Posterior marginalization accelerates Bayesian inference for dynamical systems

Elba Raimúndez [1,2], Michael Fedders [1] and Jan Hasenauer [1,2,3,*]

[1] Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

[2] Technische Universität München, Center for Mathematics, Garching, Germany

[3] Helmholtz Zentrum München - German Research Center for Environmental Health, Computational Health Center, Neuherberg, Germany

[*] Corresponding author: Jan Hasenauer (jan.hasenauer@uni-bonn.de)

Abstract

Bayesian inference is an important method in the life and natural sciences for learning from data. It provides information about parameter uncertainties, and thereby the reliability of models and their predictions. Yet, generating representative samples from the Bayesian posterior distribution is often computationally challenging. Here, we present an approach that lowers the computational complexity of sample generation for problems with scaling, offset and noise parameters. The proposed method is based on the marginalization of the posterior distribution, which reduces the dimensionality of the sampling problem. We provide analytical results for a broad class of problems and show that the method is suitable for a large number of applications. Subsequently, we demonstrate the benefit of the approach for various application examples from the field of systems biology. We report a substantial improvement up to 50 times in the effective sample size per unit of time, in particular when applied to multi-modal posterior problems. 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 [1–3], patient progression [4, 5] and epidemics [6, 7]. 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), also characteristics of the measurement process may be unknown. In systems biology, most measurement techniques, including Western blotting [8], fluorescence microscopy [9] and mass spectrometry [10], are not fully quantitative but provide only relative information. Moreover, there is often an unknown offset and/or noise level [11]. 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 [12–14].

Bayesian inference is often used to determine unknown parameters [15–17]. A particularly common approach is to employ Markov chain Monte Carlo (MCMC) algorithms, such as (adaptive) Metropolis Hastings [18], Hamiltonian Monte Carlo methods [19, 20] and parallel tempering [21], to generate representative samples from the posterior distribution. Yet, with increasing number of unknown parameters, the application of MCMC algorithms becomes challenging [22]. 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 [23]). 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 [14, 24, 25]. 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. [24]). 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 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

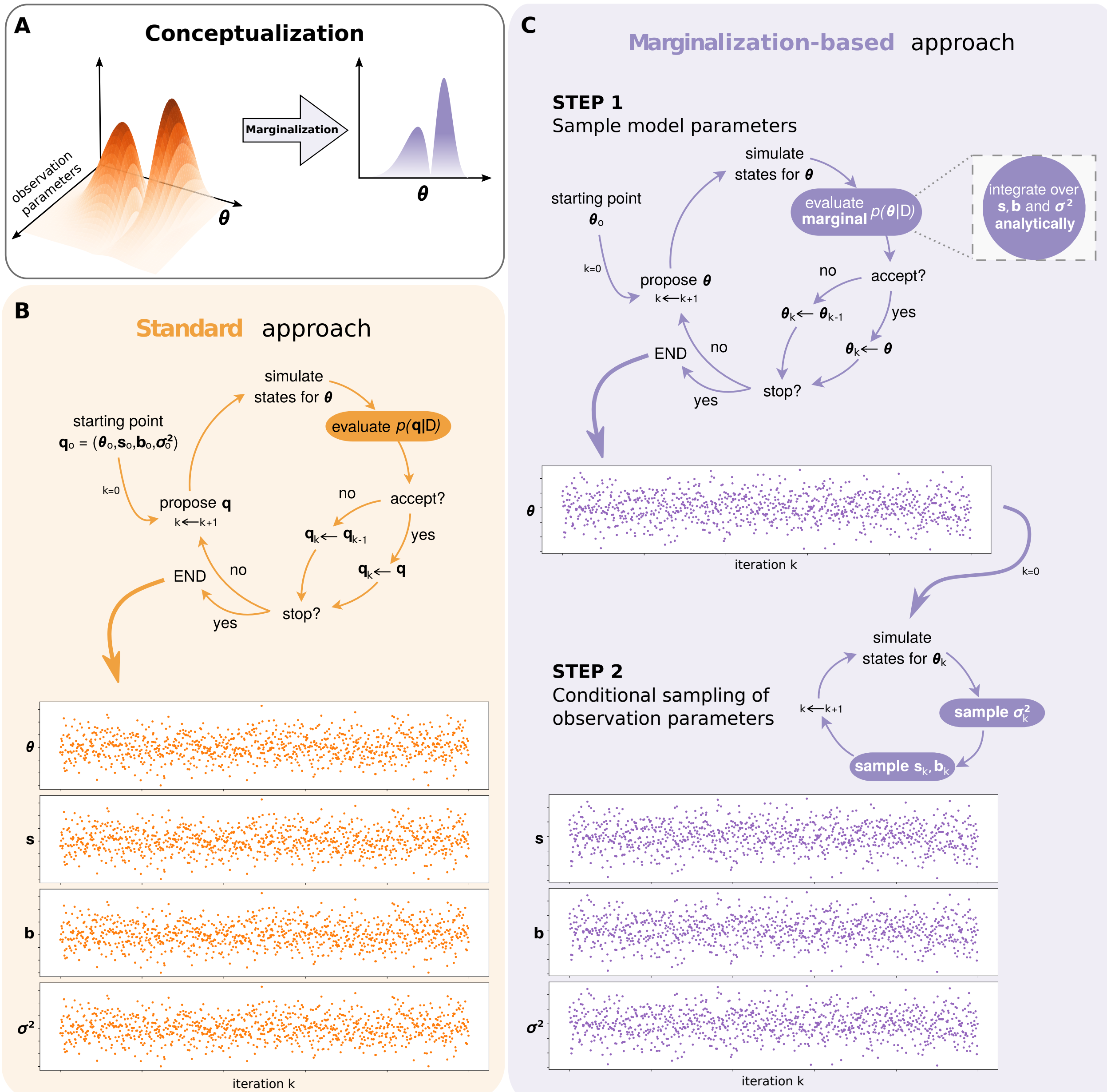

Figure 1: **Standard and marginalization-based Markov chain Monte-Carlo sampling.** (A) Illustration of the general marginalization concept. (B) Standard approach. (C) Marginalization-based approach depicting: (Step 1) the sequential integration of the observation parameters s, b and σ^2 to evaluate $p(\theta \mid \mathcal{D})$, and (Step 2) the (optional) conditional sampling of the marginalized observation parameters.

approach, we benchmark its performance with a collection of 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 a 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 ordinary differential equation (ODEs) and partial differential equation models, from data via observation models with scaling, offset and noise parameters. For the case of a mathematical model with parameter θ and time- and parameter-dependent states $x(t, \theta)$, we consider for the case of a one-dimensional observable with additive Gaussian measurement noise the observation model

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

in which $h(x, \theta)$ is the observable map, s is the scaling factor, b is the offset and σ^2 is the variance of the measurement noise. Following Bayes' theorem, the posterior distribution is given by

$$p(\theta, s, b, \sigma^2 \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta, s, b, \sigma^2)p(\theta)p(s, b, \sigma^2)}{p(\mathcal{D})}, \tag{2}$$

in which $p(\mathcal{D} \mid \theta, s, b, \sigma^2)$ denotes the likelihood of the data $\mathcal{D}$, $p(\theta, s, b, \sigma^2) = p(\theta)p(s, b, \sigma^2)$ denotes the prior distribution, and $p(\mathcal{D})$ denotes the marginal probability of the data.

