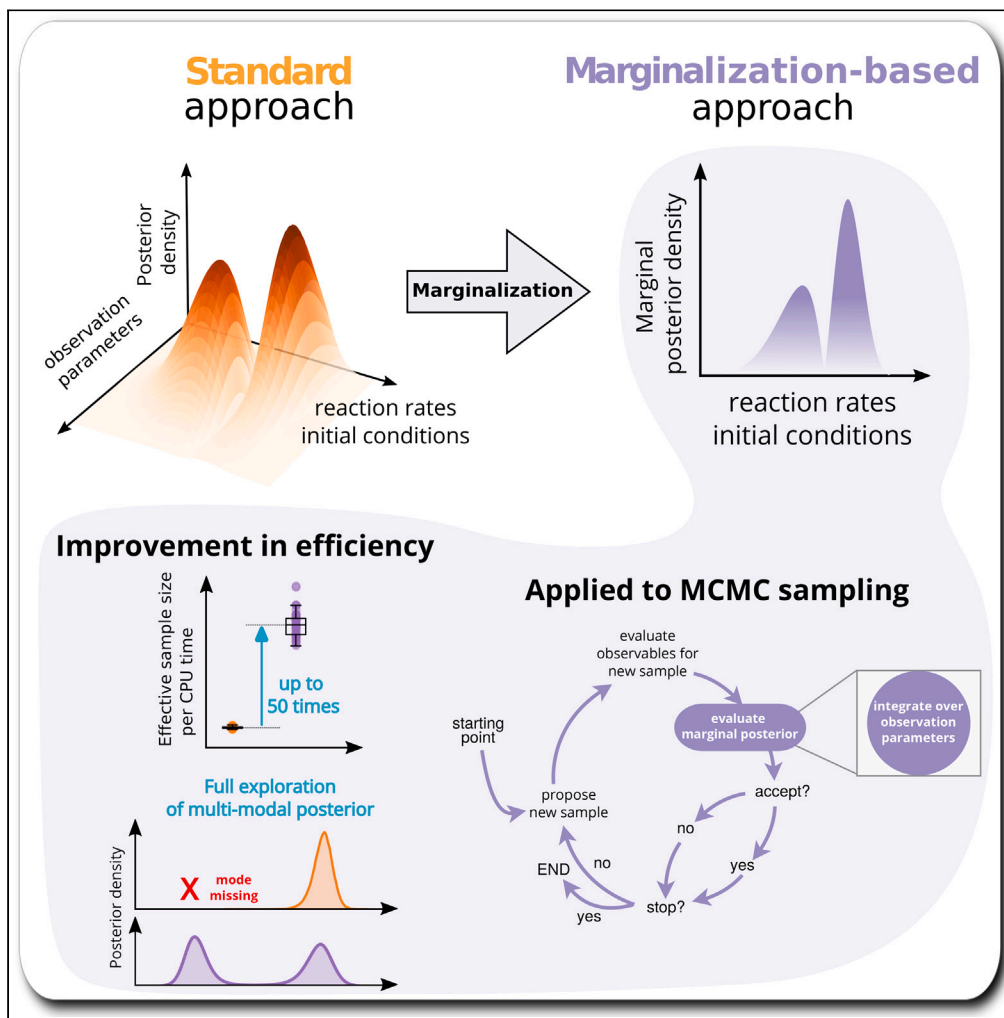


Article

# Posterior marginalization accelerates Bayesian inference for dynamical models of biological processes



Elba Raimúndez,  
Michael Fedders,  
Jan Hasenauer

jan.hasenauer@uni-bonn.de

**Highlights**  
Bayesian posterior distribution inference for mathematical models can be challenging

Our mathematical approach applies marginalization to reduce parameter dimensionality

Our method increases the effective sample size per unit of time for all tested models

Particularly beneficial for multi-modal posterior problems



## Article

## Posterior marginalization accelerates Bayesian inference for dynamical models of biological processes

Elba Raimúndez,<sup>1,2</sup> Michael Fedders,<sup>1</sup> and Jan Hasenauer<sup>1,2,3,4,\*</sup>

## SUMMARY

**Bayesian inference is an important method in the life and natural sciences for learning from data. It provides information about parameter and prediction uncertainties. Yet, generating representative samples from the posterior distribution is often computationally challenging. Here, we present an approach that lowers the computational complexity of sample generation for dynamical models with scaling, offset, and noise parameters. The proposed method is based on the marginalization of the posterior distribution. We provide analytical results for a broad class of problems with conjugate priors and show that the method is suitable for a large number of applications. Subsequently, we demonstrate the benefit of the approach for applications from the field of systems biology. We report an improvement up to 50 times in the effective sample size per unit of time. As the scheme is broadly applicable, it will facilitate Bayesian inference in different research fields.**

## INTRODUCTION

Mathematical models are important tools for understanding and predicting the dynamics of many processes, such as signaling processing in biological systems,<sup>1–3</sup> patient progression,<sup>4,5</sup> and epidemics.<sup>6,7</sup> However, the parameters of mathematical models are in general unknown and need to be inferred from experimental data. This is an inherently challenging problem and complicated by the fact that, in addition to the dynamical properties of interest (e.g., rate constants and initial conditions), characteristics of the measurement process may also be unknown. In systems biology, most measurement techniques, including western blotting,<sup>8</sup> fluorescence microscopy,<sup>9</sup> and mass spectrometry,<sup>10</sup> are not fully quantitative but provide only relative information. Moreover, there is often an unknown offset and/or noise level.<sup>11</sup> Accordingly, unknown observation parameters, such as scaling factors but also offsets and noise levels, have to be estimated along with parameters of the mathematical models.<sup>12–14</sup>

Bayesian inference is often used to estimate unknown parameters.<sup>15–17</sup> A particularly common approach is to employ Markov chain Monte Carlo (MCMC) algorithms, such as (adaptive) Metropolis-Hastings,<sup>18</sup> Hamiltonian Monte Carlo methods,<sup>19,20</sup> and parallel tempering,<sup>21</sup> to generate representative samples from the posterior distribution. Yet, with increasing number of unknown parameters, the application of MCMC algorithms becomes challenging.<sup>22</sup> This is a bottleneck that leaves sampling methods on the edge of computational feasibility. In principle, the challenge can be addressed by reducing the dimensionality of the sampling problem, e.g., by marginalizing over nuisance parameters (as, e.g., demonstrated in cosmology<sup>23</sup>). However, there is no generic and broadly applicable framework.

In frequentist inference, a template for the reduction of the dimensionality of parameter estimation problems has been provided.<sup>14,24,25</sup> Here, hierarchical optimization approaches have been developed to determine the maximum likelihood estimate. These methods exploit that the observation parameters can be computed analytically for a given set of model parameters. It has been shown that this benefits the convergence of optimization methods and the computational efficiency, while providing the same results (see, e.g., Loos et al.<sup>24</sup>). Yet, these concepts cannot be directly translated to Bayesian inference as we are not interested in only optimal point estimates, but in (marginal) posterior distributions over parameters.

In this manuscript, we introduce a generic method for improving sampling efficiency by marginalizing over observation parameters. We provide analytical results for the marginalization over complex posterior distributions for dynamical biological processes—described, e.g., by ordinary differential equations (ODEs)—with a broad class of observation models. The marginalization yields a lower dimensional posterior for MCMC sampling. Samples of the original posterior can be obtained by subsequent sampling of the observation parameters conditioned on the remaining parameters. To illustrate the properties of the proposed approach, we benchmark its performance with a collection of

<sup>1</sup>Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

<sup>2</sup>Technische Universität München, Center for Mathematics, Garching, Germany

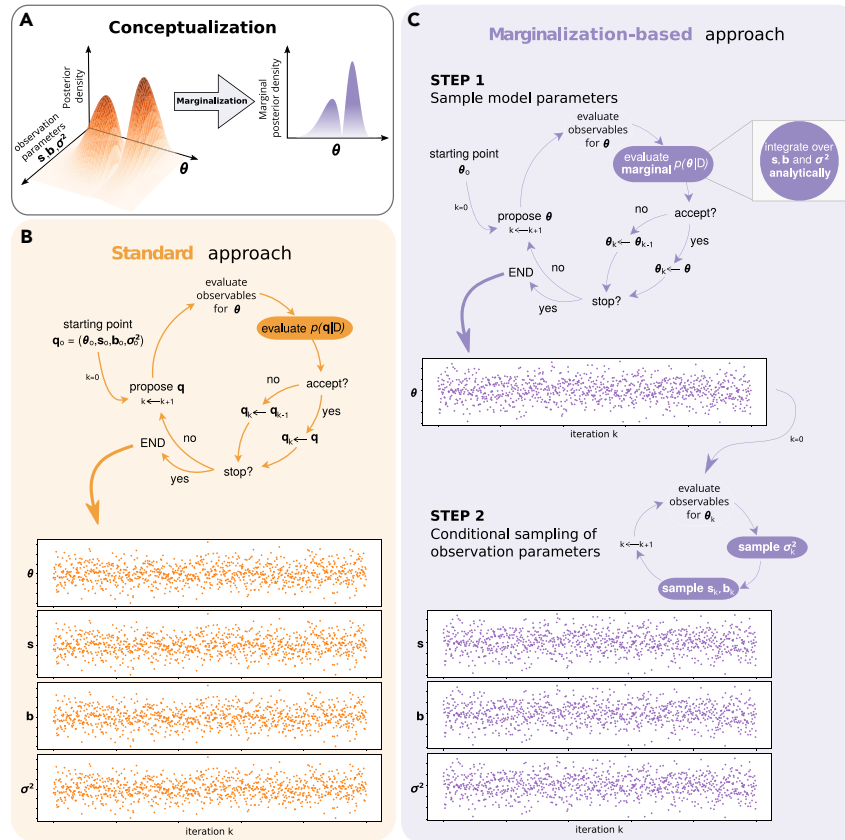
<sup>3</sup>Helmholtz Zentrum München - German Research Center for Environmental Health, Computational Health Center, Neuherberg, Germany

