beta distribution, thus $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$ is, in principle, a consistent result and considering the typical point estimator in Bayesian statistics, the mean value

$$E[p_{\text{suc}}] = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{32}$$

of $\text{beta}(\alpha_1, \alpha_2)$. The inferred $p_{\text{suc}} = 0.5$ is identical to the ML inference. However, the Bayesian uncertainty quantification is extreme if the Haldane prior is chosen.

Assuming that one has seen each outcome once in two trials, it is implausible that one probability of the two alternatives holds, e.g., $p_1 = 10^{-100}$ and the other one holds $p_2 = 1 - 10^{-100}$. Because following frequency arguments, one should see that kind of data set only with a mean frequency of once per $10^{100}$ trials. However, $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$ places the same probability density for every $p_{\text{suc}}$ value. Hence, the Haldane prior assumes that the observed data set could be anything, from a representative data set for the process to the opposite, an extreme outliers data set.

In contrast, Jeffreys and the uniform prior and even more the $\text{beta}(1.1, 1.1)$ have more regularizing tendencies towards the center of the parameter space. Thus, they express the intuition (and add to the model) that seeing both outcomes in tiny data sets means that both probabilities will be of a similar magnitude. Here, regularization tendencies are meant relative compared with the Haldane prior.

The same relations between the Jeffreys prior and uniform prior are also visible in the first panel. The uniform regularizes more than the Jeffreys towards the center of the simplex. With this sparse data set, $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$ mimics the prior extremely. Thus, it is essential to question the following; does one think that the other outcome may be impossible,

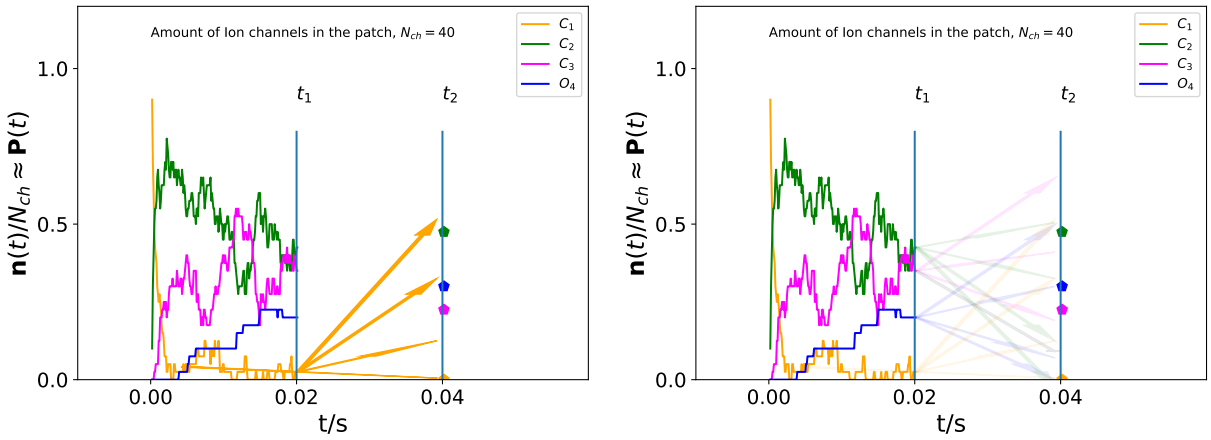


Figure 11: **The true Markov kernel for the occupation vector of an ensemble of channels is a convolution of multinomials**. The data is sampled by CRN1. The occupation vector $\mathbf{n}(t)$ fluctuates in each component around its mean occupation $\mathbb{E}(\mathbf{n}(t)) = N_{\text{ch}}\mathbf{P}(t)$. The open probabilty of the channel corresponds to $\mathbb{E}(n_{O_4}(t))/N_{\text{ch}} = P_O(t)$, and is proportional to the mean current. Assuming that one observes all states of a closed first-order CRN without noise, what would be the distribution $\mathbf{n}(t_2)$ given $\mathbf{n}(t_1)$. **a,** The subpopulation $n_1(t_1)\mathbf{e}_1$ of the single Markov state $C_1$ (orange curve) at $t_1$ distributes (orange arrows) over the entire Markov state space in a multinomial manner, thus producing a new subpopulation $\mathbf{n}^{(1)}(t_2)$ at $t_2$. **b,** The sum of the outcomes, $\sum_i \mathbf{n}^{(i)}(t+1)$, of the individual multinomial experiments of each subpopulation, indicated by the differently colored arrows, is generalized-multinomial distributed.

meaning that $p = 0$ if yes, $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$ (red) derived from Jeffreys prior seems legitimate. Otherwise, one should not work with the typical uninformative prior distribution and consider some regularization by the uniform (black posterior) or even by the $\text{beta}(1.1, 1.1)$ prior (green posterior).
For $N_{\text{trials}} = 3$ and $n = 2$ (Fig. 10 **a3**) the Haldane prior still strongly supports asymmetry, but the Jeffreys prior is a decent compromise between regularization and minimal informativeness. For larger data sets $N_{\text{trial}} = 4, 5$ with no or small asymmetry (Fig. 10 **b1-2**) in the outcome, all three posteriors are plausible. Increasing the trials and the asymmetry (Fig. 10 **c1-3**) ends up in a posterior for $N_{\text{trials}} = 6$ and $n = 5$, which still peaking at $p_{\text{suc}} \to 1$ if derived from the Haldane prior. If one knows that both probabilities must be finite, the Haldane prior delivers a posterior inconsistent with this prior knowledge. Increasing the data set to $N_{\text{trials}} = 20$ indicates (Fig. 10 **d1**) the limiting behavior that is proven in the Bernstein-von Mises theorem [69] that ML and Bayesian statistics inference deliver identical results if the data quality is high and the statistical model is regular. However, $N_{\text{trails}} = 20$ and $n = 18$ is far away from the result of the Berstein-von Mises theorem (Fig. 10 **d2**). Hence, the amount of data and the specific outcome are both relevant for understanding the prior's influence on $p(p_{\text{suc}} \,|\, \mathcal{Y}_N)$. The posterior of a $N_{\text{trails}} = 20$ and $n = 19$ data set not displayed would again be peaking for $p \to 1$. even more extreme than in panel **c3**. In summary, in most cases, the Jeffreys prior seems like a decent compromise between the regularization tendency of the uniform $\text{beta}(1, 1)$ and the sensitivity of the $\text{beta}(0, 0)$ to the data. $\text{beta}(0, 0)$ assumes that any order of magnitude of $\text{logit}(p_{\text{suc}}) = \frac{p_{\text{suc}}}{1-p_{\text{suc}}}$ is plausible.

## C The Markov Kernel of ensemble experiments is a concolution of multinomials

The Markov kernel, which samples $\mathbf{s}(t + \Delta)$ from $\mathbf{s}(t) = \mathbf{e}_j$ for a single channel, is a categorical distribution or, in other words, a multinomial distribution with only one trial.

$$\mathbf{s}(t + \Delta t) \sim \text{multinomial}(1, \mathbf{T}_{:,j}), \tag{33}$$

where $\sim$ means *sampled from* and $\mathbf{T}_{:,j}$ denotes the column from $\mathbf{T}$ with (column) index $j$. The used BF algorithm infers the elements of $\mathbf{K}$, constituting the CRN of the channel. The hidden ensemble state vector $\mathbf{n}(t) \in \mathbb{N}_0^M$ counts

$$\mathbf{n}(t) := (n_1(t), n_2(t), n_3(t), n_4(t))^\top = \sum_{i=1}^{N_{\text{ch}}} \mathbf{s}_i(t), \tag{34}$$

the number of channels in each Markov state $\mathbf{s}$. The solution of the CME for any first-order CRN, including birth, death, and mono-molecular transitions, is a convolution of Poisson and multinomial distributions [74], which reflects the first-order Markov dynamics, which a chemically reacting system is believed to follow. See App. D for an illustration of

the differences between the traditional deterministic description of such CRN by RE, which loose the first-order Markov property in contrast to the CME. For channels, only mono-molecular and death transitions are allowed or transitions that can be well approximated by first-order kinetics. Note that one can add to each CRN of a functional channel an absorbing state that collects the channels that become non-functional after some time. What would be the true Markov kernel of such an ensemble of channels?

Consistent with the findings of [74], we argued [41] that the Markov kernel of a closed CRN describing channels needs to be a $M$-dimensional discrete convolution of multinomials. The argument is as follows: From the theory of Markov models, one knows that the distribution of channels originating from state $i$ is a multinomial distribution (Fig. 11**a**)

$$p(\mathbf{n}^{(i)}(t+\Delta t)\,|\,n_i)\mathbf{e}_i) = p(n_1,\ldots,n_M\,|\,\mathbf{n}(t)) = n_i\mathbf{e}_i) = \frac{n_i!}{n_1!\cdots n_M!}T_{1,i}^{n_1}\cdots T_{M,i}^{n_M}. \tag{35}$$

Each of the $M$ subpopulations $n_i(t)\mathbf{e}_i$, with $i \in \{1,\ldots,M\}$, performs its own independent multinomial experiment at $t$. Hence, the sum,

$$\mathbf{n}(t+\Delta t) = \sum_i^M \mathbf{n}^{(i)}(t+\Delta t), \tag{36}$$

is distributed by the $M$-dimensional discrete convolution of multinomials which is defined in [41] as

$$\mathbf{n}(t+\Delta t) \sim \text{generalized multinomial}(\mathbf{n}(t),\mathbf{T}) \tag{37}$$

To sample time traces of the ion channels governed by example CRNs, we use the Gillespie [125, 135] implemented in cube and an in-house algorithm. In that way, the time traces contain the MM property. Alternatively, one could sample the time traces using the generalized-multinomial distribution, given small time steps.