The **standard approach** is to use MCMC methods to obtain representative samples from the joint posterior distribution for model parameters θ and observation parameters s, b and σ^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 depends on θ but not on s, b or σ^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. [26, 27]).

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 σ^2, yielding

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

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

$$p(\mathcal{D} \mid \theta) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(\mathcal{D} \mid \theta, s, b, \sigma^2) p(s, b, \sigma^2) \,\mathrm{d}s \,\mathrm{d}b \,\mathrm{d}\sigma^2. \tag{3}$$

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 σ^2,

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

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 $\Gamma^{-1}(\cdot)$ the Inverse-Gamma function. Here, we obtain for observations $\bar{y}_i$ with $i = 1, \ldots, n_t$ the closed-form expression for the marginal likelihood as

$$p(\mathcal{D} \mid \theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa\tau}{(n_t + \kappa)\left(\tau + \sum_{i=1}^{n_t} h_i^2\right) - \left(\sum_{i=1}^{n_t} h_i\right)^2}} \tag{4}$$

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

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

The combination of additive Gaussian noise and Normal-Inverse-Gamma prior is a common choice of conjugate distributions, which allow for an analytically tractable marginal likelihood. 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 [28]. Supplementary Tables S1–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 *Supplementary Material*.

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

The samples of model parameters θ from $p(\theta \mid \mathcal{D})$ allow for the assessment of the model properties and its uncertainties. In this regard, there is no difference of sampling the marginalized posterior distribution $p(\theta \mid \mathcal{D})$ compared to projecting the full posterior distribution $p(\theta, s, b, \sigma^2 \mid \mathcal{D})$ onto the θ 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 $p(s, b, \sigma^2 \mid \theta, \mathcal{D})$. 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 θ can be obtained by drawing from Gamma and Normal distributions:

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

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

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

$$\left.\lambda' = \lambda\left(\tau + \sum_{i=1}^{n_t} h_i^2 - \frac{\left(\sum_{i=1}^{n_t} h_i\right)^2}{(n_t + \kappa)}\right)\right),$$

with h_i and C being evaluated for model parameter θ. 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 *Supplementary Material*. As the conditional sampling can be performed independently and does not require model simulation, it is computationally efficient. For additional observation models see Supplementary Tables S1- S2.

In summary, a broad spectrum of sampling problems occurring in scientific disciplines, such as systems and computational biology, can be reformulated by performing an analytically tractable marginalization of their observation parameters. Sampling of this lower dimensional posterior distribution for the model parameters θ 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.

Table 1: **Key numbers and features of the considered toy and benchmark models.** The number of unknown model parameters n_θ, unknown scaling parameters n_s, unknown offset parameters n_b and unknown noise parameters n_σ, which are effectively sampled, are reported.

Model ID	n_θ	n_s	n_b	n_σ	Description	Reference
Toy	2	1	1	1	Conversion reaction	-
M1	13	3	-	-	EGF-AKT pathway	[29]
M2	6	3	-	3	STAT5 dimerization	[30]
M3	3	1	-	1	mRNA transfection	[31]
M4	26	31	-	-	Gastric cancer signaling	[27]

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 [28, 32] 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, θ_1, and the backward rate B to A, θ_2; and three observation parameters: the scaling s, the offset b and the noise variance σ^2 (Table 1). A detailed description of the model is provided in the *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 (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 Hasting method [18]. 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 θ_1 and θ_2 (Figure 2E–F), and the conditionally sampled observation parameters (Figure 2G–I).

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

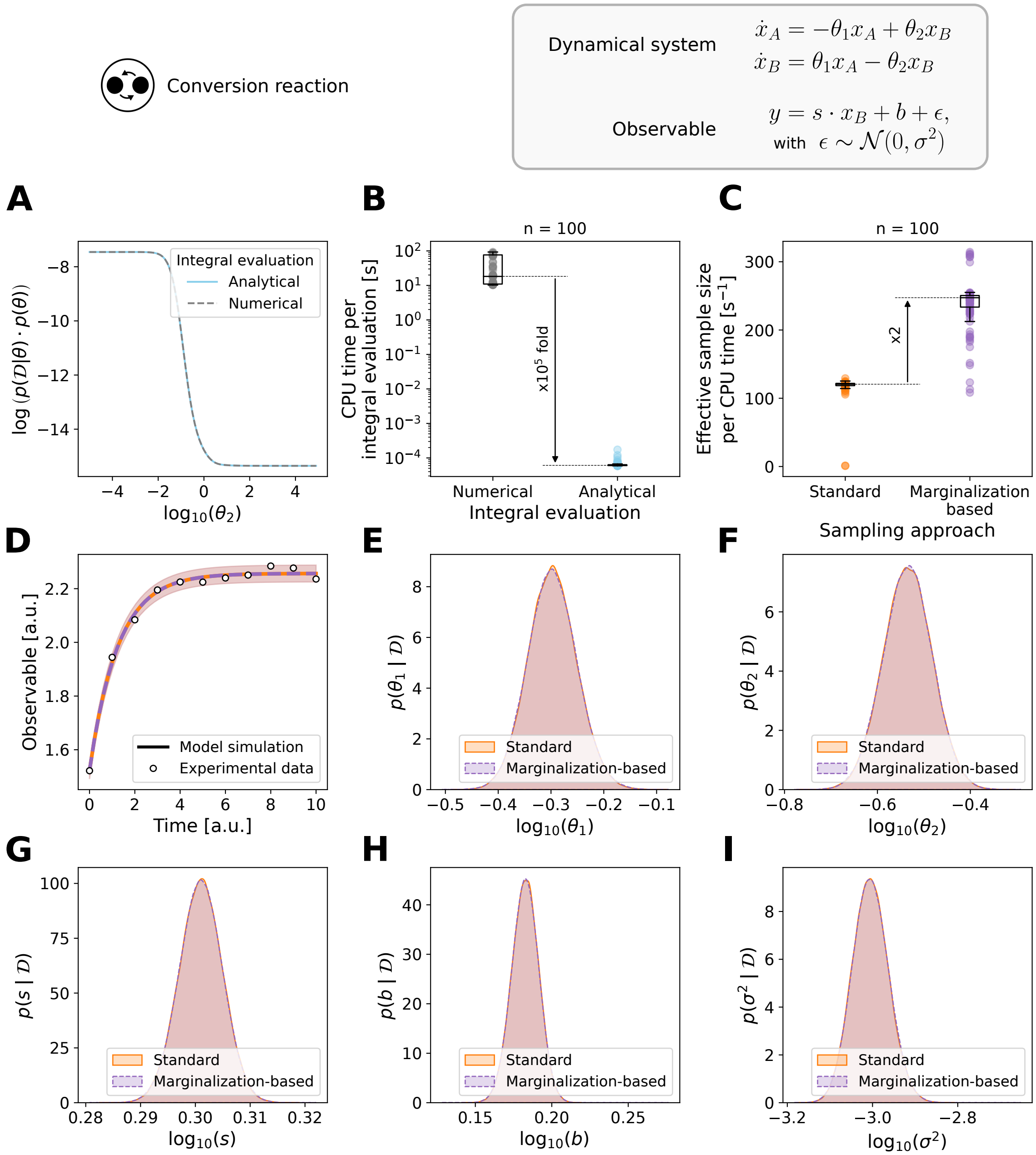

Figure 2: **Evaluation of the standard and marginalization-based approach for the toy model.** (A) Comparison of analytical vs. numerical integration. (B) Time comparison of analytical vs. numerical integration. (C) Effective sample size per unit of time for 100 independent runs. (D) Model fit of the 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) θ_1 and (F) θ_2, and the conditionally sampled observation parameters: (G) scaling factor s, (H) offset b, and (I) noise variance σ^2.