<sup>4</sup>Lead contact

\*Correspondence: [jan.hasenauer@uni-bonn.de](mailto:jan.hasenauer@uni-bonn.de)

<https://doi.org/10.1016/j.isci.2023.108083>





**Figure 1. Standard and marginalization-based Markov chain Monte Carlo sampling**

(A) Illustration of the general marginalization concept.

(B) Standard approach.

(C) Marginalization-based approach depicting: (Step 1) the sequential integration of the observation parameters  $s$ ,  $b$ , and  $\sigma^2$  to evaluate  $p(\theta|D)$ , and (Step 2) the (optional) conditional sampling of the marginalized observation parameters.

published models, including models for which current available sampling strategies are computationally infeasible. We demonstrate that the proposed method achieves higher sampling efficiencies by reducing the auto-correlation of the samples and increasing the transition probabilities between posterior modes. Indeed, it turns computationally infeasible sampling problems feasible, increasing the set of problems which can be tackled using Bayesian inference.

## RESULTS

### Many model structures allow for analytical marginalization of parameters and sampling in lower dimensional space

To facilitate Bayesian inference for mathematical models with observation parameters, we developed and implemented a marginalization-based sampling approach (Figure 1). The approach allows for inferring the parameters of mathematical models, such as ODEs and partial differential equation models, from data via observation models with scaling, offset, and noise parameters. For a mathematical model with parameter  $\theta$  and time- and parameter-dependent states  $x(t, \theta)$ , consider the case of a one-dimensional observable with additive Gaussian measurement noise and observation model

$$\bar{y} = (s \cdot h(x(t, \theta), \theta) + b) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (\text{Equation 1})$$

in which  $h(x, \theta)$  describes the measured quantity,  $s$  is the scaling factor ( $\in \mathbb{R}$ ),  $b$  is the offset ( $\in \mathbb{R}$ ), and  $\sigma^2$  is the variance of the measurement noise ( $\in \mathbb{R}_+$ ). A collection of measurements  $\bar{y}_k$  at time points  $t_k$ , with  $k = 1, \dots, n_t$ , is denoted as data  $D = \{(\bar{y}_k, t_k)\}_{k=1}^{n_t}$ . Following Bayes' theorem, the posterior distribution of the parameters  $(\theta, s, b, \sigma^2)$  given the data  $D$  is

$$p(\theta, s, b, \sigma^2 | D) = \frac{p(D | \theta, s, b, \sigma^2) p(\theta, s, b, \sigma^2)}{p(D)}, \quad (\text{Equation 2})$$

in which  $p(D | \theta, s, b, \sigma^2)$  denotes the likelihood,  $p(\theta, s, b, \sigma^2)$  denotes the prior distribution, and  $p(D)$  denotes the marginal probability.

The **standard approach** is to use MCMC methods to obtain representative samples from the joint posterior distribution for model parameters  $\theta$  and observation parameters  $s, b$ , and  $\sigma^2$  (2) for subsequent analysis (Figure 1B). All parameters are sampled jointly, disregarding their nature (Figure 1B); in particular note that the state  $x(t, \theta)$  and the value of the observation map  $h(x(t, \theta), \theta)$  only depend on  $\theta$  but not on  $s, b$ , or  $\sigma^2$ . This approach is often challenging and even infeasible for models with large datasets since the number of observation parameters can easily exceed the number of model parameters (see, e.g., Bachmann et al. and Raimúndez et al.<sup>26,27</sup>).

To simplify the sampling process, we propose a **marginalization-based approach**, which exploits a decomposition of the sampling problem in two steps (Figure 1C). In Step 1, we consider the marginalization of the posterior distribution (2) with respect to the observation parameters  $s, b$ , and  $\sigma^2$ , yielding

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

with  $p(\mathcal{D}|\theta)$  as the marginal likelihood given by

$$p(\mathcal{D}|\theta) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(\mathcal{D}|\theta, s, b, \sigma^2) p(s, b, \sigma^2|\theta) ds db d\sigma^2, \quad (\text{Equation 3})$$

assuming that the prior can be written as  $p(\theta, s, b, \sigma^2) = p(s, b, \sigma^2|\theta)p(\theta)$ . For various choices of noise models and prior distributions (in particular conjugate priors), this marginal likelihood can be computed in closed form. This is for instance the case for the combination of additive Gaussian noise with a joint prior distribution for  $s, b$ , and  $\sigma^2$ ,

$$p(s, b, \sigma^2|\theta) = \mathcal{N}(s|\nu, \sigma^2/\tau) \cdot \mathcal{N}(b|\mu, \sigma^2/\kappa) \cdot \Gamma^{-1}(\sigma^2|\alpha, \beta),$$

in which  $\nu, \mu \in \mathbb{R}$  and  $\tau, \kappa, \alpha, \beta \in \mathbb{R}_+$  denote hyperparameters of the Normal-Inverse-Gamma-distributed joint prior, and  $\Gamma^{-1}(\cdot)$  denotes the Inverse-Gamma function. The hyperparameters might depend on  $\theta$ . Here, we obtain for observations  $\bar{y}_i$  with  $i = 1, \dots, n_t$  the closed-form expression for the marginal likelihood as

$$p(\mathcal{D}|\theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa\tau}{(n_t + \kappa)\left(\tau + \sum_{i=1}^{n_t} h_i^2\right) - \left(\sum_{i=1}^{n_t} h_i \bar{y}_i\right)^2}} \quad (\text{Equation 4})$$

with  $h_i := h(x(t_i, \theta), \theta)$  and parameter-dependent constant

$$C := \beta + \frac{1}{2} \left( \kappa\mu^2 + \tau\nu^2 + \sum_{i=1}^{n_t} \bar{y}_i^2 - \frac{(\kappa\mu + \sum_{i=1}^{n_t} \bar{y}_i)^2}{n_t + \kappa} - \frac{((\kappa\mu + \sum_{i=1}^{n_t} \bar{y}_i)(\sum_{i=1}^{n_t} h_i) - (n_t + \kappa)(\tau\nu + \sum_{i=1}^{n_t} h_i \bar{y}_i))^2}{(n_t + \kappa)\left((n_t + \kappa)\left(\tau + \sum_{i=1}^{n_t} h_i^2\right) - \left(\sum_{i=1}^{n_t} h_i \bar{y}_i\right)^2\right)} \right).$$

As the Normal-Inverse-Gamma prior is a conjugate prior for additive Gaussian noise, the marginal likelihood is analytically tractable. There are various other cases, including multiplicative Gaussian noise and even distributions with outliers. For the latter, Laplacian noise has shown to be more robust against measurement outliers.<sup>28</sup> Tables S1 and S2 summarize ten practically relevant cases for which we obtained closed-form expressions, and we are certain that many more are possible. For details on the derivation of all individual results (including two cases for Laplace distributed noise), we refer to the [supplemental data](#).

Given the marginalized likelihood function  $p(\mathcal{D}|\theta)$  and the prior  $p(\theta)$ , the posterior distribution  $p(\theta|\mathcal{D})$  of the parameters of the mathematical model can be sampled using MCMC and related methods. The sampling can be performed in the space of  $\theta$ , as the observation parameters are implicitly considered (Figure 1C).