## D Differences of the full CME compared to its deterministic approximation the RE

Traditionally, chemically, biochemically or physically reacting ensemble of molecules are described by the deterministic RE approach [75], which is a mean value $\mathbb{E}[\cdot]$ based description neglecting the discrete fluctuations originating from the intrinsic noise. A typical example would be the radioactive decay model

$$\mathbf{X} \xrightarrow{k_\mathrm{d}} \emptyset \tag{38}$$

with $\mathbf{X} \in \{0, 1\cdots N_0\} \subset \mathbb{N}_0$. The QM interpretation of the disintegration of a radioactive atom is that these events occur stochastically with a rate $k_d$ reducing the occupation number $n \in \mathbb{N}_0$ (the number of atoms $X$) in discrete steps by one. However, the RE approach assumes $\mathbb{E}[n] \in \mathbb{R}_0$ and a continous deterministic decay

$$\frac{d}{dt}\mathbb{E}[n(t)] = -k_d\mathbb{E}[n(t)]. \tag{39}$$

which is a justified assumption for the forward model as long the number of particles is large. The solution of Eq. 39 including the initial value $N_0$ is $\mathbb{E}[n(t)] = N_0\exp(-\alpha t)$. One differential equation (ODE) for the CRN is the essence of the RE approach. The RE model alone is unable to make any statement about the distribution $\Pr n(t)$ of radioactive particles at $t$ only the mean value $\mathbb{E}[n(t)]$ is modeled. Unfortunately, the fact that one refers here to a mean value is often suppressed such that this notation $\frac{d}{dt}n(t) = -\alpha n(t)$ is common.

What would the CME for the radioactive decay look like? It is a system of $\mathbb{N}_0+1$ coupled ODEs representing the probability masses for each of the $\mathbb{N}_0+1$ possible Markov states of the system, which has to be solved simultaneously. For $n \in \{1, 2\cdots N_0-1\}$

$$\frac{\partial}{\partial t}\Pr n,t = k_\mathrm{decay}[(n+1)\Pr n+1,t - n\Pr n,t]. \tag{40}$$

For the special Markov state $n = N_0$ one has $\frac{\partial}{\partial t}\Pr N_0,t = -k_\mathrm{decay}[N_0\Pr N_0,t]$ and for $n = 0$ $\frac{\partial}{\partial t}\Pr 0,t = k_\mathrm{decay}[\Pr 1,t]$. Essentially, the CME is a system of continuity equations for the probability of each (Markov) state the system can be in. Note that this system of $N_0$ atoms, there are not only two Markov states of the single atom. The system exists in a lattice of states being identified by the lattice vector $(n_1,n_2)^T$ with $n_1+n_2 = N_0$.

One can solve this equation by the following trick. Consider one radioactive atom. The CME for this particle is only $\frac{\partial}{\partial t}\Pr 1,t = -k_d\Pr 1,t]$. Solving $\Pr 1,t$ immediately solves $\Pr 0,t$ because of $\Pr 1,t = 1-\Pr 0,t$. The solution is $\Pr 1,t = \exp(-k_d t)$ if the initial condition is deterministic that the particle is in state $\mathbf{X}$. Now, if at $t_o$ $N_0$ particles are in state $\mathbf{X}$ they all have the same probability to be decayed at $t$ such that the result of is a time-dependent binomial distribution $\Pr n,t = \binom{N_0}{n}\exp(-k_d t)^n(1-\exp(-k_d t))^{N_0-n}$.

Assume the CRN of protein production and degradation, called the chemical birth-death process. There exist two chemical reactions: the production of proteins

$$\emptyset \xrightarrow{k_\mathrm{birth}} \mathbf{X} \tag{41}$$

and

$$\mathbf{X} \xrightarrow{k_{\text{death}}} \emptyset \tag{42}$$

The CME of that system would be a system of infinitely many ODEs because this open system can, in principle, have any number of proteins. However, its deterministic RE model would still be a single ODE. The corresponding CME has a solution [74] but this example indicates the rise in complexity for an analytical solution of the CME model. A simple example of a CRN comparable to ion channel dynamics has two reversible reactions

$$\mathbf{Y} \xrightarrow{k_{\text{b}}} \mathbf{X} \tag{43}$$

and

$$\mathbf{X} \xrightarrow{k_{\text{d}}} \mathbf{Y}. \tag{44}$$

Then would the probability distribution of finding $n_1$ molecules in state $\mathbf{X}$ be distributed by a convolution of multinomials [74, 41] if $\mathbf{Y}$ and $\mathbf{X}$ are both populated at $t = 0$.

For example, a well-stirred and closed ensemble of LGICs transitions from one chemical state into another. If one assumes that the amount of ligands in the bulk is orders of magnitude larger $N_{\text{ligand}} \gg N_{\text{ch}}$ than the number of ion channels in the patch, then one could, in principle, describe the mean evolution of the ensemble of ion channels by a linear ordinary differential equation
Assume the CRN of protein production and degradation model, called the chemical birth-death process. There exist two chemical reactions in the production

$$\emptyset \xrightarrow{k_{\text{birth}}} \mathbf{X} \tag{45}$$

of a protein $\mathbf{X}$ which increase $n_{X,\text{new}} = n_{X,\text{old}} + 1$ the number of proteins and the degradation of a protein

$$\mathbf{X} \xrightarrow{k_{\text{death}}} \emptyset \tag{46}$$

which decreases $n_{X,\text{new}} = n_{X,\text{old}} - 1$ the amount of proteins by one. The CME describes the probability $\Pr n, t$ to have at a given time $n = 0$, $n = 1$, $n = 2$... $n = \infty$

$$\frac{\partial}{\partial t} \Pr n, t = k_{\text{birth}}[(n-1) \Pr n-1, t - n \Pr n, t] + k_{\text{death}}[(n+1) \Pr n+1, t - n \Pr n, t] \tag{47}$$

For each $n \in \mathbb{N}_0$, there exists a linear differential equation which is coupled to the differential equations of $n - 1$ and $n + 1$.

Let $n(t)$ be the number of proteins, the deterministic RE description for this system is then

$$\frac{d}{dt}\mathbb{E}[n(t)] = k_{\text{birth}} - k_{\text{death}}\mathbb{E}[n(t)] \tag{48}$$

which is an inhomogeneous linear differential equation. The solution of Eq. 48 alone allows no statement about the fluctuations of the intrinsic system.

## E A Bayesian model inference that uses a non-SI likelihood can still be meaningful in Bayesian context.

The definition of SI non-SI relates to ML inferences but it has consequences for Bayesian inference, too. Notably, Bayesian inferences can still be meaningful without the likelihood being SI. Given a non-SI likelihood, the posterior will not have a unique optimum unless induced by a informative prior such as, e.g., for the linear birth-death model (Eq. 13), $k_{\text{death}} \sim \text{normal}(\mathbb{E}[k_{\text{death}}], \sigma)$. Unfortunately, a prior misspecification in the sense that $\mathbb{E}[k_{\text{death}}] - k_{\text{death,true}} \neq 0$ cannot be corrected asymptotically by data if the model is non-SI. Imagine, that one can deduce a upper limit on one of the rates in the linear birth-death model. This makes the posterior proper. Notably, $p(k_{\text{death}} \,|\, \mathcal{Y}_T)$ and $p(k_{\text{birth}} \,|\, \mathcal{Y}_T)$ will not be very useful but $p(k_{\text{death}}, k_{\text{birth}} \,|\, \mathcal{Y}_T)$ is. It will demonstrate the symmetry $\xi = k_{\text{birth}} - k_{\text{death}}$ of the likelihood and report probability densities for the corresponding rate matrix $\mathbf{K}$. Of course this tool box is in this simple example not necessary because the symmetry of the likelihood $\xi$ can be identified straight away but finding its exact position given noisy data requires the Bayesian update. However, for more complex inferences such as binding curves of in principle cooperative binding sites [98, 96, 136] the symmetries of the likelihood $\xi(\boldsymbol{\theta})$ might be insight full if some prior knowledge of upper and lower limits or informative priors of some parameters can be applied.

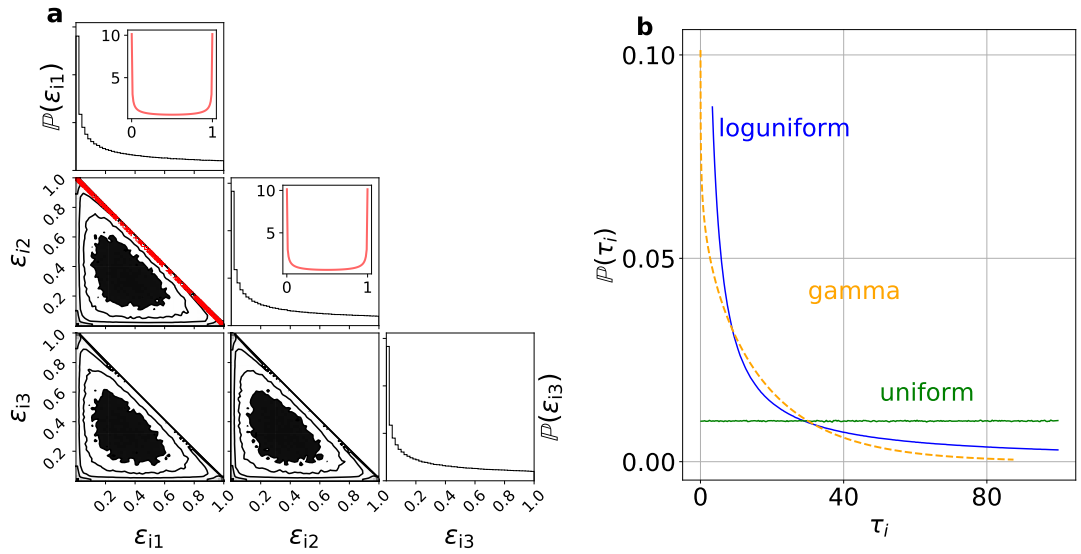


Figure 12: **MI prior for a Markov state with supposed three transitions that exit the state.** **a** A 3d Dirichlet prior $\mathrm{dir}(0.5, 0.5, 0.5)$ (black) for a supposed state which has three Markov transitions leaving the state such as state $C_4$ in CRN2 (Fig. 1). For a supposed state with two exiting Markov transitions, the corresponding $\mathrm{dir}(0.5, 0.5) = \mathrm{beta}(0.5, 0.5)$ (red) is displayed in the first two columns. Usually, the beta distribution is only considered as the distribution of success probabilities. Nevertheless, each draw from a beta distribution is a draw from a probability simplex by $P_{\text{success}}$ and $P_{\text{fail}} = 1 - P_{\text{success}}$. The constraints $\sum_i \epsilon_i = 1$ and $0 \leq \epsilon_i \leq 1$ of the probability simplex in 2d create a 1d submanifold, simply a straight line (red). However, the constraints define a 2D submanifold embedded in 3D when three outcomes exist. **b**, shows the possible alternatives for the prior on $\tau_i$ or one $k_{ji}$ of column the $i$-th state which sets the time scales of the Markov transitions leaving the state. The panel contrasts the log-uniform prior (blue) against the uniform prior (green) and the a gamma prior (yellow) which would have regularising effect if small but finite parameters are used. A simple way to draw from a log-uniform distribution is $\tau' \sim \mathrm{uniform}(a, b)$ followed by $\tau = 10^{\tau'}$. This way of drawing log-uniform random numbers shows that each decade is equally probable.