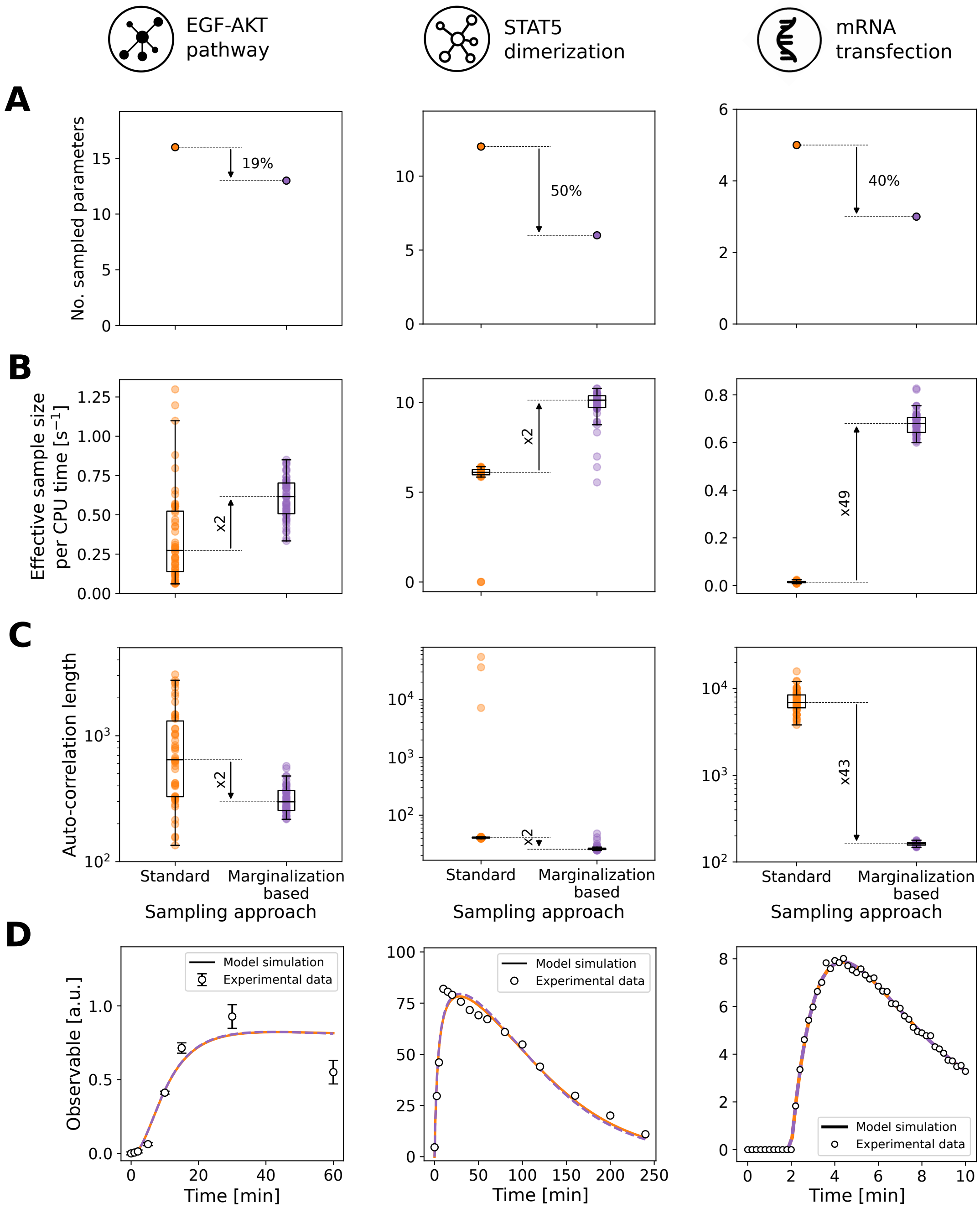

Figure 3: **Evaluation of the standard and marginalization-based approach for the benchmark models.** Models M1–M3 are shown from left to right. (A) Number of sampled parameters. (B) Effective sample size per unit of time. (C) Auto-correlation length. (D) Model fit of the best sample found during sampling. A subset of the experimental data is shown for M1 and M2. Complete datasets are depicted in Supplementary Figures S2 and S4.

(Table 1 and *Methods* section). The models M1 to M3 describe cellular processes: (M1) EGF-induced AKT signalling; (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 (Supplementary Tables S1– S2). More importantly, the full posterior distributions exhibit different characteristics, ranging for instance from uni- to bi-modal.

For the considered application problems, the marginalization of the observation parameters reduced the dimensionality of the sampling problems by up to 50% (Figure 3A). 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 [21] after assessing the correctness of the analytical marginalized likelihood for models M1–M3 (Supplementary Figure S9). All the runs were initialized at the local optima found during multi-start optimization [12], and run for 10^6 iterations. Further details are provided in the *Methods* section. The high number of iterations allowed all MCMC runs of the standard and marginalized problem to converge according to the Geweke test [33]. 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 2 (M1 and M2) to nearly 50 (M3) times higher efficiency in the marginalization-based approach. As the computation time was similar, the core reasons 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 (Figure 3D) as well as the parameter marginal distributions (Supplementary Figures S1, S3 and S5).

In summary, test and application problems demonstrates 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.

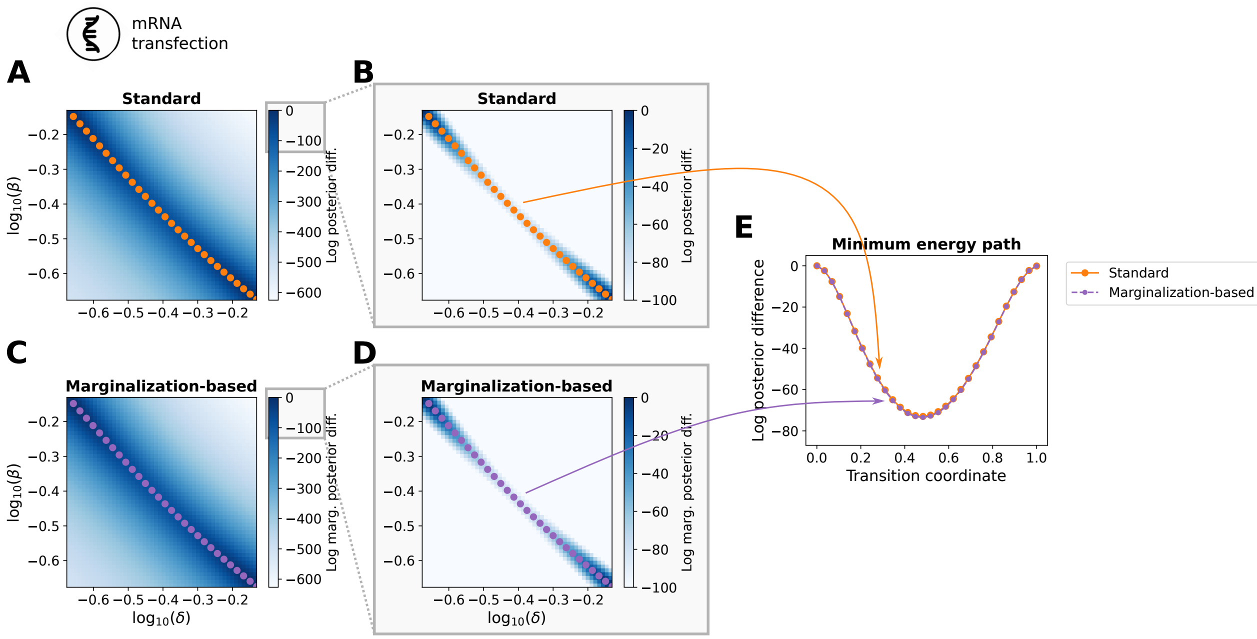

Figure 4: **Comparison of the minimum energy path for model M3.** Landscape of the optimized (A,B) posterior and (C,D) marginalized posterior for different fixed values of the model parameters β and δ. The difference with respect to the maximal posterior value is depicted. (E) Transition coordinates for the minimum energy path.