The samples of model parameters  $\theta$  from  $p(\theta|\mathcal{D})$  allow for the assessment of the model properties and its uncertainties. In this regard, there is no difference of sampling the marginalized posterior distribution  $p(\theta|\mathcal{D})$  compared to projecting the full posterior distribution  $p(\theta, s, b, \sigma^2|\mathcal{D})$  onto the  $\theta$  component. However, tasks like the assessment and plotting of the model-data mismatch also require the posterior of the observation parameters. These can be obtained by sampling from the conditional distribution  $p(s, b, \sigma^2|\theta, \mathcal{D})$ . As the observation parameters only influence the observation model (1) and not the calculation of state  $x(t, \theta)$  and observable map  $h(x, \theta)$ , the conditional distribution can be expressed in closed form and sampled efficiently. For the aforementioned case, a matching sample of observation parameters for a given model parameter  $\theta$  can be obtained by drawing from Gamma and Normal distributions:






$$\sigma^2 = 1/\lambda \quad \text{with} \quad \lambda \propto \Gamma\left(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C\right),$$

$$b \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu + \left(\sum_{i=1}^{n_t} \bar{y}_i - h_i\right)}{\kappa + n_t}, \lambda' = \lambda(n_t + \kappa)\right), \text{ and}$$

$$s \propto \mathcal{N}\left(\mu' = \frac{(\kappa + n_t)(\tau\nu + \sum_{i=1}^{n_t} h_i \bar{y}_i) - (\kappa\mu + \sum_{i=1}^{n_t} \bar{y}_i)(\sum_{i=1}^{n_t} h_i)}{(\kappa + n_t)\left(\tau + \sum_{i=1}^{n_t} h_i^2\right) - \left(\sum_{i=1}^{n_t} h_i \bar{y}_i\right)^2}, \lambda' = \lambda\left(\tau + \sum_{i=1}^{n_t} h_i^2 - \frac{\left(\sum_{i=1}^{n_t} h_i \bar{y}_i\right)^2}{(n_t + \kappa)}\right)\right),$$



**Table 1. Key numbers and features of the considered toy and benchmark models**

Model ID	$n_\theta$	$n_s$	$n_b$	$n_\sigma$	Description	Reference
Toy 	2	1	1	1	Conversion reaction	–
M1 	13	3	–	–	EGF-AKT pathway	Fujita et al. <sup>37</sup>
M2 	6	3	–	3	STAT5 dimerization	Boehm et al. <sup>38</sup>
M3 	3	1	–	1	mRNA transfection	Leonhardt et al. <sup>39</sup>
M4 	26	31	–	–	Gastric cancer signaling	Villaverde et al. <sup>40</sup>

The number of unknown model parameters  $n_\theta$ , unknown scaling parameters  $n_s$ , unknown offset parameters  $n_b$ , and unknown noise parameters  $n_\sigma$ , which are effectively sampled, are reported.

with  $h_i$  and  $C$  being evaluated for model parameter  $\theta$ . This conditional sampling can be proven to provide the same correlation structure as directly sampling the full posterior distribution. For details on the derivation of the conditional sampling for the observation parameters we refer to the [supplemental data](#). As the conditional sampling can be performed independently and does not require model simulation, it is computationally efficient. For additional observation models see [Tables S1](#) and [S2](#).

In summary, a broad spectrum of parameter estimation problems can be reformulated by performing an analytically tractable marginalization of their observation parameters. Sampling of this lower dimensional posterior distribution for the model parameters  $\theta$  in combination with conditional sampling for the observation parameters allows the construction of samples from the full posterior distribution. Accordingly, the original sampling problem is decomposed in two sub-problems, of which the conditional sampling is optional.

### Marginalization-based approach yields same results at lower computational cost

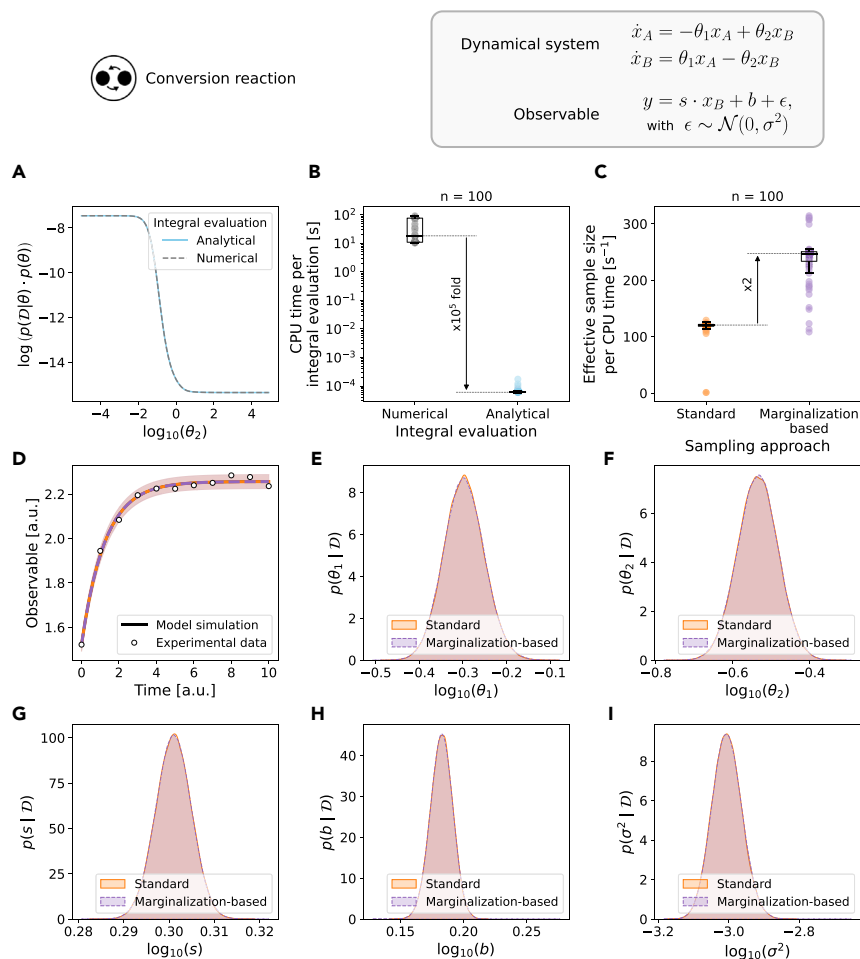
To compare the performance for the standard and marginalization-based approach, we performed a range of studies using (i) a simple test problem and (ii) published models and datasets.

As a simple test problem we considered a model of a conversion reaction process,  $A \rightleftharpoons B$ . This process was considered in various other publications<sup>28,29</sup> and can be described using a two-dimensional system of ODEs, with the concentrations of  $A$  and  $B$  as state variables. Here, we considered that the abundance of  $B$  is measured up to an unknown scaling, offset, and noise level. Accordingly, the mathematical model possesses two model parameters, the forward rate  $A$  to  $B$ ,  $\theta_1$ , and the backward rate  $B$  to  $A$ ,  $\theta_2$ , and three observation parameters, the scaling  $s$ , the offset  $b$ , and the noise variance  $\sigma^2$  ([Table 1](#)). A detailed description of the model is provided in the [STAR methods](#) section.

In the first step, we used the model to assess the correctness of the analytical marginalized likelihood (4) by comparing its agreement with numerical integration of [Equation 3](#). The results show a perfect match for a range of different parameter values ([Figure 2A](#)). Yet, the evaluation of the analytical marginalized likelihood was five orders of magnitude faster than the numerical integration ([Figure 2B](#)), which highlights the importance of the analytical derivations. In the second step, we performed 100 independent MCMC sampling runs for the standard and marginalization-based approach. The runs employed a state-of-the-art adaptive Metropolis-Hastings method.<sup>18</sup> We found a superior performance of the marginalization-based approach, as the observed effective sample size per unit of time was twice as high as for the standard approach ([Figure 2C](#)). This indicates that the marginalization-based approach facilitates already for simple problems the mixing of the MCMC chains and, hence, provides a more efficient exploration of the posterior. Moreover, the model fit for the best sample found (i.e., maximizing the posterior) coincided for both approaches ([Figure 2D](#)) as well as the marginal distributions for the model parameters  $\theta_1$  and  $\theta_2$  ([Figures 2E](#) and [2F](#)), and the conditionally sampled observation parameters ([Figures 2G–2I](#)).

Following the promising results for the test problem, we evaluated the performance of the proposed marginalization-based approach for three already published models and datasets ([Table 1](#) and [STAR methods](#) section). The models M1 to M3 describe cellular processes: (M1) epidermal growth factor (EGF)-induced protein kinase B (AKT) signaling; (M2) phosphorylation-dependent STAT5 dimerization; and (M3) mRNA transfection. The numbers of model and observation parameters differ, and so do the observation functions. Accordingly, different closed-form expressions for the marginalized likelihood function are used ([Tables S1](#) and [S2](#)). More importantly, the full posterior distributions exhibit different characteristics, ranging for instance from uni- to bimodal.