## F Illustration of the prior for a column of the rate matrix with two Markov transitions leaving the state

Here, we discus the shape of the MI prior corresponding to one column of $\mathbf{K}$ for two different scenarios. If the CRN assumes, e.g., two existing Markov transitions, $W_i = 2$, leaving the $i$-th state, the corresponding prior needs to be a beta distribution (2d Dirichlet distribution, red, inset of Fig. 12**a**). The probability drawn from the beta distribution is $\epsilon_{ji}$, and because there is only one alternative transition, the corresponding probability of that transition equates to $1 - \epsilon_{ji}$. Three assumed Markov transitions, $W_i = 3$, require a 3d Dirichlet distribution (Fig. 12**a**); four connections are distributed by a 4d Dirichlet distribution, et cetera. The only exception is a Markov state at some CRN's end. This Markov state is only connected by one transition to the network. In that case, $k_{ji} = \tau_i^{-1}$ such that there is only the log-uniform distribution (Fig. 12**b**) present for this column of $\mathbf{K}$. To understand the difference in the assumptions between a log-uniform distribution for $\tau_i^{-1}$ (Fig. 12**b**, blue curve) and a uniform prior on $\tau_i$ (green curve) consider the following. A uniform prior states that one has ten times more probability in $[10^1 - 10^2]$ than in $[10^0 - 10^1]$ that the actual value is in that decade. This relationship continues for each decade. In contrast, if one assumes a log-uniform prior such that the exponent is uniformly distributed, it means that each decade has the same probability. Notably, both distributions, uniform and log-uniform, are improper (Eq. 2), which causes the posterior to be improper if the likelihood is too flat. Similar to the uniform distribution, the log-uniform distribution is only a distribution if restricted to some interval $[a, b] \subset (0, \infty)$ with $0 > a > b < \infty$. That has potential risks for any statistical inference. Below, App. G.1, indicates that it is a general feature of HMMs that the likelihood is flat in a space of infinite volume and only in a finite volume around $\mathbf{K}_{\text{true}}$ is the likelihood, not flat. That means using improper priors should lead to an improper posterior, which we can reproduce (Fig. 5). So, careful checking of the posterior's tails is needed to understand how sensitive the results of the inference are to the typically chosen subjective limits of the sampling box.

### F.1 The non-PI problem with other states beeing observed.

If a second orthogonal observable is added, such as $y_2(t) \propto \mathbb{E}[n_B(t)]$, this intrinsic non-PI problem for $k_1$ disappears for this CRN but not for $k_2$ because for the general solution

$$\mathbb{E}[n_B(t)] = n_A \cdot (\frac{k_1 \cdot L}{k_1 \cdot L - k_2} \cdot \exp(-k_2 \cdot t) - \frac{k_1 \cdot L}{k_1 \cdot L - k_2} \cdot \exp(-k_1 \cdot L \cdot t)) \tag{49}$$

including the mentioned initial conditions assuming $k_1 \cdot L \gg k_2$ produces $\mathbb{E}[n_B(t)] = n_A \cdot (\exp(-k_2 \cdot t) - \exp(-k_1 \cdot L \cdot t))$. Hense, both exponential decays are preserved. However assuming $k_2 \gg k_1 \cdot L$ for all $L$ eliminates both amplitudes, which is an intuative result because $\mathbf{B}$ is hardly populated. Observing only $\mathbb{E}[n_B(t)]$ has no PI problem in the large $k_1$ direction but still in the large $k_2$ direction. Having two signals $y_1(t) \propto \mathbb{E}[n_O(t)]$ and $y_2(t) \propto \mathbb{E}[n_A(t)]$ resolves the non PI for $k_1$ but does not resolve it for $k_2$, since

$$\mathbb{E}[n_A(t)] = n_A \cdot \exp(-k_1 \cdot t) \tag{50}$$

holds. Note, that Eq. 50 has no dependency concerning the rates in the amplitude.
To eliminate the pathologies of the likelihood where $k_2 \gg k_1$ holds in $\boldsymbol{\theta}$, one needs to observe data from a mixed population $\mathbf{n}(t_0) = (n_A, n_B, 0)$ and a signal that contains two orthogonal observables $y_1 \propto \mathbb{E}[n_A(t)]$ and $y_2 \propto \mathbb{E}[n_B(t)]$ or alone $y \propto \mathbb{E}[n_O(t)]$ because $\mathbb{E}[n_O(t)]$ contains with these initial conditions, in this case, both time scales. Note that because of the initial conditions, the solution of the CME is not a multinomial distribution anymore but a convolution of two multinomials. In other words, the group of ion channels starting in $\mathbf{A}$ have a different $p_O(t)$ compared to the group of ion channels starting in $\mathbf{B}$.

## G Sensitivity analysis of the solution of the RE and the

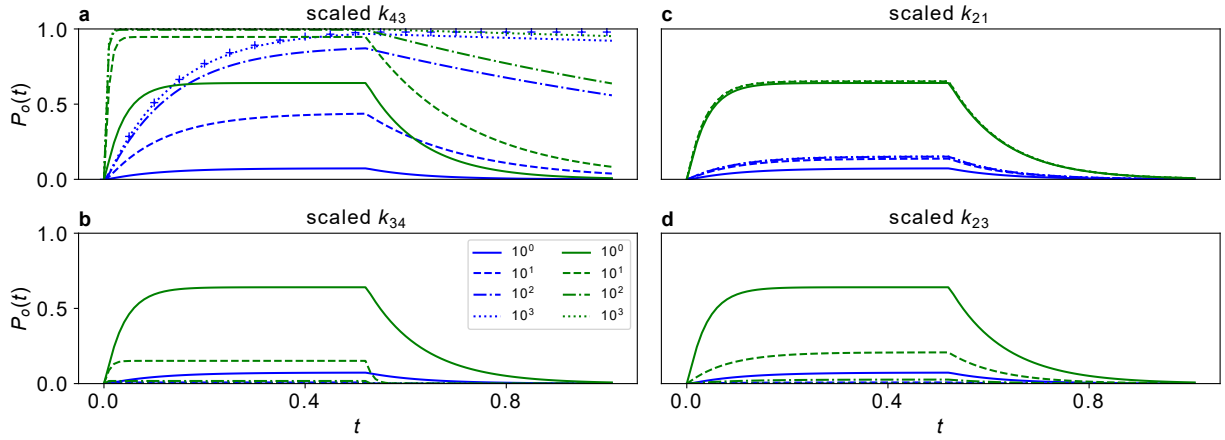


Figure 13: **Scaling the chemical rate constants of CRN1 shows that the influence on the mean signal fades, particularly when scaling the ligand-binding rate.** In each panel, one chemical rate constant $k_{ij}$ is scaled by $[10^1, 10^2, 10^3]$ to display the sensitivity of the mean kinetic time evolution of $P_o(\mathbf{K}, t)$ to that parameter scaling. Modeling $P_o(\mathbf{K}, t)$ by a deterministic RE is identical to experimenting with a noise-free experimental setup that probes an infinite number of ion channels. Two different ligand concentrations $[0.125, 1]\mu M$ are encoded in blue and green, respectively. The different lines indicate the different scaling. In panel **a**, a $10^6$ scaling (blue crosses) for the lower ligand concentration is included. It overlaps with $10^3$ scaling. Indicating that the sensitivity of kinetics ultimately also fades for $k_{43}$.

In chemical physics, it is a well-known phenomenon that often the overall rate of a process in a CRN relies heavily on the slowest chemical reaction, called the rate-limiting step. Rate limitations in inverse modeling become a relevant problem due to the aggregation of Markov states in the signal. Ideally, the mean of an observable of a Markov system would not be aggregated, which means it would report the entire $\mathbf{p}(\mathbf{K}, t) = \mathbb{E}[\mathbf{n}(t)]/N_{\text{ch}}$ the population probability vector of all Markov states. In that case, rate limitation would be a minor problem because the kinetics of each state alone allow one to dissect better the contributions of each chemical rate. Only the bandwidth of the data, the signal-to-noise ratio, and the quantity of data would limit the precision of the inference, given a correctly specified likelihood. However,

one faces $P_o(\mathbf{K},t) = \tilde{\mathbf{H}}\mathbf{p}(\mathbf{K},t)$ as the observable in PC data, with $\tilde{\mathbf{H}} = (0,0,0,1)$ given CRN1 (Fig. 1). Due to the projection, $\tilde{\mathbf{H}}$, $P_o(\mathbf{K},t)$ can be *locally* insensitive to changes in some $k_{ij}$. That means that

$$\frac{\partial P_o(\mathbf{K},t)}{\partial k_{ij}} = 0 \tag{51}$$

for all $t$.

Instead of *locally*, the *global* 1d sensitivity is scanned (Fig. 13) by scaling in each panel one of the $k_{ij}$ by $[10^1, 10^2, 10^3]$ and observing the changes of $P_o(\mathbf{K},t)$ for all $t$ while leaving the other parameters unchanged. Additionally, some finite experimental time is assumed. Note that multi-dimensionality must be accounted for in a full sensitivity analysis. However, for intuition, it seems sufficient to consider one-dimensional scalings. The deterministic $P_o(\mathbf{K},t)$ of CRN1 (Fig. 1) can be calculated by the RE approach [101]. Qualitatively, the larger the changes of $P_o(\mathbf{K},t)$ compared with $P_o(\mathbf{K}_{true},t)$ of the CRN (Fig. 13, solid lines), that was used to simulate the data; the smaller can be the data quality to infer that parameter with a satisfactorily small uncertainty, given a correctly specified likelihood. Note that in the following, a likelihood will be constructed [101] for unrealistic best-case data scenarios, whose essential parameter to predict data is $P_o(\mathbf{K},t)$. Thus, differences in $P_o(\mathbf{K},t)$ for each scaling of the $k_{ij}$, at least for some $t$, are crucial that the model does not have some degree of non-PI. Even in that best-case scenario, one encounters a non-PI problem.