190 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 [34] for the standard (Figure 4A,B) and marginalized posterior (Figure 4C,D) (see details in the *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.

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 [18] with 10^6 iterations. This simplified the interpretation as it excludes the possibility of chain swaps. Yet, we found that for the given number of iterations this single-chain algorithm does essentially not transition between the modes (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 for the tempered posterior. We found that the marginalization-based approach allows already at lower temperatures for transitions between the modes unlike the standard sampling approach (Figure 5A and Supplementary Figures S7–S8). 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 (Figure 5B,C), while for the marginalization-based approach on average 1.6×10^4 transitions occurred (Figure 5D,E). As

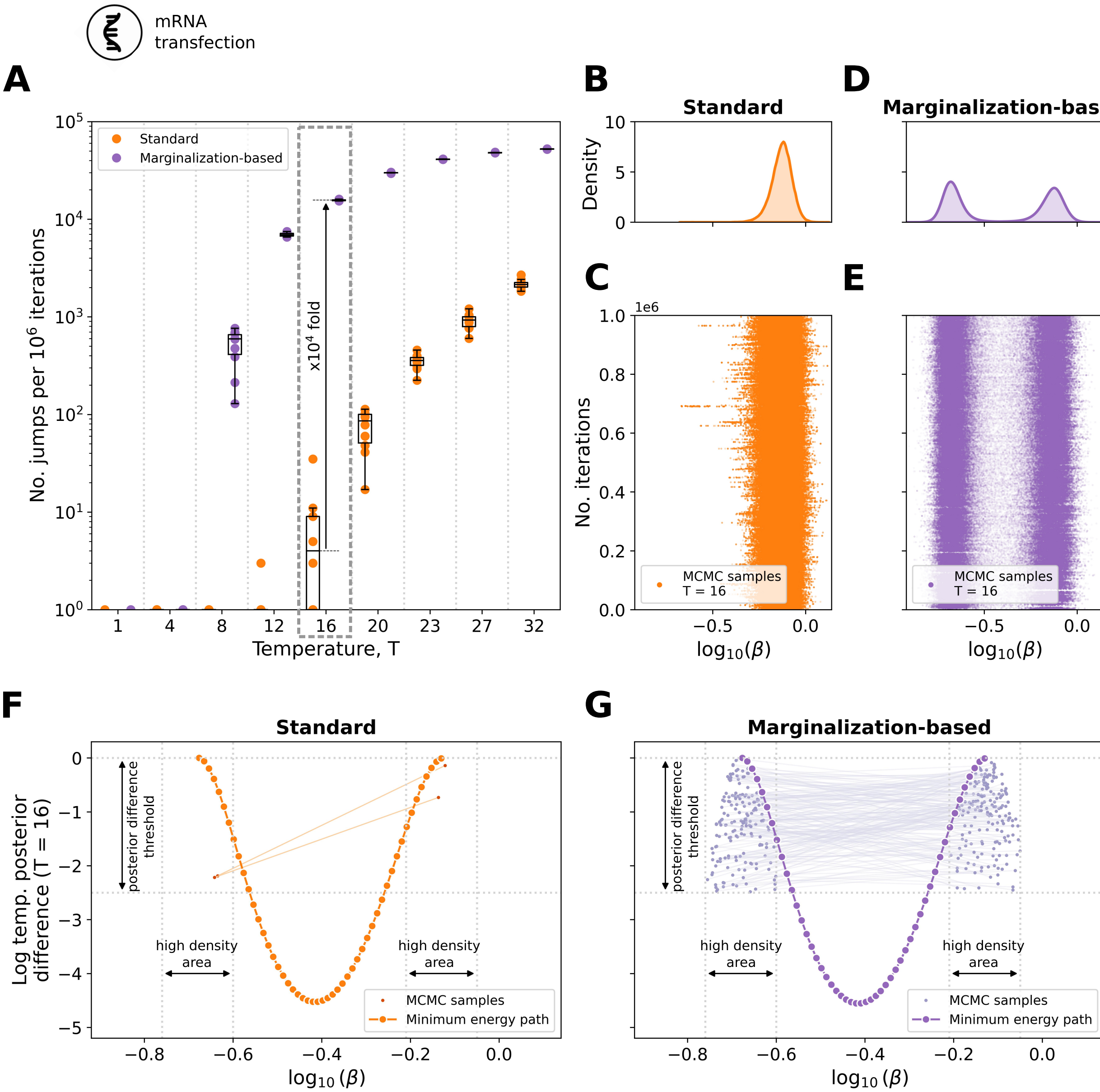

Figure 5: **Quantification of the transitions between the posterior modes for different temperatures T for model M3.** (A) Number of transitions per 10^6 iterations for a range of temperatures for the standard (orange) and marginalization-based (purple) approach. A total of 10 chains per temperature value are depicted. (B,D) Marginal distribution computed using a kernel density estimate and (C,E) parameter trace for the model parameter β of a representative chain obtained with the (B,C) standard and (D,E) marginalization-based approach for $T = 16$. (F,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 $T = 16$.

the minimum barrier energy is conserved also for higher temperatures (Supplementary Figure S6), 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 (Figure 5F,G), meaning that it transitions between 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 behaviour 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 [27]. 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 [18] and the adaptive parallel tempering algorithm [21] employed in the previous sections were run 10