For the considered application problems, the marginalization of the observation parameters reduced the dimensionality of the sampling problems by up to 50% (ranging from 19% to 50%) ([Figure 3A](#)). The validity of the analytical expressions for marginalized likelihoods was again confirmed using numerical integration ([Figure S1](#)). To evaluate the impact of this reduction on the sampling efficiency, we performed 50 independent MCMC sampling runs using the parallel tempering algorithm with 10 temperatures.<sup>21</sup> All the runs were initialized at parameter values maximizing the posterior probability which were found using multi-start optimization.<sup>12</sup> For M1 and M2, these maximum a posteriori (MAP) estimates were unique, while for M3, two MAP estimates were found with identical posterior values. The sampling was run for  $10^6$  iterations. Further details are provided in the [STAR methods](#) section. The high number of iterations allowed all MCMC runs of the standard and marginalized problem to converge according to the Geweke test.<sup>30</sup> Yet, the marginalization-based approach achieved a higher effective sample size per unit of computation time than the standard approach ([Figure 3B](#)). The improvement was problem dependent and ranged from



**Figure 2. Evaluation of the standard and marginalization-based approach for the toy model**

(A) Comparison of analytical vs. numerical integration.

(B) Time comparison of analytical vs. numerical integration.

(C) Effective sample size per unit of time for 100 independent runs.

(D) Model fit of the best sample found during sampling from the standard (orange) and marginalization-based (purple) approach.

(E–I) Parameter marginal posterior distributions computed using a kernel density estimate for the model parameters (E)  $\theta_1$  and (F)  $\theta_2$ , and the conditionally sampled observation parameters: (G) scaling factor  $s$ , (H) offset  $b$ , and (I) noise variance  $\sigma^2$ .

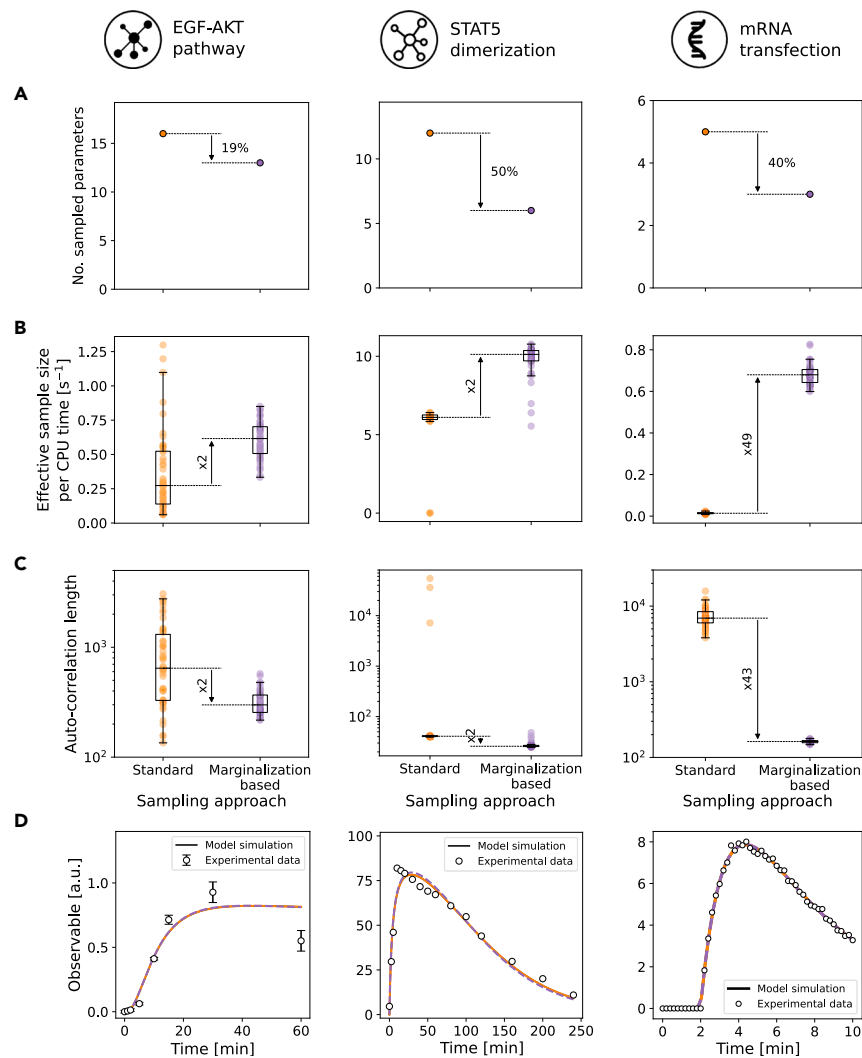
2 (M1 and M2) to nearly 50 (M3) times higher efficiency in the marginalization-based approach. As the computation time was similar, the core reason for this is a reduction in the auto-correlation length (Figure 3C). The model fits for the best sample found were identical for both approaches (Figures 3D, S2 and S3) as well as the parameter marginal distributions (Figures S4–S6).

In summary, test and application problems demonstrate the acceleration potential of the marginalization-based approach. The improvement was problem specific, with no clear dependence on the degree of dimensionality reduction, but in all cases substantial.

### Marginalization-based approach improves transition rates between posterior modes

To understand for which problems the marginalization-based approach is expected to achieve a large acceleration, we considered the model M3. The posterior distribution for M3 is bimodal, and a simple explanation for the acceleration would have been that the bimodality is eliminated. Yet, this is not the case as the bimodality is related to a symmetry in model parameters. Numerical simulations as well as analytical results reveal that the observable trajectory remains unchanged when the mRNA and protein degradation rates are interchanged. As long as the optimal point is not located on the line of equal degradation rates, standard and marginalized posterior are bimodal.

We hypothesized that the large efficiency improvement is related to a lower minimum energy path for the transitions in the marginalized posterior. To assess this, we computed the minimum energy paths<sup>31</sup> for the standard (Figures 4A and 4B) and marginalized posterior (Figures 4C and 4D) (see details in the STAR methods section). To our surprise, the minimum energy path is almost identical for both approaches (Figure 4E). Hence, there is at least no difference in the minimum energy path.



**Figure 3. Evaluation of the standard and marginalization-based approach for the benchmark models**

Models M1–M3 are shown from left to right.

(A) Number of sampled parameters.

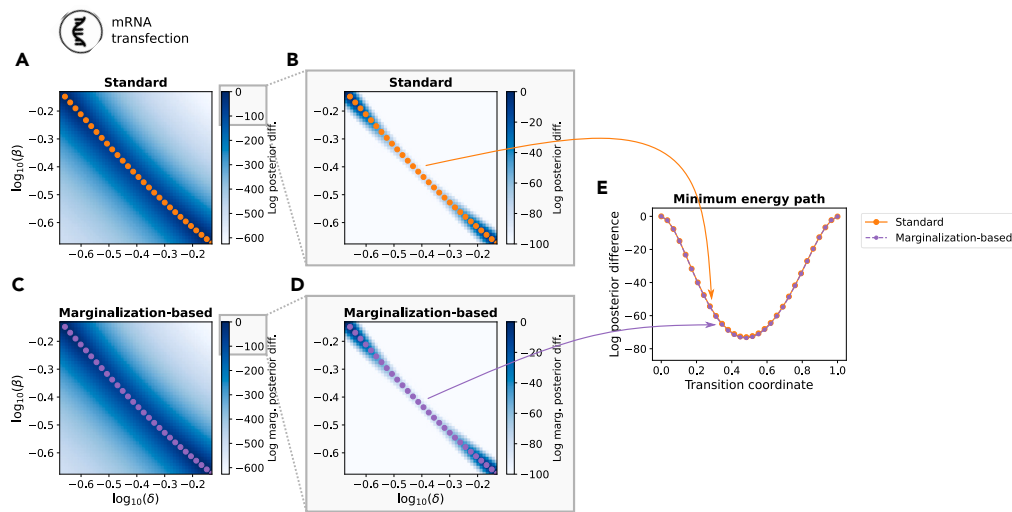
(B) Effective sample size per unit of time.

(C) Auto-correlation length.

(D) Model fit of the best sample found during sampling. A subset of the experimental data is shown for M1 and M2. Complete datasets and parameter marginal distributions are depicted in [Figures S2–S6](#). A comparison of analytical vs. numerical integration is shown in [Figure S1](#).

In order to understand the improvement observed for runs of adaptive parallel tempering methods, we performed 10 runs of a single-chain adaptive Metropolis algorithm<sup>18</sup> with  $10^6$  iterations for exploring the posterior ( $T = 1$ ). We expected this to simplify the interpretation. Yet, the adaptive Metropolis algorithm was essentially unable to transition between the two modes of the posterior, meaning that efficiency improvements could not be assessed with reasonable computation time (see  $T = 1$  in [Figure 5A](#)). To assess the relative complexity of the sampling problem for standard and marginalization-based approach, we repeated the evaluation with the single-chain adaptive Metropolis algorithm for the tempered posterior, keeping the temperature fixed for a specific run. We found that the marginalization-based approach allows already at lower temperatures for transitions between the modes unlike the standard sampling approach ([Figures 5A, S8, and S9](#)). For temperatures such as  $T = 16$ , the standard approach showed an average number of only 5 transitions between the modes with many runs only sampling from a single mode ([Figures 5B and 5C](#)), while for the marginalization-based approach on average  $1.6 \times 10^4$  transitions occurred ([Figures 5D and 5E](#)). As the minimum barrier energy is conserved also for higher temperatures ([Figure S7](#)), this increase in the transition rate by four orders of magnitude for the same algorithm implies a lower overall complexity of the marginalization-based sampling problem.