$P_o(\mathbf{K},t)$ is very sensitive (Fig. 13 **a,c**) to the two gating rates $k_{43}$ and $k_{34}$, given CRN1. Firstly, because $P_o(\mathbf{K},t=\infty)$ of both of them can be scaled from 0 to 1 by scaling them from zero to infinity, respectively. Note that for visual clarity (Fig. 13 **a,c**) only scaling to larger values is considered. Secondly, both are rate-limiting in the time trace's respective activation and deactivation parts. However, the experiment's finite time resolution and duration add global sensitivity limits caused by rate-limiting effects. This can be seen for the smaller ligand concentration $L = 0.125\,\mu\mathrm{M}$, the scaling of $k_{43}$ (Fig. 13 **a**, blue curves) leads to four visually different kinetics, indicating high sensitivity. However, adding a fourth scaling $10^6$ (blue crosses), the third and the fourth are visually identical, although they are different by three orders of magnitude. For at least the applied $10^3$ scaling $\frac{\partial P_o(\mathbf{K},t)}{\partial k_{43}}|_{k_{23\geq 10^3 k_{43,true}}} \approx 0$. That means the scaled $k_{43}$ is instantaneous relative to the other involved transitions. Any higher scaling has little effect on the kinetics of the observable in the given amount of experimental time. However, changes would be detectable in the forward model for an infinite amount of relaxation time because $P_o(\mathbf{K},t=\infty)$ changes with each scaling. Thus, predictions of the model, $P_o(\mathbf{K},t)$, are rate-limited by the other $k_{ij}$ when $k_{43} > 10^3 \cdot k_{43,true}$. Despite, being locally, around $k_{43,true}$ $P_o(\mathbf{K},t)$, highly sensitive, the global sensitivity is constrained. From the experimental perspective, that sounds paradoxically because $k_{43}$ and $k_{34}$ are around $k_{43,true} = 50$ and $k_{34,true} = 10$ locally the rate-limiting steps themselves for the activation and deactivation time series (unless $L$ is very small in the activation data).

These rate-limitations are practically irrelevant for $k_{43}$ and $k_{34}$ because they are roughly of the order $10^3$ away from their true value, $k_{43,\mathrm{true}}$ and $k_{34,\mathrm{true}}$. However, in principle, they exist even for $k_{ij}$, which are locally around $k_{ij,true}$ the rate-limiting step, or at least for a set of time traces very influential on $P_o(\mathbf{K},t)$.

The global sensitivity of $P_o(\mathbf{K},t)$ changes are *visually* distinguishable in two ways when the scaling is applied on the ligand binding rate $k_{21}$ (Fig. 13 **c**). Firstly the sensitivity of $P_o(\mathbf{K},t)$, is reduced (without any effects from the boundary at $P_o = 1$) for two reasons For $P_o(\mathbf{K},t=\infty)$, the scaling reduces the probability to be in $C_1$ more and more, but the ratios between the other states $C_2$, $C_3$, and $O_4$ remain the same; hence, the sensitivity of amplitude of the current is constrained from above to smaller values than $P_o(\mathbf{K},t=\infty) = 1$. Also, $P_o(\mathbf{K},t)$ loses its sensitivity due to rate-limitations; however in this case by the actual transition $k_{43}$, which is, from the experimental perspective, the rate-limiting step of the activation. Secondly, for the higher ligand concentration (Fig. 13 **c**, green curves), one can only distinguish $10^0$ and $10^1$ but not the other scalings. For the lower ligand concentration (Fig. 13 **c**, blue curves) one can at least distinguish $10^0$ and $10^1$ and $10^2$ visibly. In contrast to the scaling of $k_{43}$ (where the larger $L$ has a higher global sensitivity), the smaller $L$ reveals a higher global sensitivity of $P_o(\mathbf{K},t)$ for scaling $k_{21}$. The smaller $L$, the larger $L \cdot k_{21}$ can be scaled without being instantaneous relative to the other $k_{ij}$ in the kinetic scheme. Note that scaling $k_{21}$ to smaller values would have no global sensitivity limit as $P_o(\mathbf{K},t=\infty)$ would approach 0. Further at some point $k_{21}$ would be the rate-limiting step in the activation for any $L$ but this time in contrast to $k_{43}$ up the finite limit of the used relaxation time is not even theoretically relevant.

One of the advantages of fitting multiple concentrations (Fig. 13 **a**, green curves) of ligands (or, depending on the molecule, other series of stimuli) that the limiting values for $k_{ij}$, above which rate-limitations eliminate all effects of altering $k_{ij}$ on the likelihood can be tweaked by changing the magnitude of the stimulus. Hence, the local and global sensitivity of $P_o(\mathbf{K},t)$ to altering some $k_{ij}$ is changed.

Similarly, $P_o(\mathbf{K},t)$ has a high sensitivity for scaling up the ligand-unbinding rate $k_{23}$ (Fig. 13 **d**) because the scaling pushes $P_o(\mathbf{K},t) \to 0$ for all $t$.

Summing up, there is a global sensitivity limit of the mean kinetics and equilibrium to scaling up $k_{21}$ and similar $k_{32}$. However, for scaling $k_{43}$ up, the global sensitivity limit also requires a finite relaxation time in the experiment. However, a global sensitivity limit for scaling up $k_{34}$, $k_{23}$, and $k_{12}$ does not exist. And similarly, scaling one of the down $k_{21}$, $k_{32}$ and $k_{43}$ will decelerate the coupled reaction more and more and $P_o(\mathbf{K},t=\infty)$ will approach 0 such

that the sensitivity of $P_o(\mathbf{k}, t) \approx 0$ to scaling these rate constants to smaller values is high. These asymmetries in the sensitivity will produce highly asymmetric likelihood profiles (Fig. 14). If one had one independent observable for each Markov state, meaning that the full mean $\mathbf{p}(\mathbf{K}, t)$ is an observable (i.e., no aggregation of states in the fictitious signal); the limitations due to partial observability would be alleviated. In this optimal case, only the signal-to-noise ratio and the bandwidth would induce non-PI structure in the likelihood. Adding orthogonal signals fluorescent signals such as done with cPCF [89] is the attempt to expose more of $\mathbf{p}(\mathbf{K}, t)$ than $P_o(\mathbf{K}, t)$ does.

### G.1 HMMs, even under unrealistic optimal data and prior knowledge scenarios, have a degree of a non-PI problem.

The following section scrutinizes the likelihood of PC data of the 4-states-1-open-state model for unrealistic optimistic data and prior knowledge scenarios. It is demonstrated that in these best-case scenarios, the likelihood has a degree of a non-PI problem, which indicates that HMM inference has, in general, a non-PI problem.
A non-PI problem exists due to an infinite area in $\boldsymbol{\theta}$ where the likelihood is insensitive to changes of some $k_{ij}$'s or $\tau_i^{-1}$'s. The reasons for the non-sensitivity are further investigated in App. G by comparing the deterministic mean kinetics, $P_o(\mathbf{K}, t)$, for different $\mathbf{K}$. The combination of rate-limitations with the limited influence of some $k_{ij}$ on the equilibrium open probability, $P_o(\mathbf{K}, t = \infty)$, causes $P_o(\mathbf{K}, t)$ to lose its sensitivity towards changes of the chemical rates in an infinite volume in $\boldsymbol{\theta}$. One might consider it as intrinsic inferential limits of the protein's model likelihood. Inferring some of the $k_{ij}$'s with a low uncertainty defines for the other more uncertain $k_{ij}$'s where the likelihood becomes completely insensitive in $\boldsymbol{\theta}$ for the value of these $k_{ij}$'s. where in $\boldsymbol{\theta}$ the likelihood becomes completely insensitive for these $k_{ij}$'s. Ultimately, this causes the likelihood of partially observed CRNs to have a degree of non-PI even for completely experimental noise-free data.

### G.2 The fading sensitivity of some chemical rates on the observed kinetics due to rate limitations induces practical non-identifiability of the model.

When modeling experimental data, the intrinsic inferential limits interact with the data's bandwidth, experiment duration, and signal-to-noise ratio. Nevertheless, it is shown herein that the intrinsic inferential limits of a CRN might become the bottleneck for the degree of non-PI for many $k_{ij}$'s or $\tau_i^{-1}$'s the entire model.
In the ML context, it is one prerequisite for entire PI in contrast to non-PI [120] that the parameter values of likelihood surfaces of constant value form compact sets. In a 1D setting, one could formalize this prerequisite to that

$$\lim_{k_{ij} \to \infty} \log(p(y \,|\, k_{ij})) = -\infty \tag{52}$$

holds. Similarly, $\lim_{k_{ij} \to 0} \log(p(y \,|\, k_{ij})) = -\infty$. In other words, the model's predicted statistics, given the set of parameters, have to be further and further away from the true statistics of the data if some parameters are more and more diverging. When sampling a posterior, given a minimal informative prior, with a finite sampling box, one assumes that Eq. 52 is sufficiently fulfilled at the limits of the sampling box.
The discussion of the forward model (App. G) indicates that the likelihood of macroscopic HMM data beyond trivial examples has the opposite of Eq. 52 as a general feature: If the global sensitivity of $P_o(\mathbf{K}, t)$ for some scaling magnitude is vanishing, then the derivative of the log-likelihood must approach $0$.
In the following, a simulated best-case data and prior knowledge scenario shows that non-PI features arise in a likelihood which is an exact representation of the statistics of the simulated data. The non-PI problem becomes essential for maximum likelihood estimation or Bayesian statistics at poor data quality. Nevertheless, it is always present in principle. Again, this does not mean there is no optimal parameter set in the posterior or likelihood, but the corresponding uncertainty quantification faces severe problems (Fig. 2). Suppose the likelihood of the model used for the inference has a unique global optimum but becomes flat in some directions. In that case, the inference quality must be judged by how high the likelihood's maximum is relative to its flat area. In the case of Bayesian statistics with the applied Jeffreys prior, one can transform this criterion into how much the posterior decays (5 **a,b**) before its slope is approximately the slope of the log-uniform prior.
Corresponding to the profile likelihood technique [95], one may define that, for example, $\mathbb{R}(\tau_i^{-1})$ needs to be decayed to $10^{-2}$ relative to the value of the posterior's peak before its slope drops to $1.05$ of the log uniform prior slope to call the parameter identified. Then $N_{ch} = 10^2$ (5 **a**) would be by definition not be considered identified, even if the posterior has some interval which is clearly distinguished from the flat rest of $\boldsymbol{\theta}$. In contrast, if we define that the posterior needs to decay at least to $10^{-1}$ before its slope drops to $1.05$, then it would be considered as identified. One needs this definition based on how high is the peak of the posterior relative to the prior dominated area to justify the limits of the sampling box after one has done the sampling. That is because outside of the sampling box is always an *infinite amount of probability mass* due to the non-PI problem, given that an improper prior (Eq. 2) is used.