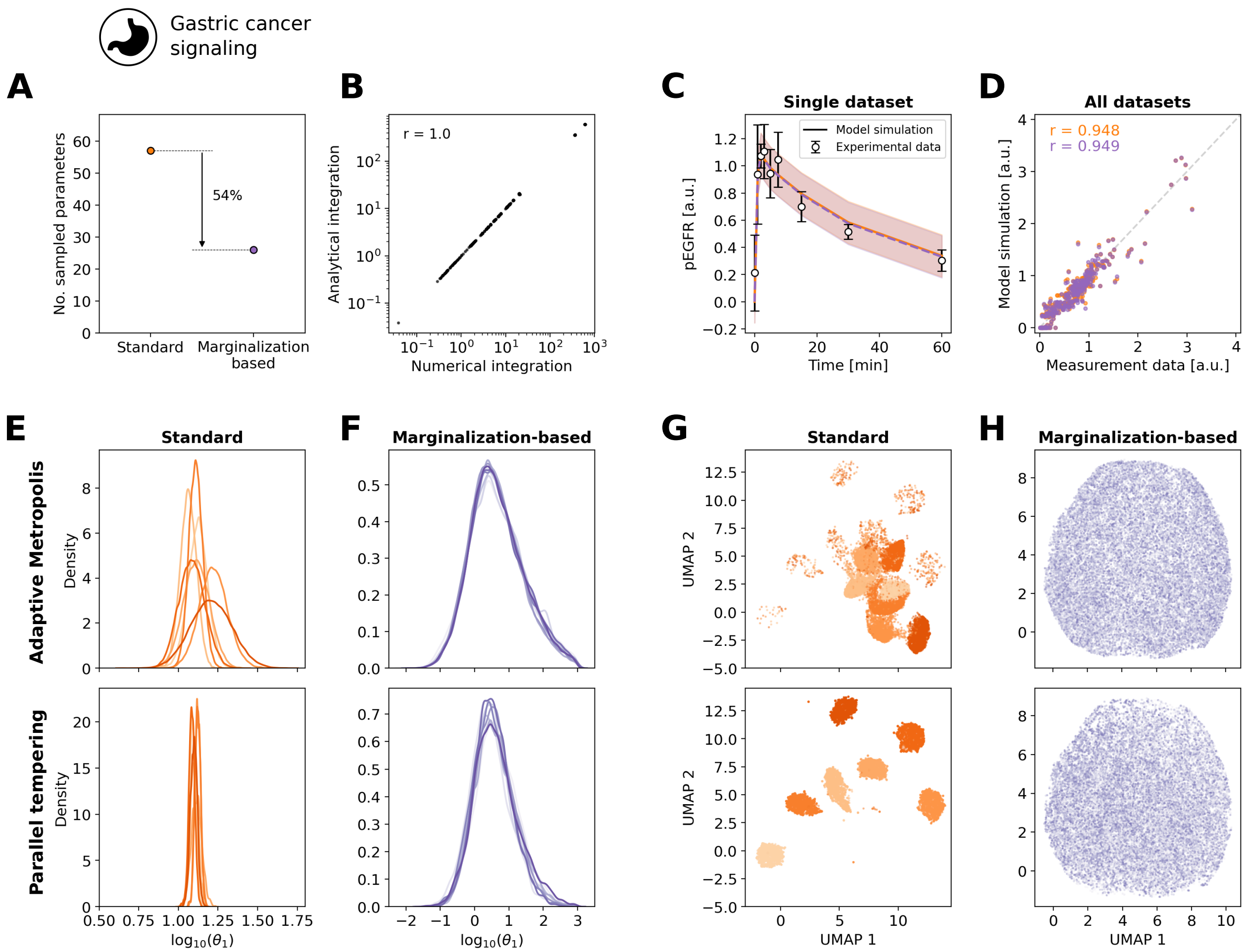

Figure 6: **Convergence of the marginalization-based approach for model M4.** (A) Number of sampled parameters. (B) Scatter plot for the agreement of analytical and numerical integration. (C,D) Model fit of the best sample found during sampling for, (C) a subset of the experimental data and (D) the complete dataset in form of a scatter plot, the standard (orange) and marginalization-based approach (purple). (E–H) Results from adaptive Metropolis (top) and parallel tempering (bottom) are shown. (E,F) Parameter marginal posterior distribution obtained using the (E) standard and (F) marginalization-based approach computed using a kernel density estimate for model parameter θ_1. (G,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` [35].

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. The maximum a posteriori estimates observed in the different runs provided similar fits (Figure 6C,D). In contrast, the marginal distributions of the model parameters differed, with the marginalization-based approach mostly providing broader parameter distributions than the standard approach (Figure 6E,F). The assessment of the reproducibility of the marginal distributions revealed a high variability between different runs performed using the standard approach (Figure 6E and Supplementary Figure S10). On the contrary, for the marginalization-based approach a good agreement between runs was observed (Figure 6F and Supplementary Figure 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 [35]. 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 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.

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 CPU time of 150 hours rendering its application impracticable. Furthermore, our study provides improved estimates for the parameters (Supplementary Figure S12) 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 large fraction of the overall parameters [12]. This complicates sampling and can render the generation of representative samples practically infeasible. Here, we address this challenge by introducing a framework which employs (analytical) marginalization. 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 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. This is consistent with previous finding for hierarchical optimization [25]. 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.

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 (Supplementary Tables S1 and S2). We provide the detailed derivation in the *Supplementary Material*. 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 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.

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 [19, 20]. As the marginal likelihood is differentiable, merely the derivation and implementation of the gradient is required. The usage of methods which exploit the Riemann geometry of the parameter space of statistical models, e.g., Metropolis-adjusted Langevin algorithm [36], 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 [37].

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 [16]. Accordingly, we expect that it might contribute to resolving reliability problems of Bayes prediction uncertainty analysis encountered in recent studies [38]. Furthermore, the proposed approach is not limited to ODEs, but directly applicable for other deterministic models, e.g. partial differential equations. As well, the idea might be incorporated in likelihood-free inference schemes used for stochastic and multi-scale models [39, 40]. Among other things, it might be used in exact Approximate Bayesian Computation schemes [41] by reformulating the acceptance probability.

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.

Methods

Mechanistic modeling of biological systems

We consider models based on ODEs of the form

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

in which the vector field $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$ 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, θ includes reaction rate constants and initial amounts of species. Here, n_x is the total number of modeled species, and n_θ the total number of model parameters. The states $x(t,\theta)$ and model parameters θ are linked to the observables via the observation map $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_y}$, 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 [8, 9] and, frequently, measurements are noise corrupted. Hence, to obtain the measurements $\bar{y}$ (i) the model observables must be rescaled 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

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

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 [11] and Laplace distributed noise models [28]. 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 Supplementary Tables S1–S2 and *Supplementary Material*.

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

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 [28] and describes a reversible chemical reaction, which converts a biochemical species A to a species B with rate θ_1, and B to A with rate θ_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 Supplementary Material.

M1: Model of EGF-dependent AKT pathway

The model of EGF-dependent AKT pathway has been introduced in [29] 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 Supplementary Figure S2.

M2: Model of STAT5 dimerization

The model of STAT5 dimerization has been introduced in [30] 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 Supplementary Figure S4.

M3: Model of mRNA transfection

The model for mRNA transfection has been introduced in [31] 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 Supplementary Material.

M4: Model of gastric cancer signaling

The model for gastric cancer signalling has been introduced in [27]. 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.

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 [12, 42]. For local optimization, we used the trust-region optimizer `fides` [43]. Parameters were $\log_{10}$-transformed to improve numerical properties [42, 44, 45]. 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 [46]. The MAP estimates were used to initialize the MCMC chains [46]: the full optimal vector $(\theta, s, b, \sigma^2)^*$ to initialize the standard approach runs, while for the marginalization-based approach runs the corresponding subset θ^* from $(\theta, s, b, \sigma^2)^*$ was used.

The parameter posterior distribution was sampled using the adaptive Metropolis [18] and parallel tempering [47, 48] algorithms implemented in the Python toolbox pyPESTO [49]. For the parallel tempering algorithm, we used 10 chains initialized at the 10 best local optima found during multi-start optimization for both approaches.

Convergence after burn-in was assessed using the Geweke test [33] and auto-correlation length using Sokal's adaptive truncated periodogram-estimator [50], both also available under pyPESTO. 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 ρ_τ is the estimated auto-correlation at lag τ.

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

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

and

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

with temperature T.