As the increased transition rate is not caused by an altered energy path, we studied the transition paths. This revealed that the employed single-chain algorithm facilitates jumps over the valley in the objective function ([Figures 5F and 5G](#)), meaning that it transitions between



**Figure 4. Comparison of the minimum energy path for model M3**

(A–D) Landscape of the optimized (A and B) posterior and (C and D) marginalized posterior for different fixed values of the model parameters  $\beta$  and  $\delta$ . The difference with respect to the maximal posterior value is depicted.

(E) Transition coordinates for the minimum energy path.

high-probability regions around the local optima. These direct transitions appear at a high rate for the marginalization-based approach (Figure 5G), while they rarely happen for the standard approach (Figure 5F). For the latter, most transitions are along low-energy paths with posterior probabilities dropping below the minimum energy path. Accordingly, the transition behavior is for the marginalization-based approach more efficient than for the standard approach.

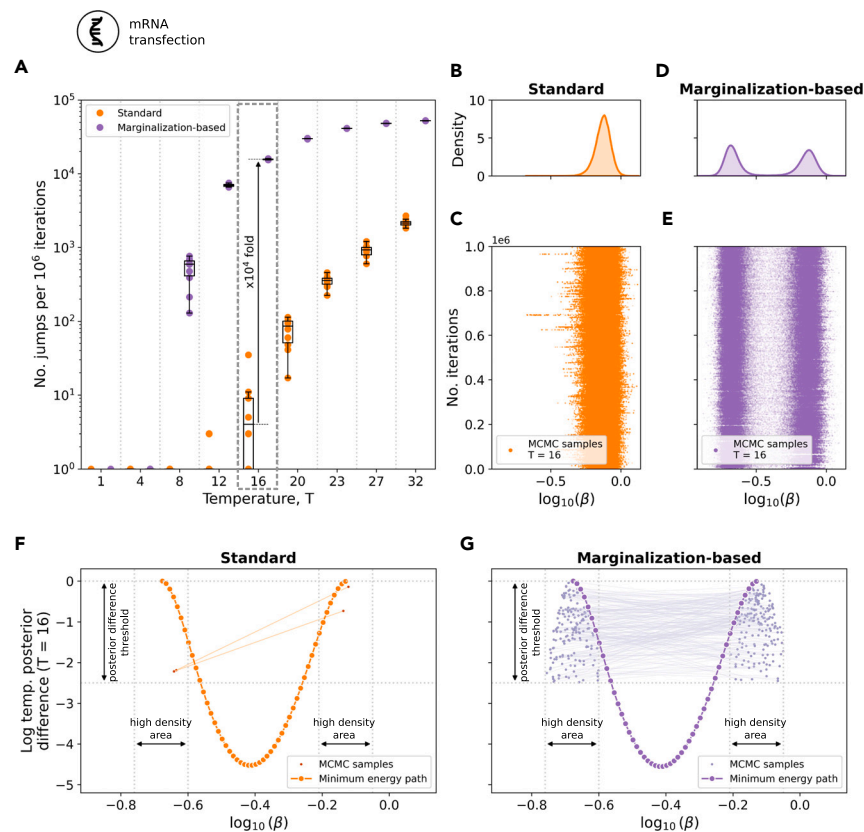
In summary, the in-depth study of the mRNA transfection model (M3) showed that the marginalization-based approach can achieve substantial accelerations as the structure of the sampling problem is simplified, e.g., by facilitating transitions between modes. The improvements are related to the interplay of sampling approach and problem geometry. In particular for challenging (e.g., multi-modal) problems a much greater improvement could be observed.

### Marginalization-based approach enables Bayesian inference for large models

As the marginalization-based approach appeared beneficial for challenging problems, we assessed in a next step whether it enables Bayesian inference for problems for which standard approaches did not provide reproducible results in a reasonable time frame. Specifically, we considered an ODE model for signal transduction in gastric cancer cells (cell line MKN1) that was developed to unravel response and resistance markers.<sup>27</sup> This model possesses in total 57 unknown parameters, of which 26 are model parameters and 31 are observation parameters (Table 1, M4).

The application of the marginalization-based approach resulted in a reduction of the dimensionality of the sampling problem by over 50% (Figure 6A). For the 26 model parameters which remain to be sampled, we compared the marginal likelihoods as computed using the previously derived analytical formulas and numerical integration (Figure 6B). The agreement of the results (Pearson correlation  $r \approx 1.0$ ) confirmed the correctness of our analytical integration.

To determine the parameters of the model, we performed sampling using standard and marginalization-based approach. The adaptive Metropolis-Hastings algorithm<sup>18</sup> and the adaptive parallel tempering algorithm<sup>21</sup> employed in the previous sections were run 10 times with different starting points and random seeds for  $10^6$  iterations for the adaptive Metropolis-Hastings and  $10^5$  iterations for the adaptive parallel tempering algorithm. We found that while all runs in the marginalization-based approach (and for both sampling algorithms) successfully finished within a run time limit of 7 days, only 7 out of 10 runs successfully finished for the standard approach for each sampling algorithm. The MAP estimates observed in the different runs provided similar fits (Figures 6C and 6D). In contrast, the marginal distributions of the model parameters differed, with the marginalization-based approach mostly providing broader parameter distributions than the standard approach (Figures 6E and 6F). The assessment of the reproducibility of the marginal distributions revealed a high variability between different runs performed using the standard approach (Figures 6E and S10). On the contrary, for the marginalization-based approach a good agreement between runs was observed (Figures 6F and S11), indicating reproducibility. To verify that the behavior observed for the individual parameters is maintained in the full parameter space, we analyzed the overall agreement of all parameter samples across all runs for the standard and marginalization-based approach by visualizing the samples using the uniform manifold approximation and projection (UMAP) representation.<sup>32</sup> We found that the individual runs of the standard approach represent individual clusters in the UMAP (Figure 6G), while the individual runs of the marginalization-based approach were indistinguishable (Figure 6H). This finding was supported by the distribution of the nearest neighbors (Figure S12). This revealed that: (i) in the marginalization-based approach all the individual runs sample from the same distribution and (ii) the standard approach failed for both algorithms considered here.



**Figure 5. Quantification of the transitions between the posterior modes for different temperatures  $T$  for model M3**

(A) Number of transitions per  $10^6$  iterations for a range of temperatures for the standard (orange) and marginalization-based (purple) approach. A total of 10 chains per temperature value are depicted.

(B and D) Marginal distribution computed using a kernel density estimate and (C, E) parameter trace for the model parameter  $\beta$  of a representative chain obtained with the (B, C) standard and (D and E) marginalization-based approach for  $T = 16$ .

(F and G) Direct transitions between the posterior modes of a representative chain along with the minimum energy path obtained with the (F) standard and (G) marginalization-based approach for  $T = 16$ . See also [Figures S7–S9](#).

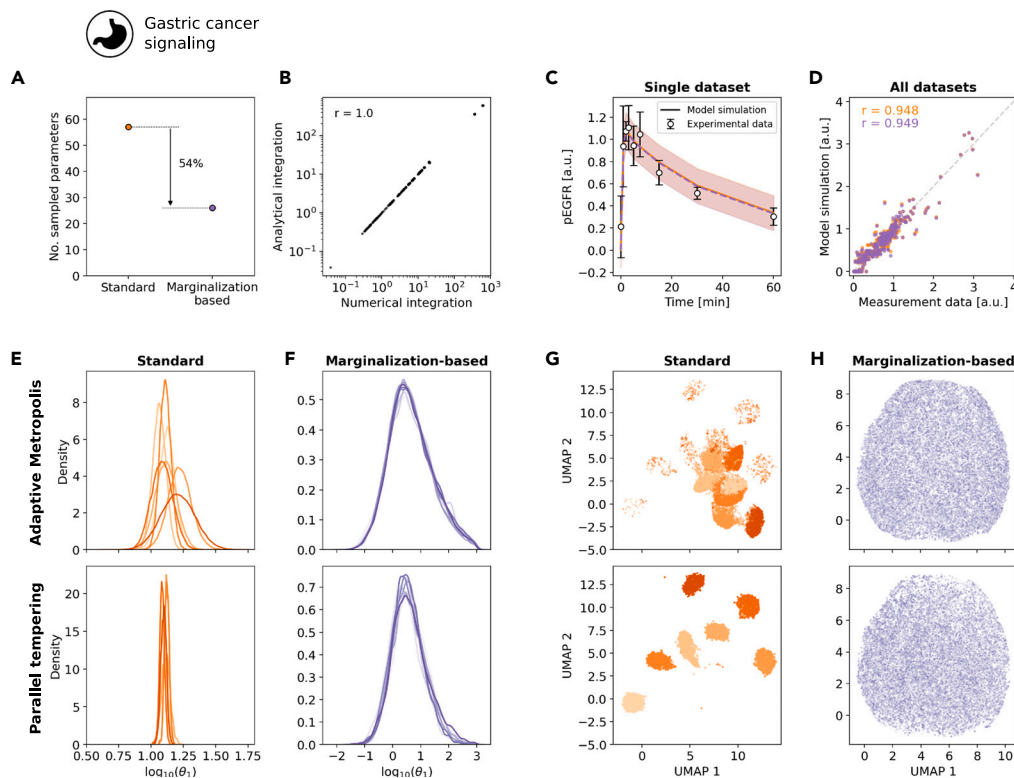
The study of the model of signal processing in gastric cancer cells revealed that marginalization-based approach allows for reproducible sampling in problems, where the standard approach failed. While for the marginalization-based approach all runs provided consistent results, the standard approach failed to converge within an average central processing unit (CPU) time of 150 h rendering its application impracticable. Furthermore, our study provides improved estimates for the parameters ([Figure S13](#)) of important processes of a drug used in clinical practice.