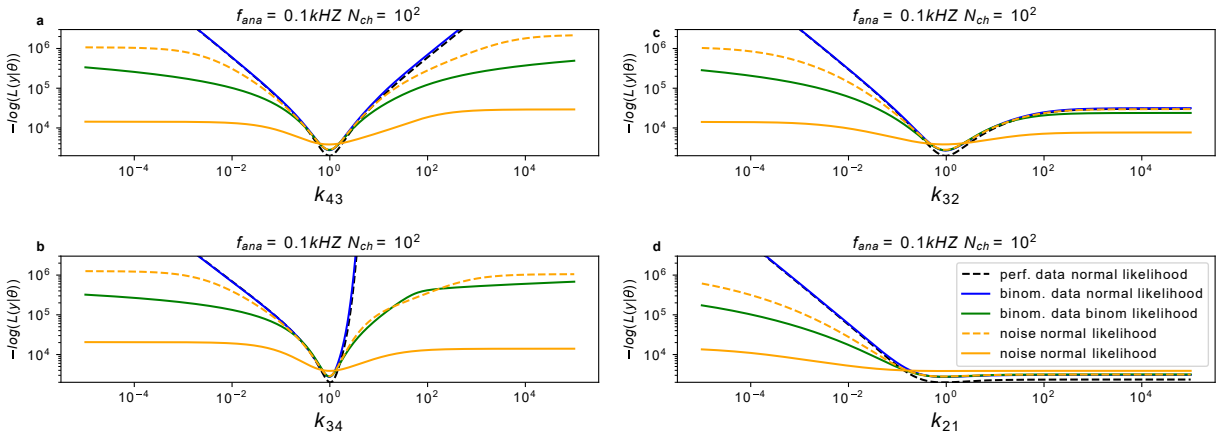


Figure 14: **Likelihood Profiles of four different parametersof CRN1 reveal varying degrees of non-PI even with instrumental noise-free data** Here, we show likelihood profiles, given different data settings and corresponding likelihoods for different $k_{ij}$. **a-d,** The black dashed curve indicates the value of a normal log-likelihood analyzing the perfect data (no experimental and intrinsic noise). The perfect data are the result of calculating the deterministic solution of the RE. The blue line represents binomial data analyzed with a normal likelihood. Green represents the binomial data analyzed with binomial likelihood. We add white normal distributed noise to the binomial data (orange curves) with $\sigma = 1$ (dashed curve) $\sigma = 10$ (solid curve).

### G.3 The profile likelihood reveals flat parts even in a best-case scenario.

We generate idealized model data sets from CRN1 (Fig. 1) and then add in a step-wise manner intrinsic fluctuations of the ensemble and instrumental noise to the synthetic signal. We calculate for each time point the open probability $P_o(\mathbf{K}, t)$ of the model and derive from it the mean signal

$$\mathbb{E}[y(t)] = N_{ch} i P_o(t) \tag{53}$$

and its variance

$$\text{var}[y(t)] = \sigma^2_{intrin} = N_{ch} i^2 P_o(t)(1 - P_o(t)) \tag{54}$$

utilizing the multinomial approximation of the distribution of stochastic gating and binding of ion channels in macroscopic patches [101]. The initial preparation of the relaxation experiment assumes that all channels are at $C_1$ because the ligand concentration holds $L = 0$. Further, only one data point per time trace is used per activation and relaxation experiment With these conditions, the multinomial assumption used in [101] precisely describes the data's statistics. Otherwise, the actual likelihood, given the Markov assumption, must be based upon a convolution of multinomial distributions [74, 41].

As asymptotically ideal data, we use the true model's mean current signal (Eq. 53) without any intrinsic or extrinsic noise (Fig. 14, black dashed curves). Then, we additionally sample the stochastic time evolution by drawing from a binomial (Eq. 28) distribution $n(t) \sim \text{binomial}(N_{ch}, P_o(\mathbf{k}, t))$ current data for each time step (Fig. 14, blue and green curves) and finally add some experimental white noise with constant variance (Fig. 14, orange curves). In that way, we generate data sets that match the white noise assumption of the RE approach [101], which differs from simulating data by the Gillespie algorithm. In contrast to the Markov property (i.e., included in the Gillespie algorithm) of actual experimental time series data, we switch the MM property off. The MM property lowers the information content of a time trace (see, [101] Fig. 10). Hence, the used data sets contain more information than actual HMM data with the same sampling frequency. Consequently, the following analysis is a best-case data scenario corresponding to a massive bootstrapping of the data: one data point per time series (see, [101] Fig. 10).

These data sets are analyzed by assuming different likelihoods. For the perfect data set (Fig. 14, black dashed curves), we use a normal log-likelihood which inherits its mean signal and variance from the above-stated multinomial assumption of the RE approach [101]. The blue curve represents the log-likelihood value for binomial data analyzed with a normal log-likelihood and, correspondingly, analyzed with a binomial log-likelihood (green curve). We fix all parameters but one to their true values and plot the profile log-likelihood vs. the scaled $\tilde{k}_{ij} = k_{ij}/k_{ij,true}$ parameter. This way, a very informed prior knowledge of the rate matrix is assumed, as every other relevant parameter except the scaled one is precisely known. Our inverse modeling scenario corresponds to the forward modeling scenario discussed in (Fig. 13). In contrast to [95, 137, 136], the other parameters are not optimized such that the profile likelihood herein

displays the contour along a straight line in $\boldsymbol{\theta}$ while their procedure searches for the slowest decay of the likelihood given one fixed parameter as a constraint, which is not necessarily a straight line; it is some lower dimensional, potentially curved submanifold in $\boldsymbol{\theta}$. However, the approach used herein seems sufficient for displaying pathologies of the 1D likelihood. Note that the profile of the log-likelihood is identical to the profile of a 1D log posterior derived from a uniform prior distribution. Nevertheless, if more chemical rates are unknown, likelihood profiles are generally not identical to the 1D marginalized posterior.

In the first column (Fig. 14 **a,b**) we investigate $k_{43}$ and $k_{34}$ the opening and closing rates of the channel, respectively. Even if it is not displayed, there exist inevitably an upper limit of $-\log(p(y\,|\,k_{ij}))$ for scaling $k_{43}$ because of the observed loss of sensitivity of $P_o(\mathbf{K},t)$ "occurring" above scalings of $10^3$ (Fig. 13 **a**). That is because the normal likelihood uses only the mean signal and its variance as parameters, which are functions of $P_o(\mathbf{K},t)$. Ultimately, $P_o(\mathbf{K},t)$ becomes insensitive even to changes of large $k_{43}$ and $k_{34}$ (Fig. 13 **b**, green curves). Hence $-\log(p(y\,|\,k_{ij}))$ cannot go to infinity. Correspondingly to the large interval, in which $P_o(\mathbf{K},t)$ is sensitive (Fig. 13 **a**) the limiting behavior of $-\log(p(y\,|\,k_{ij}))$ is beyond the limits of panel Fig. 14 **a**.

Note that the more accurate binomial log-likelihood reveals a slower increase of its negative log-likelihood values (Fig. 14 **a-d**). This means that the normal approximation rules out large deviations from the true value more strictly than the true likelihood. The difference between the binomial and normal log-likelihood is reminiscent of the breakdown of the normal approximation to the binomial distribution when $P_o(\mathbf{K},t)$ is close to 1 or 0. The vanishing sensitivity of $P_o(\mathbf{K},t)$ of other rate constants for $k_{34}$ and $k_{43}$ are (Fig. 14 **a,b**), in the considered case, not the bottleneck as additional experimental noise (Fig. 14 **a** orange curves) strongly affects the depth of the minimum. If we add some noise with $\sigma_{exp}$, the total variance of the signal will be $\mathrm{var}[y] = \sigma^2_{intr} + \sigma^2_{exp}$. With a small amount of extra experimental noise (Fig. 14) **a,b**, orange dashed curves), the likelihood profile for $k_{34}$ and $k_{43}$ becomes flat, giving each of the parameters a more noticeable degree of non-PI. As expected, a larger magnitude of the noise (orange, solid curves) makes the minimum more shallow.

Of course, the resort in practice is that the minimum negative log-likelihood is deep enough. It is still almost an order of magnitude deep, and the minimum is of the order of $10^4$ such that in the space of $\exp(\log(p(y\,|\,k_{ij})))$ where one evaluates parameter non-identifiability, the corresponding relative height of the likelihood's maximum is even higher above the flat parts. That enables almost arbitrary high-confidence thresholds [95] in the probability space. The word 'almost' seems not even worth a technical note, given that the data used to calculate the log-likelihood profile creates such a low minimum in panels a and b. Still, it is impossible to choose an arbitrarily high confidence threshold (Fig. 2, black dashed lines) in the maximum likelihood framework that delivers a finite confidence region for the parameter $k_{43}$. Bayesian analysis with a Jeffreys prior would technically have the same non-PI feature, leading to an improper posterior on $k_{43} \in (0, \infty)$.

The non-PI aspect, induced by the rate limitations due to other finite chemical rates, of the HMM becomes relevant when considering $k_{32}$ and $k_{21} = \tau_1^{-1}$ (Fig. 14 **c,d**), respectively under identical experimental conditions. The right-hand side of the likelihood is flat even under the perfect data assumption because $P_o(\mathbf{K},t)$ reached its global sensitivity limit (Fig. 13, **b**). In other words, it is more suited for inverse modeling if there is a global sensitivity limit, the likelihood based on a rate matrix with an infinitely scaled $k_{32}$ or $k_{21} = \tau_1^{-1}$ does not become an infinitely bad predictor of the data, but that is what the Eq. 52 requires in order that the model has no degree of non-PI. Still, if the data quality is enough, the minimum of the negative log-likelihood might be deep enough.