Implementation and data availability

Models M1, M2 and M4 were taken from the `PEtab` benchmark collection [51] which is based on [44] and available at `https://github.com/Benchmarking-Initiative/Benchmark-Models-PEtab`. As model M3 is analytically solvable, we implemented the solution in Python code. For ODE integration (models M1, M2 and M4) we used the Python toolbox AMICI [52]. For optimization and sampling, we used the Python toolbox pyPESTO [49]. pyPESTO already offers an interface to the `fides` optimizer [43]. For the UMAP visualizations and the minimum energy path calculation, we used respectively the Python packages `umap` [35] and `mep` https://github.com/chc273/mep.

All code and models used in this study are available from the Zenodo database at `https://doi.org/10.5281/zenodo.7199473`.

Acknowledgements

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.

Competing interests

The authors declare no competing interests.

Supplementary Figures and Tables

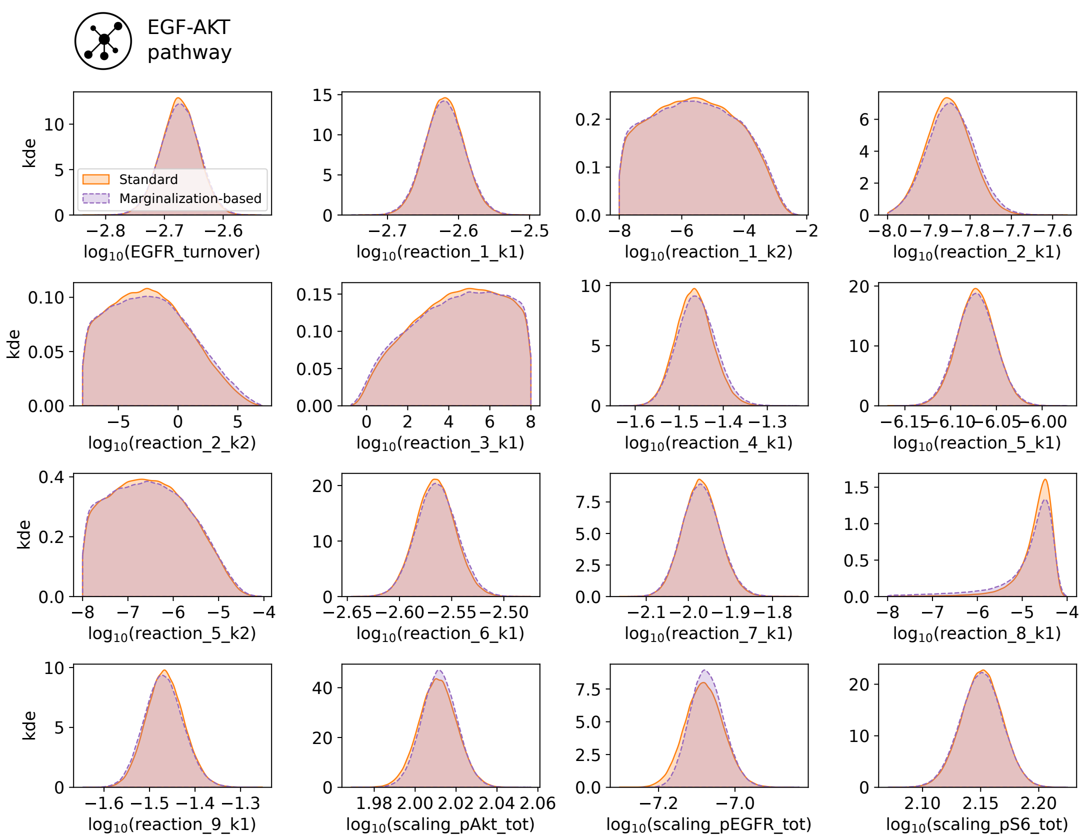

Figure S1: **Parameter marginal posterior distributions computed using a kernel density estimate for model M1.** The marginalized parameters, which are conditionally sampled, correspond to those denoted with *scaling_**.

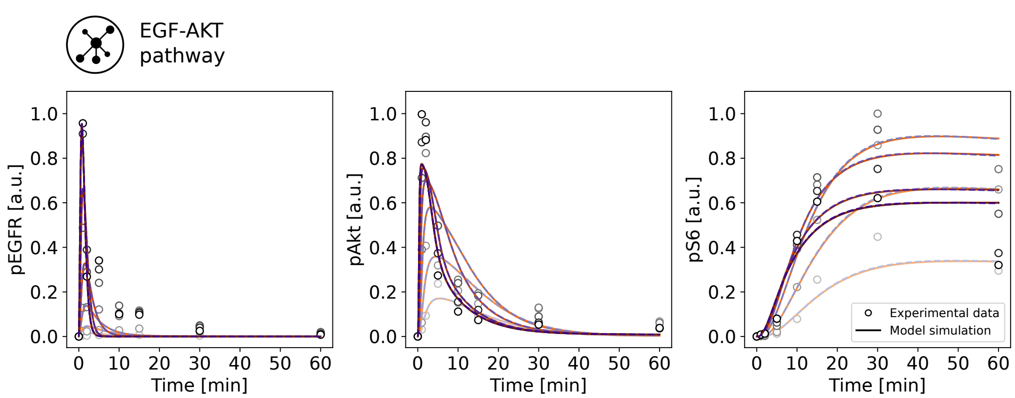

Figure S2: **Complete dataset and model fit for model M1.** Model simulation of the best sample found for the standard approach is depicted in orange and for the marginalization-based approach in purple. Different shades indicate different experimental conditions.

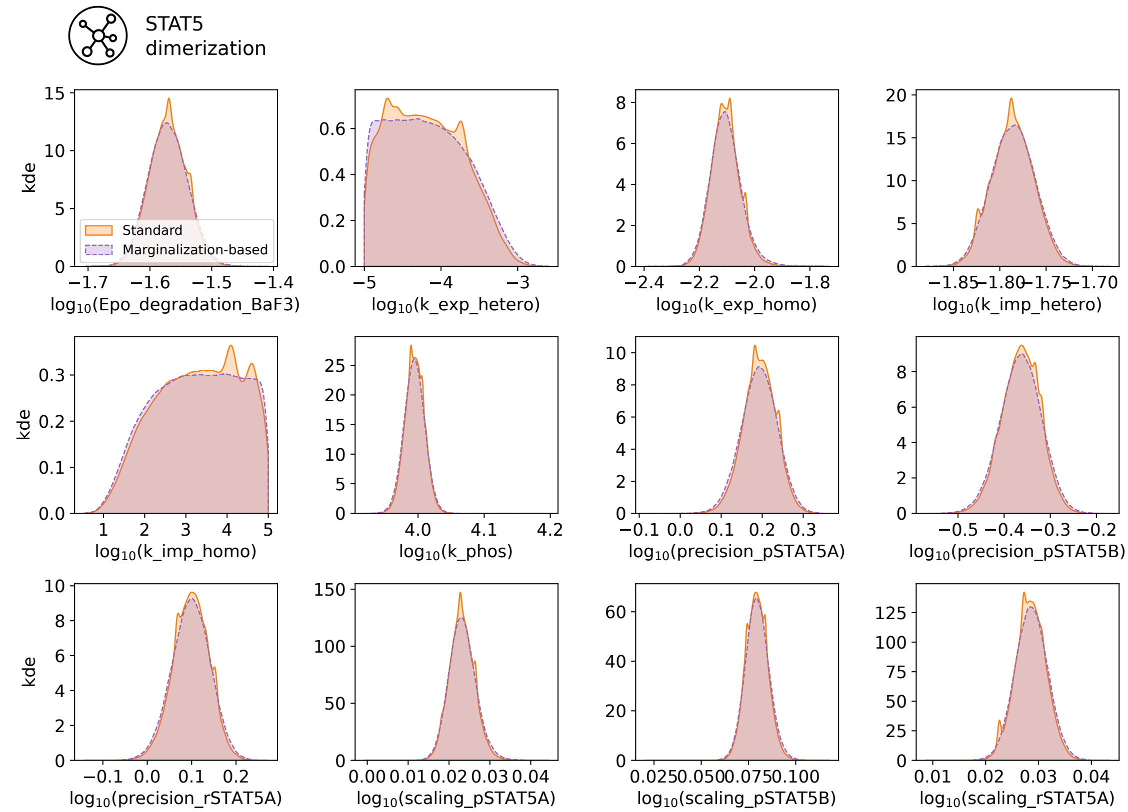

Figure S3: **Parameter marginal posterior distributions computed using a kernel density estimate for model M2.** The marginalized parameters, which are conditionally sampled, correspond to those denoted with *scaling_** and *precision_**.