In summary, the application of our marginalization-based approach to Bayesian inference for models with relative measurement data shows consistently that our approach yields the same marginal distributions for the parameters as the standard approach, while being highly more efficient in exploring the parameter space and enabling Bayesian inference of larger models, which was not possible before with the standard approach.

## DISCUSSION

Bayesian inference for models of biological processes requires the consideration of parameters of the dynamical systems as well as the measurement process. The unknown scaling factors, offsets, and noise levels often resemble a large fraction of the overall parameters.<sup>12</sup> This complicates sampling and can render the generation of representative samples practically infeasible. Here, we address this challenge by proving that an (analytical) marginalization of the posterior even for dynamical models, for common observation and noise models, and plausible priors. We provide analytical result for additive normal, multiplicative log-normal, and additive Laplace noise for different choices of unknown parameters and priors. This approach allows for the construction of a sample from the full posterior by (i) sampling a marginalized posterior for the parameters of the dynamical systems and (ii) conditional sampling of the observation parameters.

We evaluated the performance of our marginalization-based approach and compared it to the standard approach for four published models, with differences in their complexity. This revealed an increased effective sample size per unit of time, and increased transition



**Figure 6. Convergence of the marginalization-based approach for model M4**

(A) Number of sampled parameters.

(B) Scatterplot for the agreement of analytical and numerical integration.

(C and D) Model fit of the best sample found during sampling for, (C) a subset of the experimental data represented as mean  $\pm$  standard deviation and (D) the complete dataset in form of a scatterplot, the standard (orange) and marginalization-based approach (purple). (E–H) Results from adaptive Metropolis (top) and parallel tempering (bottom) are shown.

(E and F) Parameter marginal posterior distribution obtained using the (E) standard and (F) marginalization-based approach computed using a kernel density estimate for model parameter  $\theta_1$ .

(G and H) Dimensionality reduction for all samples from all runs for the (G) standard and (H) marginalization-based approach using the UMAP representation. Different shades correspond to individual runs. The UMAPs were constructed using the Python package `umap`.<sup>32</sup> See also [Figures S10–S12](#).

probabilities between posterior modes. The marginalization-based approach was for all considered problems more efficient than the standard approach, but—more importantly—it also enabled the assessment of the posterior distribution for larger models for which the standard approach failed to converge in the considered time frame. Interestingly, there was no strong relation between the reduction of the problem dimensionality and the improvement in efficiency. The improvement seems to rather depend in the characteristics of the marginalized posterior and the interplay of these characteristics with the employed sampling algorithm. This is consistent with previous finding for hierarchical optimization,<sup>25</sup> where a minimal reduction of the problem dimensionality was shown to substantially improve the conditioning of the optimization problem. Based on our observations we expect the sampling behavior to benefit substantially even from the removal of a small number of parameters, as (i) the likelihood value is often very sensitive to them, which produces narrow rims in the posterior distribution, and as (ii) the removal of a small number of parameters can result in a substantially increased probability to jump between modes. The latter was observed for the model of mRNA transfection. A review of the PETA benchmark collection<sup>33</sup> showed that 20 out of 30 dynamical models used in systems biology and medicine possess unknown observation parameters. Hence, a large number of modeling projects could profit from the approach.

The approach presented here is not limited to relative measurement data, but also applicable to absolute measurements. As for these, the noise parameters would still have to be inferred ([Tables S1 and S2](#)). We provide the detailed derivation in the [supplemental data](#). Accordingly, our approach can be used for combinations of relative and absolute data. Also, it is applicable to different measurement process functions and noise models to the ones considered here. We hypothesize that also an extension to correlated noise is possible, but this remains to be assessed.

The choice of conjugate priors for the marginalized parameters eased the analytical derivation of the marginal posterior. This implies in our case that observable and noise parameters are not independent under the prior. Mostly, this is not a problem since both parameters are related to the measurement process. However, in some cases, there might be known parameters to be independent; therefore, other prior distribution assumptions must be considered. It should be noted that the concept of marginalization is not restricted to integrals that are

analytically solvable, but also numerical integration schemes can be considered. However, this would increase the required computation time (as observed in Figure 2B), but very likely the improved mixing properties would be maintained. Whether the improved mixing out-weights the increased computational cost will be problem dependent but might not be unlikely (and would have been the case for the mRNA transfection model (M3)) as the numerical integration over observation parameters will not require numerical simulations of the model. In future research projects this question should be tackled via a comprehensive benchmarking. Similarly, while in this manuscript only cases were presented in which the conditional sampling of observation parameters was straightforward due to the use of conjugate priors, the approach is also applicable if this does not hold. In this case the sampling of the parameters  $\theta$  is not impaired, but MCMC sampling or rejection sampling might need to be used to obtain sample for the observation parameters.

The proposed method was beneficial in combination with adaptive Metropolis-Hastings and adaptive parallel tempering algorithms. We expect that the same will hold true for sampling algorithms exploiting gradient information, such as Hamilton Monte Carlo sampling.<sup>19,20</sup> As the marginal likelihood is differentiable, merely the derivation and implementation of the gradient are required. The usage of methods which exploit the Riemann geometry of the parameter space of statistical models, e.g., Metropolis-adjusted Langevin algorithm,<sup>34</sup> might be slightly more involved. This requires the derivation of the marginalized Fisher information matrix. While we assume that this can be derived in closed form or at least be accurately approximated, the corresponding results are not yet available. Alternatively, automatic differentiation could be employed to obtain gradients.<sup>35</sup> The assessment of the impact of posterior marginalization on the performance of these samplers as well as other sampling methods would be highly beneficial but is beyond the scope of this work.

In this study, we focused on the assessment of parameter uncertainties for ODE models. Yet, as the marginalization-based approach provides a complete parameter sample, it facilitates also the evaluation of prediction uncertainties.<sup>16</sup> Accordingly, we expect that it might contribute to resolving reliability problems of Bayes prediction uncertainty analysis encountered in recent studies.<sup>36</sup> Furthermore, the proposed approach is not limited to ODEs, but directly applicable for other deterministic models, e.g., partial differential equations.

In summary, the marginalization-based approach provides a new tool for Bayesian inference for models with observation-related parameters. It substantially benefits the efficiency of sampling-based approaches and renders the generation of representative posterior samples for large models possible. As it is agnostic to the structure of the underlying dynamical model, it is widely applicable to mathematical models from different research fields, such as engineering, physics, and ecology.

### Limitations of the study

This study has three main limitations. The first limitation is the number of models that were considered in the study. The extrapolation of these results when testing on 4 published models may be not applicable to all models, and behavioral exceptions may occur. However, when selecting our candidate models we tried to cover different degrees of complexities and structures. Similarly, this applies to the sampling algorithms used. Secondly, our approach may be in principle applicable to other model types, such as partial differential equations. While we expect to get similar results, this remains to be evaluated. Lastly, the specification of such a “constrained” observation model is another limitation. Ideally, the approach could be combined with automatic-differentiation schemes to flexibly facilitate the use of multiple observation models.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**
  - Mechanistic modeling of biological systems
  - Benchmark models
  - Parameter optimization
  - Bayesian parameter inference

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2023.108083>.

### ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Education and Research (Grant no. 031L0159C; J.H.), the University of Bonn (via the Schlegel Professorship; J.H.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy EXC 2047/1 - 390685813 (E.R., M.F., J.H.); EXC 2151 - 390873048 (E.R., J.H.); TRR 333/1 - 450149205 (E.R., J.H.); SFB 1454 - 432325352 (M.F.); and 443187771 (J.H.).