If one extrapolates these observations, one can conclude the following: On the one hand, the more unknown parameters the rate matrix has, the more flexibility might the likelihood have to repeal the effect of one misplaced parameter in the signal by tweaking other parameters; hence, more or higher-dimensional submanifolds might emerge in $\boldsymbol{\theta}$ on which the likelihood profile (surface) is flat. If one needs to learn the full rate matrix, the $k_{mn}$, which influences the data the most, given the tested conditions, will have the sharpest marginal posterior. That defines where the marginal posteriors of the other $k_{ij}$ become flat. Because at some value of $k_{ij}$, the likelihood will have lost all sensitivity to the exact value of $k_{ij}$. This does not mean that these $k_{ij}$ do not have a peak somewhere else. On the other hand, it is straightforward to construct with more unknown chemical rates a direction in the higher dimensional $\boldsymbol{\theta}$ where the rate limitation of a single chemical constant by other chemical rate constants is repealed by scaling others, too. Assuming a scaling acting simultaneously on $k_{21}$, $k_{32}$, and $k_{43}$ and that the other chemical rates are fixed at their true value, The scaling means moving the chemical rates along a diagonal in $\boldsymbol{\theta}$. Moving along this direction the likelihood has the desired $\lim_{k_{21}\to\infty}\lim_{k_{32}\to\infty}\lim_{k_{43}\to\infty} p(y\,|\,k_{21},k_{32},k_{43}) = 0$ for data, which is uncorrupted from experimental noise. In this limit, the kinetics become a step function from $P_o(t=0) = 0$ to $P_o(t=\Delta t) = 1$. Given the finite bandwidth of real experimental data, this happens already for finite $k_{21}$,$k_{32}$, and $k_{43}$. In contrast, for each direction along the parameter axis $k_{12}$, $k_{23}$, or $k_{34}$ alone $\lim_{k_{ij}\to\infty} p(y\,|\,\theta)) = 0$. So, we can deduce from the model structure by the forward model at least four trivial directions in $\boldsymbol{\theta}$ for which the likelihood behaves rather PI-like, given ideal data without any experimental noise. The existence of some PI-like directions in the parameter space is, of course, not sufficient to make the model entirely PI.

In conclusion, in a Bayesian framework, the posterior sampled in a sufficiently large sample box with enough samples would always be dominated by the prior in the right tails of the inverse dwell times. Thus, if we try to be uninformative

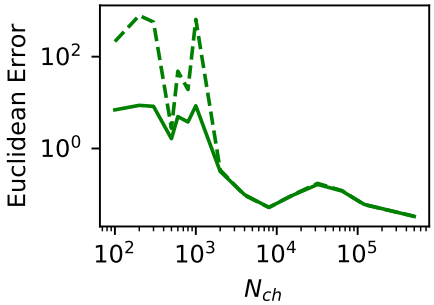


Figure 15: **Comparing the two RSME definitions based on the previously used CRN1 PC data.** The solid line corresponds to the RSME on the log space. The posterior was sampled using a uniform prior. The two RSME definitions are equivalent for $N_{ch} \geq 2 \cdot 10^3$. The two definitions deliver indistinguishable results at $N_{\mathrm{ch,crit}}$.

by equipping the time scales with a log-uniform prior and setting the sampling box limits to huge values, we end up with a posterior, which depends accidentally on the nonphysical limits of the sampling box. Strictly speaking, this is true for all parameters. The depth of the log-likelihood minimum around the actual value ensures that this is unnoticed in most data scenarios. That said, posteriors derived from the minimal informative prior still have a peak (Fig. 5**a-b** around the true value and cover the true value within a considerable part of their peak; thus, we still gained information. Notably, in the strong data, case [41], without knowing that the likelihood becomes flat in the inverse time-like parameters and without any elaborated prior selection, we showed in [41] that the HDCVs have the expected coverage frequency over the true values. In other words, the calculated Bayesian uncertainty of the parameters matches the frequency of finding the parameters within a volume to a high degree if the algorithm correctly maps the first-order Markov property of the data. Thus, cutting away infinite "probability mass" at some data strength is just fine because the rest behaves as if the posterior is well-defined. In contrast, in the weak data case where the posterior drops only a limited ratio before it is "flat," the limits of the samples are more critical. One could define thresholds that sort out samples in the flat likelihood area and define the posterior from the remaining, equivalent to sampling it with a smaller sampling box. It is beyond the scope of this article to investigate the necessary numerical criterion, what is flat enough that one should consider the part of the posterior as dominated from the prior. The performance of such an approach is also out of the scope of this article.

### G.4 Why is the improper posterior usually unnoticed?

The improper posterior is usually undetected as a limited amount of samples, such as $10^5$ of the posterior, would hardly have explored the volumes in $\boldsymbol{\theta}$, which are of the order $p \approx 10^{-7}$ or even smaller. Note that the improper posterior appears for the sampler as a proper posterior due to the finite sampling box. That is because posterior samplers explore $\boldsymbol{\theta}$ such that in the limit of infinite samples, each volume in $\boldsymbol{\theta}$ is visited proportional to the probability mass inside that volume. Any finite sample amount leads to high errors for volumes in $\boldsymbol{\theta}$ with low probability mass (identical to histograms with too narrow bins). Often, the sampler will not even have visited once this part of $\boldsymbol{\theta}$. The distribution that a sample hits a specific volume (bin) in $\boldsymbol{\theta}$ is always multinomial if one assumes that the sample trace has been thinned until the autocorrelation of the sample trace is eliminated. This is independent of the actual shape of the posterior. The probability of hitting that bin for an individual sample is the integral over the posterior inside the limits of the bin.

## H Some considerations when the RSME should be evaluated on the logarithm of the parameters

To calculate the RSME of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, the samples are reparametrized back to $k_{ij}$'s for the following reason. The probability, $\epsilon_{ji}$, which Markov transition is chosen, exists on a probability simplex (Eq. 31). Relative errors from $\mathbf{K}_{\mathrm{true}}$ cannot be unambiguously merged into the RSME as long as the boundaries are close. Nevertheless, similar problems arise with chemical rate constants at $k_{ij} = 0$. A 10 times larger parameter value contributes with a relative error of 9 according the definition in [41] while a 10 times smaller parameter value contributes only with 0.9 relative error. However, it is reasonable to care more for the multiplicative deviation $\mathbb{E}[k_{ij}/k_{ij,\mathrm{true}}]$ because $k_{ij}$'s are, in the statistical sense, inverse scaling parameters. When the information in the data is weak, one expects large multiplicative

deviations from $\mathbf{K}_{\text{true}}$ and correspondingly a diffuse $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, which are influenced by the mathematical constraints. $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, will be far away from multinomial distributions. Hence, a redefinition (Eq. 24) compared to the definition of the RSME of [41], seems obligatory. The boundaries at zero are eliminated by log transformation. The mean value is the standard point estimator in Bayesian statistics. It converges towards the peak of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ if the distribution becomes monomodal and symmetrical. Nevertheless, it is shown (inset, Fig. 5) that reporting the peak of the marginal distributions due to the non-PI problem makes sense.
Note that for small multiplicative deviations (i.e., $\mathbb{E}[k_l]/k_{l,\text{true}} \approx 1$) and using a Taylor series expansion, one can show that

$$\text{Error}(\log \mathbb{E}[\boldsymbol{k}]) = \sqrt{\sum_l (\log(\mathbb{E}[k_l]/k_{l,\text{true}}))^2} \approx \sqrt{\sum_l (\mathbb{E}[k_l]/k_{l,\text{true}} - 1)^2} = \text{Error}(\mathbb{E}[\boldsymbol{k}]), \tag{55}$$

holds. Hence, the redefinition becomes equivalent to the of the RSME in [41]. In aid of illustration, we use the PC data (Fig. 1) and compare $\text{Error}(\mathbb{E}[\boldsymbol{k}])$ of the definition in [41] with the new definition. As previously done, $N_{\text{ch}}$ is used as a measure for the PC data quality (Fig. 15); thus, also as a proxy for the information contained in the data for a given model. The transition (Fig. 15) that both definitions become equivalent occurs roughly at $N_{\text{ch}} < 2 \cdot 10^3 = N_{\text{ch,crit}}$. Below $N_{\text{ch,crit}}$, the RSME of the logspace is compared to the error on chemical rate space $1-2$ orders of magnitude smaller. Only minor differences can be distinguished above $N_{\text{ch,crit}}$.

### H.0.1 The learning of the model on the non-log parameter space

The influence of the Jeffreys rule inspired prior with diffusion limit on the single parameter level (Fig. 16) is visualized by HDCIs (Eq. 7) vs. $N_{\text{ch}}$. For clarity, we plot for the uniform prior (green dashed lines) and for the Jeffreys prior (gray area) only the 0.95-HDCI. However, a series of HDCIs is plotted for the diffusion limit prior (blue area). The log-space is not used; instead, a log-log scaling is applied to indicate the skewness of the marginal posteriors. Due to the skewness of the marginal posterior distributions, the HDCIs differ from quantiles. The posterior of $\tilde{\tau}_1^{-1}$, which is also identical to $p(\tilde{k}_{12} \,|\, \mathcal{Y}_T)$ (blue posterior), since $\tau_1^{-1} = k_{21}$ is cut by the diffusion limit unless the data quality is very high. The additional information allows confining the samples to a much more restricted space producing decent inferences even at $N_{\text{ch}} = 10^2$. The bias towards too large $\tau_1^{-1}$ of $p(\tau_1^{-1} \,|\, \mathcal{Y}_T)$ using the minimally informative prior now turns into a bias to too small $\tau_1^{-1}$. The applied constraints on $k_{32}$ have also for $p(\tau_2^{-1} \,|\, \mathcal{Y}_T)$, to which the second binding rate $k_{32}$ contributes by $\tau_2 = k_{32} + k_{12}$, a regularizing influence. That is true even though it cuts only through the far tail $p(k_{32} \,|\, \mathcal{Y}_T)$ (not shown here). In the regime $N_{\text{ch}} \in [10^2, 10^3]$, the constraint makes a very vague $p(\tau_1 \,|\, \mathcal{Y}_T)$ where the 0.95-HDCI covers up to two orders of magnitude, including a potential non-PI problem, to a decent $p(\tau_1 \,|\, \mathcal{Y}_T)$. The 0.95-HDCI spans $[0.4, 1.16]$ in units of $\tilde{\tau}_1$. The highly skewed distributions justify the redefinition of the Euclidean error, the log space.