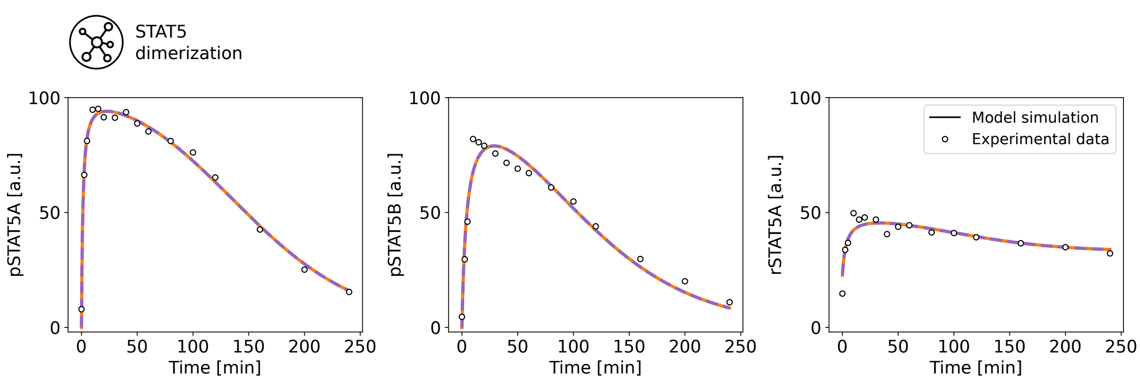

Figure S4: **Complete dataset and model fit for model M2.** Model simulation of the best sample found for the standard approach is depicted in orange and for the marginalization-based approach in purple.

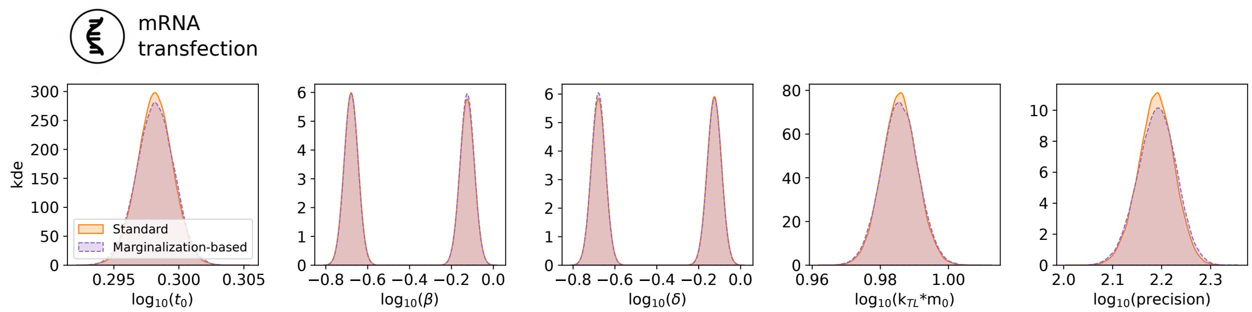

Figure S5: **Parameter marginal posterior distributions computed using a kernel density estimate for model M3.** The marginalized parameters, which are conditionally sampled, correspond to $k_{TL} * m_0$ and *precision*.

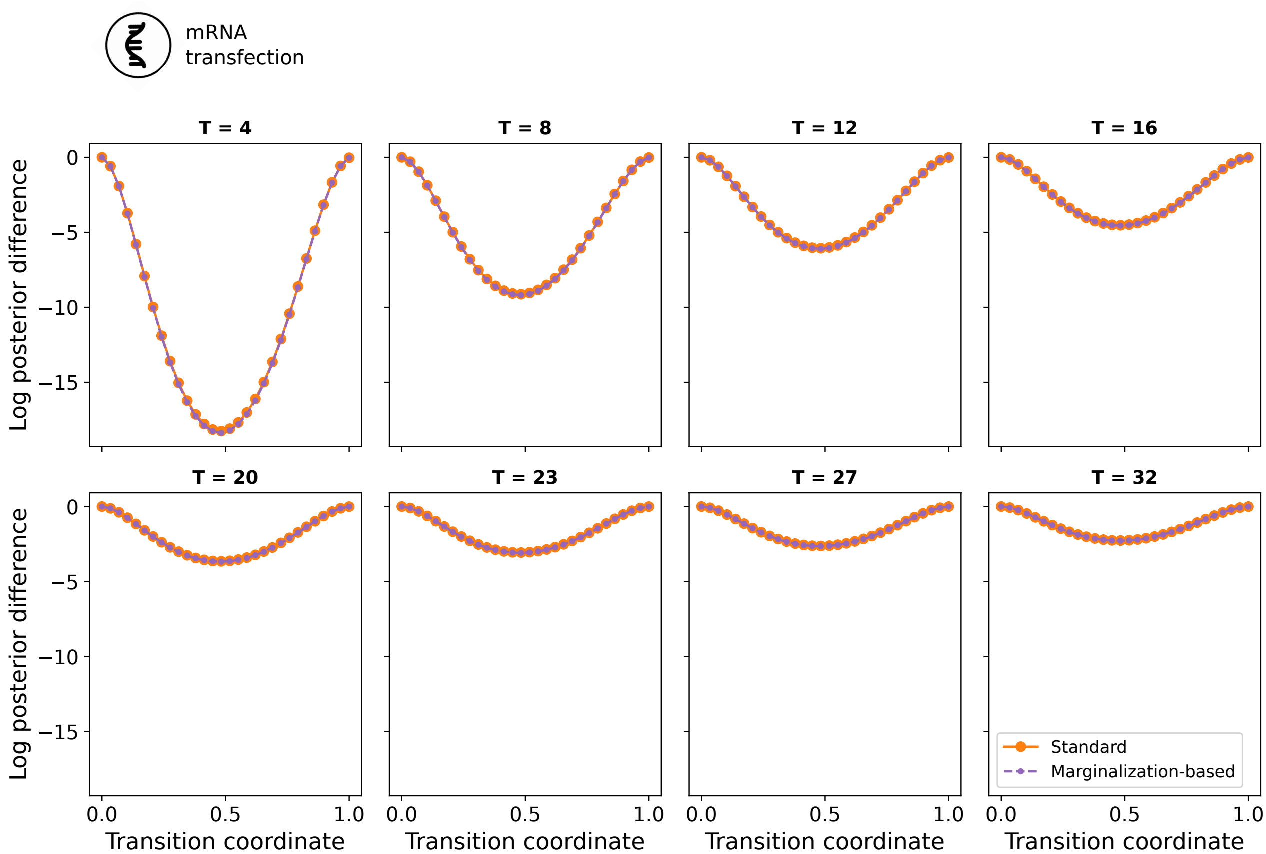

Figure S6: **Minimum energy path of the tempered posteriors for a range of temperatures considered in Figure 5A for model M3.**