## AUTHOR CONTRIBUTIONS

Conceptualization: J.H., E.R.; Methodology: J.H., E.R., M.F.; Software: E.R.; Formal analysis: E.R.; Investigation: E.R., M.F.; Data curation: E.R.; Writing – original draft: J.H., E.R.; Writing – review and editing: all authors; Visualization: E.R.; Supervision: J.H., E.R.; Funding acquisition: J.H.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 5, 2023

Revised: July 16, 2023

Accepted: September 25, 2023

Published: September 28, 2023

## REFERENCES

1. Kitano, H. (2002). Systems biology: A brief overview. *Science* 295, 1662–1664.
2. Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2005). *Systems Biology in Practice* (Wiley-VCH).
3. Schöberl, B., Pace, E.A., Fitzgerald, J.B., Harms, B.D., Xu, L., Nie, L., Linggi, B., Kalra, A., Paragas, V., Bukhalid, R., et al. (2009). Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Sci. Signal.* 2, ra31.
4. Fey, D., Halasz, M., Dreidax, D., Kennedy, S.P., Hastings, J.F., Rauch, N., Munoz, A.G., Pilkington, R., Fischer, M., Westermann, F., et al. (2015). Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.* 8, ra130.
5. Hass, H., Masson, K., Wohlgemuth, S., Paragas, V., Allen, J.E., Sevecka, M., Pace, E., Timmer, J., Stelling, J., MacBeath, G., et al. (2017). Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ Syst. Biol. Appl.* 3, 27.
6. Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., and Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* 26, 855–860.
7. Zhao, S., and Chen, H. (2020). Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant. Biol.* 8, 11–19.
8. Renart, J., Reiser, J., and Stark, G.R. (1979). Transfer of proteins from gels to diazobenzoyloxymethyl-paper and detection with antisera: A method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. USA* 76, 3116–3120.
9. Sanderson, M.J., Smith, I., Parker, I., and Bootman, M.D. (2014). Fluorescence Microscopy. *Cold Spring Harb. Protoc.* 2014, pdb.top071795.
10. Blasi, T., Feller, C., Feigelman, J., Hasenauer, J., Imhof, A., Theis, F.J., Becker, P.B., and Marr, C. (2016). Combinatorial histone acetylation patterns are generated by motif-specific reactions. *Cell Syst.* 2, 49–58.
11. Kreutz, C., Bartolome Rodriguez, M.M., Maiwald, T., Seidl, M., Blum, H.E., Mohr, L., and Timmer, J. (2007). An error model for protein quantification. *Bioinformation* 23, 2747–2753.
12. Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelker, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B.D., Theis, F.J., et al. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One* 8, e74335.
13. Degasperis, A., Fey, D., and Kholodenko, B.N. (2017). Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *NPJ Syst. Biol. Appl.* 3, 20.
14. Weber, P., Hasenauer, J., Allgöwer, F., and Radde, N. (2011). Parameter estimation and identifiability of biological networks using relative data. In *Proc. of the 18th IFAC World Congress, 44Proc. of the 18th IFAC World Congress*, pp. 11648–11653.
15. Xu, T.-R., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, G.S., Ketley, D., Dunlop, A.J., Milligan, G., Houslay, M.D., and Kolch, W. (2010). Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.* 3, ra20.
16. Raue, A., Kreutz, C., Theis, F.J., and Timmer, J. (2013). Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philos. T. Roy. Soc. A* 371, 20110544.
17. Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F.J. (2013). High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Math. Biosci.* 246, 293–304.
18. Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 7, 223–242.
19. Graham, M.M., and Storkey, A.J. (2017). Continuously tempered Hamiltonian Monte Carlo. In *Proc. of Conference on Uncertainty in Artificial Intelligence*.
20. Hoffman, M.D., and Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1593–1623.
21. Łącki, M.K., and Miasojedow, B. (2015). State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Stat. Comput.* 26, 951–964.
22. Bellman, R.E. (1961). *Adaptive Control Processes* (Princeton University Press).
23. Taylor, A.N., and Kitching, T.D. (2010). Analytic methods for cosmological likelihoods. *Mon. Not. Roy. Astron. Soc.* 408, 865–875.
24. Loos, C., Krause, S., and Hasenauer, J. (2018). Hierarchical optimization for the efficient parameterization of ODE models. *Bioinformation* 34, 4266–4273.
25. Schmiester, L., Schälte, Y., Fröhlich, F., Hasenauer, J., and Weindl, D. (2020). Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformation* 36, 594–602.
26. Bachmann, J., Raue, A., Schilling, M., Böhm, M.E., Kreutz, C., Kaschek, D., Busch, H., Gretz, N., Lehmann, W.D., Timmer, J., and Klingmüller, U. (2011). Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.* 7, 516.
27. Raimúndez, E., Keller, S., Zwingenberger, G., Ebert, K., Hug, S., Theis, F.J., Maier, D., Lubner, B., and Hasenauer, J. (2020). Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines. *PLoS Comput. Biol.* 16, e1007147.
28. Maier, C., Loos, C., and Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformation* 33, 718–725.
29. Hasenauer, J., Hasenauer, C., Hucho, T., and Theis, F.J. (2014). ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.* 10, e1003686.
30. Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics, 4Bayesian Statistics*, pp. 169–193.
31. Henkelman, G., and Jónsson, H. (2000). Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* 113, 9978–9985.
32. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* 3, 861.
33. Schmiester, L., Schälte, Y., Bergmann, F.T., Camba, T., Dudkin, E., Egert, J., Fröhlich, F., Fuhrmann, L., Hauber, A.L., Kemmer, S., et al. (2021). PETA—interoperable specification of parameter estimation problems in systems biology. *PLoS Comput. Biol.* 17, 10086466–e1008710.
34. Girolami, M., and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. B73*, 123–214.
35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. *Proc. of the 31st Conference on Neural Information Processing Systems (NIPS)*.
36. Villaverde, A.F., Raimúndez, E., Hasenauer, J., and Banga, J.R. (2019). A comparison of methods for quantifying prediction uncertainty in systems biology. *IFAC-PapersOnLine* 52, 45–51.

37. Fujita, K.A., Toyoshima, Y., Uda, S., Ozaki, Y.i., Kubota, H., and Kuroda, S. (2010). Decoupling of receptor and downstream signals in the Akt pathway by its low-pass filter characteristics. *Sci. Signal.* *3*, ra56.
38. Boehm, M.E., Adlung, L., Schilling, M., Roth, S., Klingmüller, U., and Lehmann, W.D. (2014). Identification of isoform-specific dynamics in phosphorylation-dependent STAT5 dimerization by quantitative mass spectrometry and mathematical modeling. *J. Proteome Res.* *13*, 5685–5694.
39. Leonhardt, C., Schwake, G., Stögbauer, T.R., Rappl, S., Kuhr, J.T., Ligon, T.S., and Rädler, J.O. (2014). Single-cell mRNA transfection studies: Delivery, kinetics and statistics by numbers. *Nanomedicine* *10*, 679–688.
40. Villaverde, A.F., Fröhlich, F., Weindl, D., Hasenauer, J., and Banga, J.R. (2019). Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics* *35*, 830–838.
41. Fröhlich, F., and Sorger, P.K. (2022). Fides: Reliable trust-region optimization for parameter estimation of ordinary differential equation models. *PLoS Comput. Biol.* *18*, e1010322.
42. Hass, H., Loos, C., Raimúndez-Álvarez, E., Timmer, J., Hasenauer, J., and Kreutz, C. (2019). Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics* *35*, 3073–3082.
43. Kreutz, C. (2016). New concepts for evaluating the performance of computational methods. *IFAC-PapersOnLine* *49*, 63–70.
44. Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J., and Theis, F.J. (2017). Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst. Biol.* *11*, 63.
45. Vousden, W.D., Farr, W.M., and Mandel, I. (2016). Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Mon. Not. Roy. Astron. Soc.* *455*, 1919–1937.
46. Miasojedow, B., Moulines, E., and Vihola, M. (2013). An adaptive parallel tempering algorithm. *J. Comput. Graph Stat.* *22*, 649–664.
47. Schälte, Y., Fröhlich, F., Jost, P.J., Vanhoefer, J., Pathirana, D., Stapor, P., Lakrisenko, P., Wang, D., and Raimúndez, E. (2021). Merkt S., et al. pyPESTO: A modular and scalable tool for parameter estimation for dynamic models. Preprint at arXiv. [https://doi.org/10.48550/arXiv:2305.01821\[q-bio.QM\]](https://doi.org/10.48550/arXiv:2305.01821[q-bio.QM]).
48. Sokal, A. (1997). Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. In *Functional Integration. NATO ASI Series*, 361 (Springer).

## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Code for mathematical modeling, analysis and visualization used in the manuscript	This paper	<a href="https://doi.org/10.5281/zenodo.7199473">https://doi.org/10.5281/zenodo.7199473</a>
Software and algorithms		
Python version 3.10	Python Software Foundation	<a href="https://www.python.org">https://www.python.org</a>
AMICI (Python package)	Github	<a href="https://github.com/AMICI-dev/AMICI">https://github.com/AMICI-dev/AMICI</a>
pyPESTO (Python package)	Github	<a href="https://github.com/ICB-DCM/pyPESTO">https://github.com/ICB-DCM/pyPESTO</a>
Other		
Benchmark PESTab model repository collection	Github	<a href="https://github.com/Benchmarking-Initiative/Benchmark-Models-PEstab">https://github.com/Benchmarking-Initiative/Benchmark-Models-PEstab</a>

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jan Hasenauer ([jan.hasenauer@uni-bonn.de](mailto:jan.hasenauer@uni-bonn.de)).

#### Materials availability

This study did not generate new unique reagents.

#### Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the [key resources table](#).
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the [key resources table](#).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

### METHOD DETAILS

#### Mechanistic modeling of biological systems

We consider models based on ODEs of the form

$$\dot{x}(t, \theta) = f(x(t, \theta), \theta), x(t_0, \theta) = x_0(\theta),$$

in which the vector field  $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}^{n_x}$  determines the temporal evolution of the states  $x(t, \theta) \in \mathbb{R}^{n_x}$ . The unknown model parameters, which are estimated from the measurements, are denoted by  $\theta \in \mathbb{R}^{n_\theta}$ . Usually,  $\theta$  includes reaction rate constants and initial amounts of species. Here,  $n_x$  is the total number of modeled species, and  $n_\theta$  the total number of model parameters. The states  $x(t, \theta)$  and model parameters  $\theta$  are linked to the observables via the observation map  $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}^{n_y}$ , where  $n_y$  is the total number of observables. The observables are the measured properties of the model. Most measurement techniques only provide relative information about the absolute concentrations of interest<sup>8,9</sup> and, frequently, measurements are noise corrupted. Hence, to obtain the measurements  $\bar{y}$  (i) the model observables must be re-scaled by introducing scaling factors and offsets, and (ii) the model also must capture experimental errors by defining a noise model. Most commonly, independent and additive Gaussian distributed noise models are assumed

$$\bar{y}_{j,i} = s_{j,i} \cdot h_j(x(t_i, \theta), \theta) + b_{j,i} + \varepsilon_{j,i}, \text{ with } \varepsilon_{j,i} \sim \mathcal{N}(0, \sigma_{j,i}^2), \quad (\text{Equation 5})$$

with observable index  $j$ , time index  $i$ , scaling factors  $s \in \mathbb{R}^{n_y \times n_t}$ , offsets  $b \in \mathbb{R}^{n_y \times n_t}$ , and noise parameters  $\sigma \in \mathbb{R}^{n_y \times n_t}$ . Here,  $n_t$  denotes the total number of time points. These parameters are often unknown and, therefore, also need to be estimated along with the unknown model parameters. Other usual noise assumptions include log-normal distributed noise models<sup>11</sup> and Laplace distributed noise models.<sup>28</sup> In this study,

we focus on the case of additive Gaussian noise (5), but implementations for log-normal and Laplace distributed noise models are provided in [Tables S1](#) and [S2](#) and [supplemental data](#).

We denoted the group of all measurements as  $\mathcal{D} = \{\bar{y}_{j,i}\}_{i=1}^{j=(1,\dots,n_y)}_{i=(1,\dots,n_t)}$ .

### Benchmark models

For the evaluation of the marginalization-based approach, we employed in total five models (one toy model and four published M1–M4) and their corresponding datasets ([Table 1](#)).

#### *Toy: Model of a conversion reaction*

The conversion reaction model was introduced in<sup>28</sup> and describes a reversible chemical reaction, which converts a biochemical species A to a species B with rate  $\theta_1$ , and B to A with rate  $\theta_2$  ([Figure 2](#)). We modified the observation model to include scaling and offsets. For the evaluation of the proposed method, we generated one artificial dataset which is depicted in [Figure 2D](#). For details on the model structure and synthetic data generation we refer to the [supplemental data](#).

#### *M1: Model of EGF-dependent AKT pathway*

The model of EGF-dependent AKT pathway has been introduced in<sup>37</sup> and possesses in total 16 unknown parameters: 13 model parameters and 3 scaling factors ([Table 1](#), M1). The available experimental data are a total of 144 data points under 6 different experimental conditions for 3 observables. For each data point, the corresponding variance of the measurement noise is provided, therefore it does not need to be estimated. The complete dataset is depicted in [Figure S2](#).

#### *M2: Model of STAT5 dimerization*

The model of STAT5 dimerization has been introduced in<sup>38</sup> and possesses in total 9 unknown parameters: 6 model parameters and 3 noise parameters. To this model, we have added 3 scaling factors ([Table 1](#), M2), one per observable, for the sake of testing the proposed method. The available experimental data are a total of 48 data points for 3 observables. The complete dataset is depicted in [Figure S3](#).

#### *M3: Model of mRNA transfection*

The model for mRNA transfection has been introduced in<sup>39</sup> and possesses in total 5 unknown parameters: 3 model parameters, 1 scaling factor, and 1 noise parameter ([Table 1](#), M3). The complete dataset is depicted in [Figure 3D](#). For further details of the model structure we refer to the [supplemental data](#).

#### *M4: Model of gastric cancer signaling*

The model for gastric cancer signaling has been introduced in.<sup>27</sup> Here, we considered the Cetuximab responder cell line MKN1. The available experimental data for the responder cell line were a total of 303 data points under 106 different experimental conditions for 31 observables. For each data point, the corresponding variance of the measurement noise was provided, therefore it did not need to be estimated.

For all models we used the parameter ranges and prior distributions introduced in the original publications. The priors are mostly uninformative.

### Parameter optimization

To determine the maximum a posteriori (MAP) estimates, we minimized the negative log-posterior function. This minimization was performed using multi-start local optimization, an approach which was previously shown to be reliable.<sup>12,40</sup> For local optimization, we used the trust-region optimizer fides.<sup>41</sup> Parameters were  $\log_{10}$ -transformed to improve numerical properties.<sup>40,42,43</sup> We generated 100 starting points for local optimization, except for model M4 for which we used 500 starting points.

### Bayesian parameter inference

To perform Bayesian parameter inference, we used MCMC sampling following the pipeline presented in.<sup>44</sup> Similar to parameter optimization, sampling was performed using  $\log_{10}$ -transformed parameters. The MAP estimates  $(\hat{\theta}, \hat{s}, \hat{b}, \hat{\sigma}^2)$  for the full problem (aka without marginalization) were used to initialize the MCMC chains<sup>44</sup>: all runs for the standard sampling approach were initialized using the full optimal vector  $(\hat{\theta}, \hat{s}, \hat{b}, \hat{\sigma}^2)$  (found using multi-start local optimization); while for all runs for the marginalization-based sampling approach were initialized using the corresponding subset  $\hat{\theta}$ . Note that for the runs for the marginalization-based sampling approach also the MAP estimate for the marginalized problem could have been used, yet, differences were minor and the chosen approach allowed us to match runs of standard and marginalization-based sampling approach. The parameter posterior distribution was sampled using the adaptive Metropolis<sup>18</sup> and parallel tempering<sup>45,46</sup> algorithms implemented in the Python toolbox pyPESTO.<sup>47</sup> For the parallel tempering algorithm, we used 10 chains initialized. For all runs of the parallel tempering algorithm, we initialized the first chain – which samples the posterior – with the best optimization result found using multi-start local optimization, the second chain with the second best optimization result, and so on.

Convergence after burn-in was assessed using the Geweke test<sup>30</sup> and auto-correlation length using Sokal's adaptive truncated periodogram-estimator.<sup>48</sup> Both methods are implemented in pyPESTO and we refer to the respective original publications for technical details. The effective sample size is given by

$$n_{\text{eff}} = \frac{n}{1 + 2\sum_{\tau=1}^{\infty} \rho_{\tau}}$$

where  $n$  is the number of samples remaining after discarding burn-in period, and  $\rho_{\tau}$  is the estimated auto-correlation at lag  $\tau$ .

For all models, the prior hyperparameters for both sampling approaches were the same as used for optimization.

#### *Tempering scheme for the posterior analysis*

The posterior for standard and marginalization-based approach were tempered to assess transition characteristics (Figure 5). We used the tempered posteriors

$$p_T(\theta, s, \sigma^2 | \mathcal{D}) \propto (p(\mathcal{D} | \theta, s, \sigma^2) p(\theta, s, \sigma^2))^{1/T}.$$

and

$$p_T(\theta | \mathcal{D}) \propto (p(\mathcal{D} | \theta) p(\theta))^{1/T}.$$

with temperature  $T$ .