## I More data sets for CRN2 showing the gain of robustness by the MI prior

We continue investigating the benefits and challenges of the MI prior (continuing the discussion of Fig. 7) arising when the complexity of the CRN increases. In Fig. 17 **a** the pathological dimensions of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ derived from the uniform $p(\mathbf{K})$ (colored) against the same dimensions of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ (black) derived from the MI $p(\mathbf{K})$ are compared. The less pathological dimensions in the sense that they deliver roughly Gaussian marginal posteriors are plotted in Fig. 17 **b1-6**. We reduce the data quality for the MI prior (to create a more challenging scenario), showing the overall gain in the robustness of the inference by the MI prior. We further have to work with smaller sampling boxes (which still contains $\mathbf{K}_{\text{true}}$) for the uniform prior (because of an often not converging sampler). Using a smaller sampling box for the uniform prior also disadvantages the MI prior, see Arguments in Sec 4.6. Thus, the *sufficiently proper* posterior for $N_{ch} = 5 \cdot 10^5$ using the MI prior, shows the overall robustness gain due to the MI prior. For $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ based on a uniform $p(\mathbf{K})$ we use a set of data sets with practically *impossible* to achieve data quality $N_{\text{ch}} = [7.5 \cdot 10^4, 10^5, 2.5 \cdot 10^5]$ for PC experiments.
For the uniform prior, $p(\log_{10}(\tau_2^{-1}), \log_{10}(\tau_3^{-1}) \,|\, \mathcal{Y}_T)$, derived from $N_{\text{ch}} = 10^5$ per patch, still displays the non-PI problem visually. Hence, $\tau_2^{-1}$ and $\tau_3^{-1}$ should be considered unidentified, based on the herein-employed heuristic criterium (see, discussion around Eq. 4) to define $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ as *sufficiently proper* and the parameter identified if the existing non-PI pathologies are not detected *visually*. However, $p(\log_{10}(\tau_2^{-1}), \log_{10}(\tau_3) \,|\, \mathcal{Y}_T)$ of the data set with 20 times smaller $N_{\text{ch}}$ and based on the MI $p(\mathbf{K})$ does not display the non-PI problem. In contrast, for the next smaller data set (not shown here), with $N_{\text{ch}} = 2 \cdot 10^3$ also $p(\log_{10}(\tau_2^{-1}), \log_{10}(\tau_3^{-1}) \,|\, \mathcal{Y}_T)$ derived from the MI $p(\mathbf{K})$ starts to display the non-PI problems visually on a linear scale. For $N_{\text{ch}} = 2.5 \cdot 10^5$, $p(\log_{10}(k_{32}), \log_{10}(\tau_3) \,|\, \mathcal{Y}_T)$ updated from the uniform $p(\mathbf{K})$ does not display the non-PI problem visually anymore. Hence, the MI $p(\mathbf{K})$ increases the range

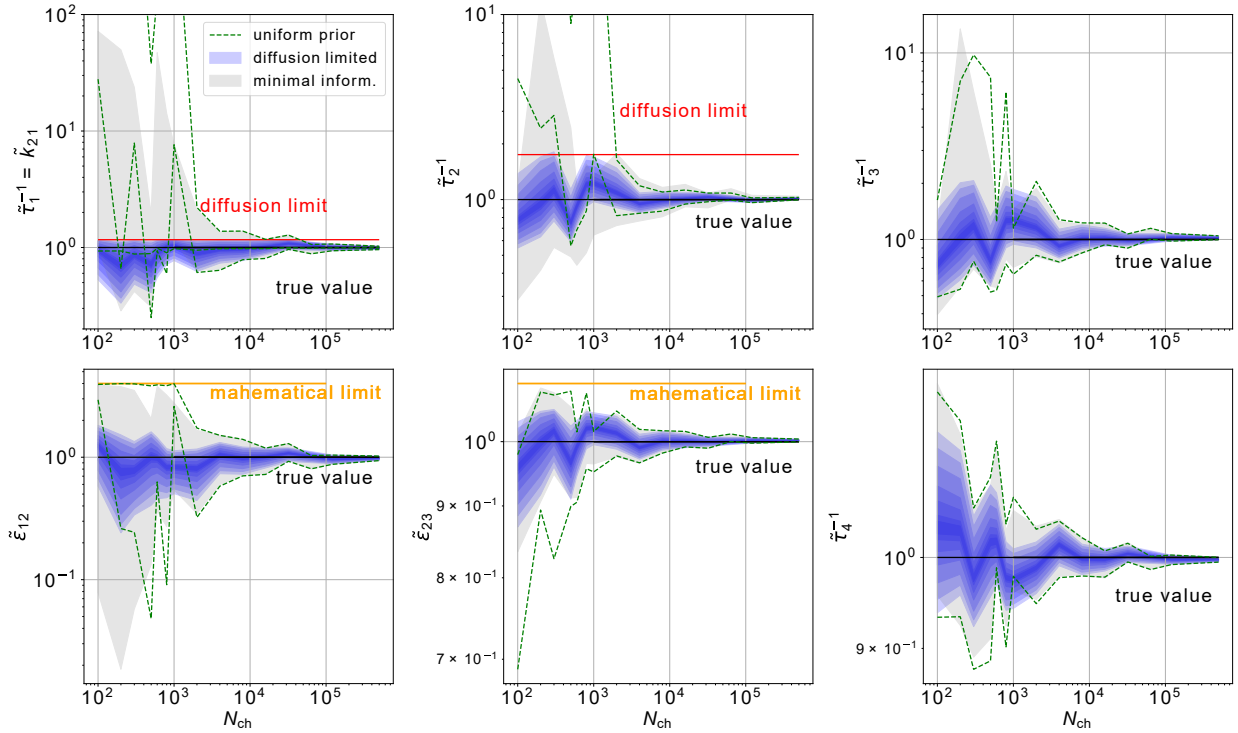


Figure 16: **The marginal posterior quantiles of CRN1 vs. $N_{\text{ch}}$ displays an influential diffusion limit even for high** $N_{\text{ch}}$. The red horizontal bars indicate the physical diffusion limit applied in sampling the blue posterior. The black horizontal bars indicate the true values. The orange horizontal bars indicate the mathematical limit that probabilities hold $< 1$. Due to the scaling to each true parameter, this limit is $\neq 1$. The green marginal is a 0.95 HDCI with a uniform prior over $\mathbf{K}$. The gray area indicates the 0.95-HDCIs using the minimally informative prior with no diffusion limit applied. The set of posterior $[0.95, 0.9, 0.8, 0.6, 0.4, 0.2, 0.1]$-HDCIs using the minimally informative prior with enforced diffusion limit is shown in blue. Due to the structure of the CRN (only one transition leaves $C_1$), $t_1^{-1} = k_{21}$ holds. Hence, the diffusion limit is identical on both parameters. For the second ligand binding $k_{32}$, $k_{32} \sim \text{log-uniform}(\text{k}_{32})$ is modeled, and the diffusion limit is enforced on this parameter. As $\tau_2 = \sum_j k_{j2}$ is shown, a probability mass above the red bar is possible.

of acceptable data for this CRN at roughly a 50 (at least a 20) fold, likely more for this DGP. Note that $\tau_1^{-1}$ is little influenced by the non-PI problem in $\tau_2^{-1}$ and $\tau_3^{-1}$. Just as naively expected, the uniform prior based $p(\log_{10}(\tau_1^{-1}) \,|\, \mathcal{Y}_T)$ are more narrow and closer to $\tau_{1,\text{true}}^{-1}$ as these posteriors originate from higher quality data. Note that $k_{21} = \tau_1^{-1}$ holds because only one Markov transition leaves $C_1$ (CRN2, Fig. 1 **d**)

### I.0.1 Finding a needle in an *almost* infinite space: The MI prior helps the sampler to converge.

To have a decently converging HMC sampler, we used for the sampling of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ based on uniform $p(\mathbf{K})$ an about ten times smaller sampling box in the direction of $k_{32}$ and $\tau_3$. It is plausible that ridges with a bimodal structure, $p(\log_{10}(k_{32}), \log_{10}(\tau_3) \,|\, \mathcal{Y}_T)$ (green and magenta Fig. 17), cause the convergence problems. The larger the simulation box in the directions of $k_{32}$ and $\tau_3$ the more would the peak around the true value and the ridge between the peaks lose its probability mass until an independently started chain [138, 139, 140] typically explores only one concentration of probability mass or the other but not both. At some size of the sampling box, the peak of the posterior around the true $\mathbf{K}$ will be like a needle in the vast constant parts of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ and not be found by the sampler, unless one changes from uniform to a MI prior, giving the constant parts of the posterior the slope of the MI prior. Inevitably, the second peak exists due to the $\log_{10}$ transformation of the flat posterior in the linear space to the posterior on the log space. The marginal posterior on this part of $\boldsymbol{\theta}$ without the logarithmic transformation is flat (Fig. 17 **c.1, c.2**). Note that a comparison with a larger sampling box would disfavor the uniform prior even more because of the non-PI problem. The relative positive effect on the range of acceptable data quality of the MI prior grows with the use of larger sampling

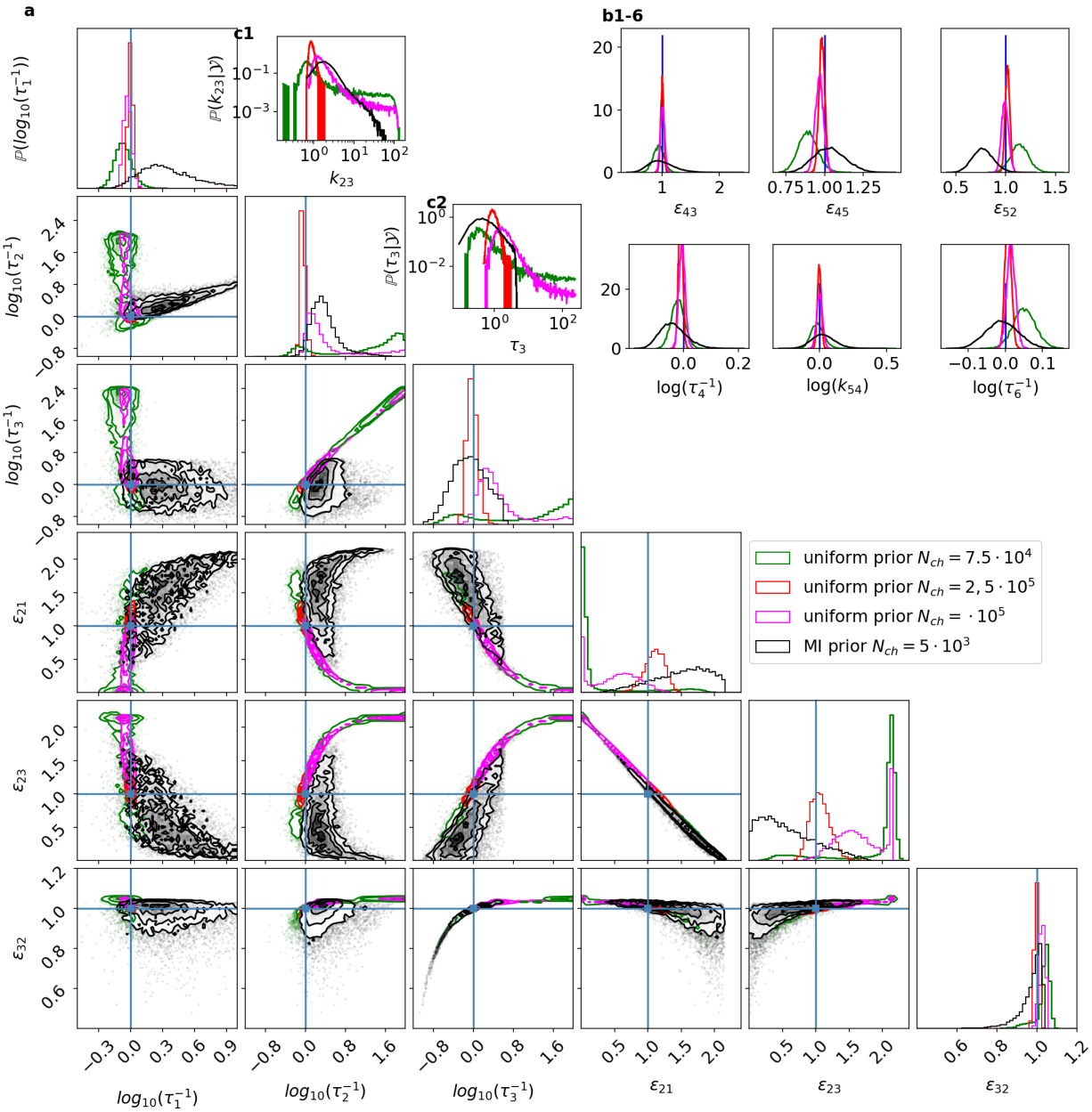


Figure 17: **The MI prior increases the data range for a proper posterior for CRN2 by a factor of more than 20. a,** Four posteriors are compared to show the overall advantage in practical identifiability (*sufficiently proper* posteriors) gained by the MI prior. Note that the colored posteriors are all using higher $N_{\text{ch}}$ and smaller simulation boxes, which makes it, in principle, easier for these posteriors to hide their non-PI pathology, resulting in an improper posterior. We suppress all marginal posteriors of somewhat boring parameters. *Boring* in the sense that they deliver more or less Gaussian-shaped marginal posteriors under all tested conditions. The black posterior derived from $N_{\text{ch}} = 5 \cdot 10^3$ is equipped with the MI $p(\mathbf{K})$. The colored posteriors use a uniform $p(\mathbf{K})$. They used data sets with $N_{\text{ch}} = [2 \cdot 10^4$ (green)$, 10^5$ (magenta)$, 2.5 \cdot 10^5$ (red)$]$ equipped with the uniform $p(\mathbf{K})$. The flat likelihood for $k_{32}$ and $\tau_3$ creates an exponential increase for the posterior of $p(\log_{10}(k_{32}) \,|\, \mathcal{Y}_T)$ and $p(\log_{10}(\tau_3) \,|\, \mathcal{Y}_T)$. Note that the simulation box for the posterior samples of $\tau_2 \in [10^{-2}, 10^{6.5}]$ updated from the minimal information is more than an order of magnitude larger than the range of the simulation box used when the uniform $p(\mathbf{K})$ is employed. The posteriors derived from the uniform $p(\mathbf{K})$ are sampled by $k_{ij}$ and then mapped to the $(\boldsymbol{\tau}, \boldsymbol{\epsilon})$-space **b,** prior that is used to allow only for minimally violated microscopic reversibility.

boxes.
As mentioned above, the limits for a proper enough $p(\mathbf{K} \,|\, \mathcal{Y}_T)$ based on the uniform prior is somewhere between $N_{\text{ch}} = 10^5$ and $N_{\text{ch}} = 2.5 \cdot 10^5$. Unfortunately, using the visual criterium based on the uniform prior requires placing the sampling box limits with a lot of luck around the peak of $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, which might be possible if before a $\boldsymbol{\theta}_{\text{ML}}$ is performed. However, this is not helpful when inferencing HMMs from actual experiments with the typical data amount typically much below $N_{\text{ch}} = 10^4$ and data imperfections.
We made this compromise of sampling the posterior based on the uniform prior in a smaller sampling box to be able to get a reasonably converging HMC sampler in the sense that four independently started sampling chains result in the same $p(\mathbf{K} \,|\, \mathcal{Y}_T)$, as judged by the Gelman-Rubin [138, 141] statistic $\hat{R}$. For $N_{\text{ch}} = 2 \cdot 10^4$ and the uniform $p(\mathbf{K})$ the sample traces (green posterior) have the desired $\hat{R} = 1$ calculated over the four independently started sampling chains for all parameters. However for $N_{\text{ch}} = 10^5$ and $N_{\text{ch}} = 2.5 \cdot 10^5$ one chain out of 4 chains starts to show a different sample trace. In principle, calculating the posterior from only one independent chain would be enough if one would know that this chain represents the typical set of the posterior [70] and has a large enough effective sample size [140]. If we only calculate the posteriors from the three seemingly convergent chains, the posteriors (magenta and red posterior) are well within the predicted probability mass of the posterior ($N_{\text{ch}} = 2 \cdot 10^4$, green). That is the expected behavior of the posterior of a much larger amount of data. Calculating $\hat{R}$ from the three similar chains $\hat{R}$, we obtain the demanded $\hat{R} = 1$. The MI prior alleviates these challenges. Note, that because for the simulated data we know the ground truth of the parameters such that we have a third independent criterion judging the remaining sampling chains. Does the posterior derived from them cover the true values.

## J The curse of dimensionality for scaling parameters

A problem to be aware of is *the curse of dimensionality*. Assume that one defines a cubic sampling box of length $L$ with $\tau_i \in (0, L)$ and uniform $p(\boldsymbol{\tau}) = \frac{1}{L}^M$ prior for $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$. Then, one may define a smaller cube inside of the sampling box with a smaller edge length $\tilde{L} = eL$, with $\tau_i \in (0, \tilde{L})$, and $0 < e < 1$. Now, we quantify how much more probability mass is in the full sampling box compared to the smaller cube inside of the sampling box. For the uniform prior, one can express the ratio of the probability masses, $R_P$, by the ratio of volumes

$$R_P(e, M) = \frac{L^M}{(e \cdot L)^M} = \left(\frac{1}{e}\right)^M \tag{56}$$

Inserting some numbers in Eq. 56 results in $R_P(0.99, 10^3) \approx 2.3 \cdot 10^4$, which is an example of what is classically termed *the curse of dimensionality*. There is essentially no probability mass in the wide cube defined by $0.99L$. Almost all probability mass resides in the thin $10^{-2}$ shell at the surface of the volume assuming a uniform distribution. Obviously, $10^3$ parameters are far beyond the complexity for a CRN that can be inferred or usefully interpreted. Making the same comparison if, e.g., one increases a CRN from 4 Markov states to 6 seems not too worrisome $R_P(0.99, 4) \approx 1.04$ and $R_P(0.99, 6) \approx 1.06$.
However, $\tau_i^{-1}$ is an inverse scaling (rate) parameter (Sec. 4.2.2), whose MI prior is a log uniform distribution (Eq. 18). In 1D, a uniform prior states that some $\tau_i$ is ten times more probable inside the decade $[0.1, 1]$ than it is in $[0.01, 0.1]$ alternatively $\tau_i$ is ten times more likely in the interval $[0, 1]$ than in $[0, 0.1]$. Hence, defining the inner cube by being $N_{\text{dec}}$ decades, $e = 10^{-N_{\text{dec}}}$, smaller, equates to

$$R_P(10^{-N_{\text{dec}}}, M) = 10^{N_{\text{dec}} M}. \tag{57}$$

One gains relatively speaking $10^{N_{\text{dec}} M}$ (Eq. 56) more prior probability mass in the volume that is added by the additional decades. Dividing by the total volume of the simulations box ensures that the probability density is normalized. The bias of the uniform prior scales exponentially with $M$ the number of Markov states. This leads, in combination with the non-PI problem of the likelihood, to a posterior whose results are susceptible to the limits of the sampling box. How does the scaling for adding decades behave if one assumes $p(\boldsymbol{\tau})$ to be log uniform (Eq. 18)? Consider a simulation box with the limits $\tau_{\min}$ and $\tau_{\max} = 10^{(N_{\text{decades}}+1)} \tau_{\min}$ in each direction of $\boldsymbol{\tau}$

$$\begin{aligned}
\int_{V(\boldsymbol{\tau})} p(\boldsymbol{\tau}) d\boldsymbol{\tau} &= \prod_i \int_{\tau_{\min}}^{10^{N_{\text{dec}}+1}\tau_{\min}} p(\tau_i) d\tau = \left[\int_{\tau_{\min}}^{\tau_{\max}} p(\tau_1) d\tau\right]^M \\
&= \left[\int_{\tau_{\min}}^{10\tau_{\min}} p(\tau_1) d\tau + \int_{10\tau_{\min}}^{10^2\tau_{\min}} p(\tau_1) d\tau + \ldots \int_{10^{N_{\text{dec}}}\tau_{\min}}^{10^{N_{\text{dec}}+1}\tau_{\min}} p(\tau_1) d\boldsymbol{\tau}\right]^M \\
&= \left[N_{\text{dec}} \int_{\tau_{\min}}^{10\tau_{\min}} p(\tau_1) d\tau\right]^M
\end{aligned} \tag{58}$$

in the last step, we used the property of the log uniform prior that it contains in each decade the same amount of probability mass. Following Eq. 58 the ratio of probability masses if one uses a $N_{\mathrm{dec}}$ decades smaller cube within the simulation box equates to

$$R_P(10^{-N_{\mathrm{dec}}}, M) = N_{\mathrm{dec}}^M. \tag{59}$$

The scaling is still exponential in $M$ but not in $N_{\mathrm{dec}}$ anymore, in contrast to the scaling when the uniform prior is applied. The bias induced by uniform prior combined with the non-PI problem demands data that exponentially concentrates the probability mass if larger simulation boxes are allowed. The posterior will generally have a non-local correlation structure because the parameters of the rate matrix might compensate for each other in various ways. See in Fig. 17 **a** $p(\log(k_{32}), \log(\tau_3^{-1}) \,|\, \mathcal{Y}_T)$ The bias induced by the uniform prior, together with the non-PI problem, induces a ridge. The values along the ridge increase towards the limit of the sampling box