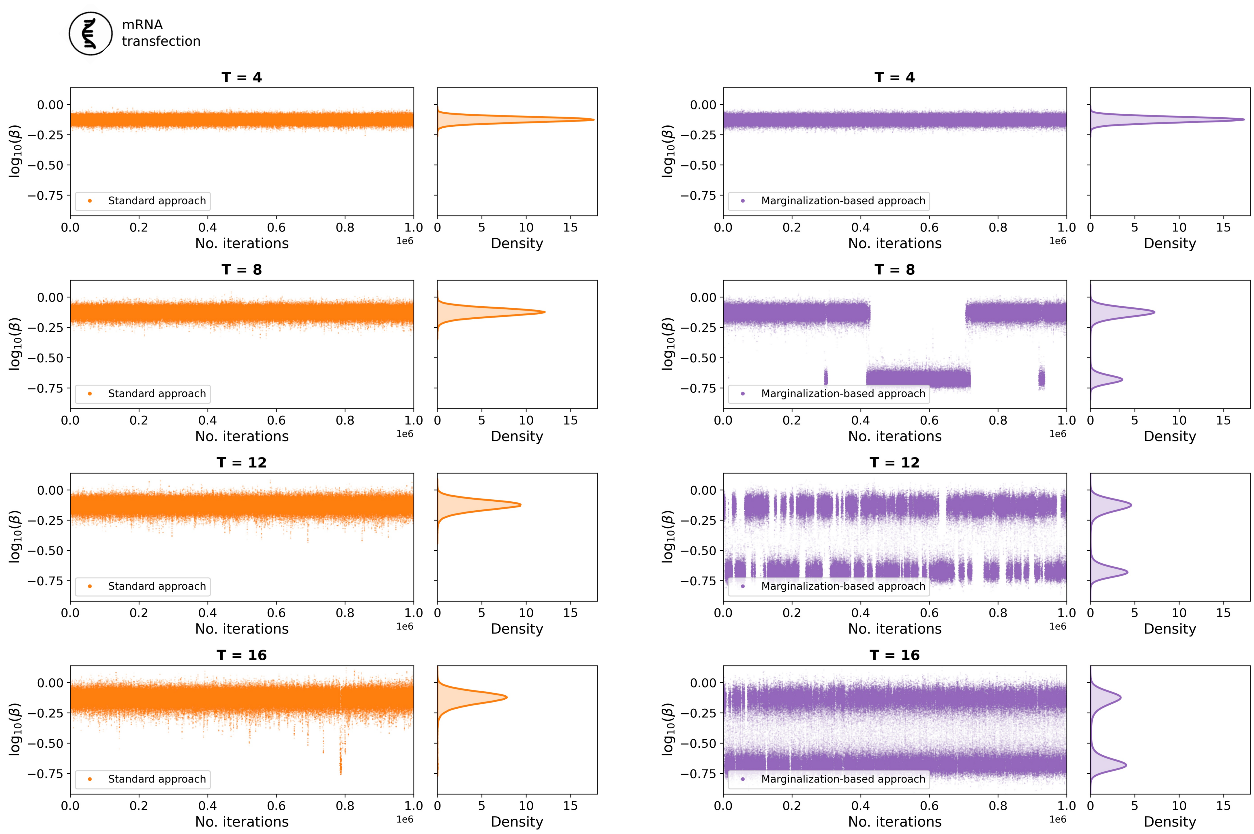

Figure S7: **Representative parameter traces for the model parameter β for a range of temperatures considered in Figure 5A for model M3.**

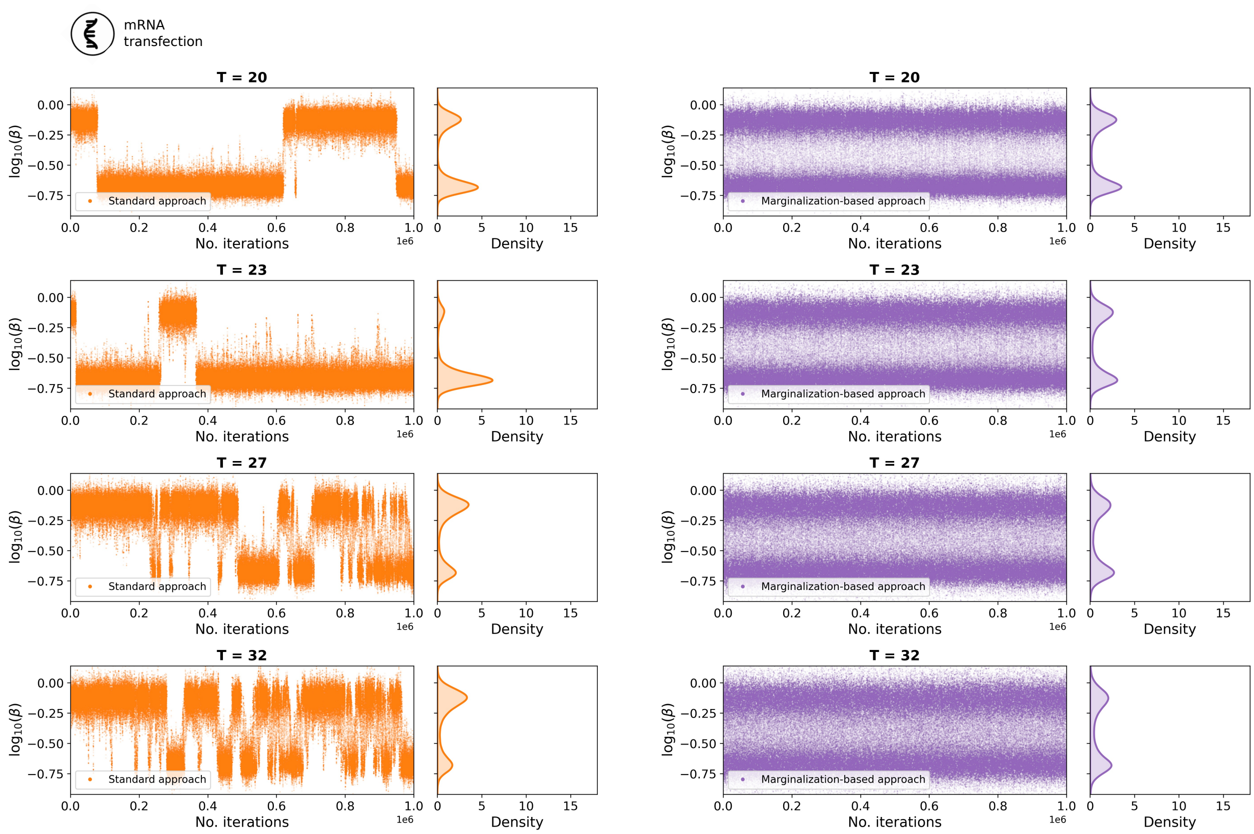

Figure S8: **Representative parameter traces for the model parameter β for a range of temperatures considered in Figure 5A for model M3.**

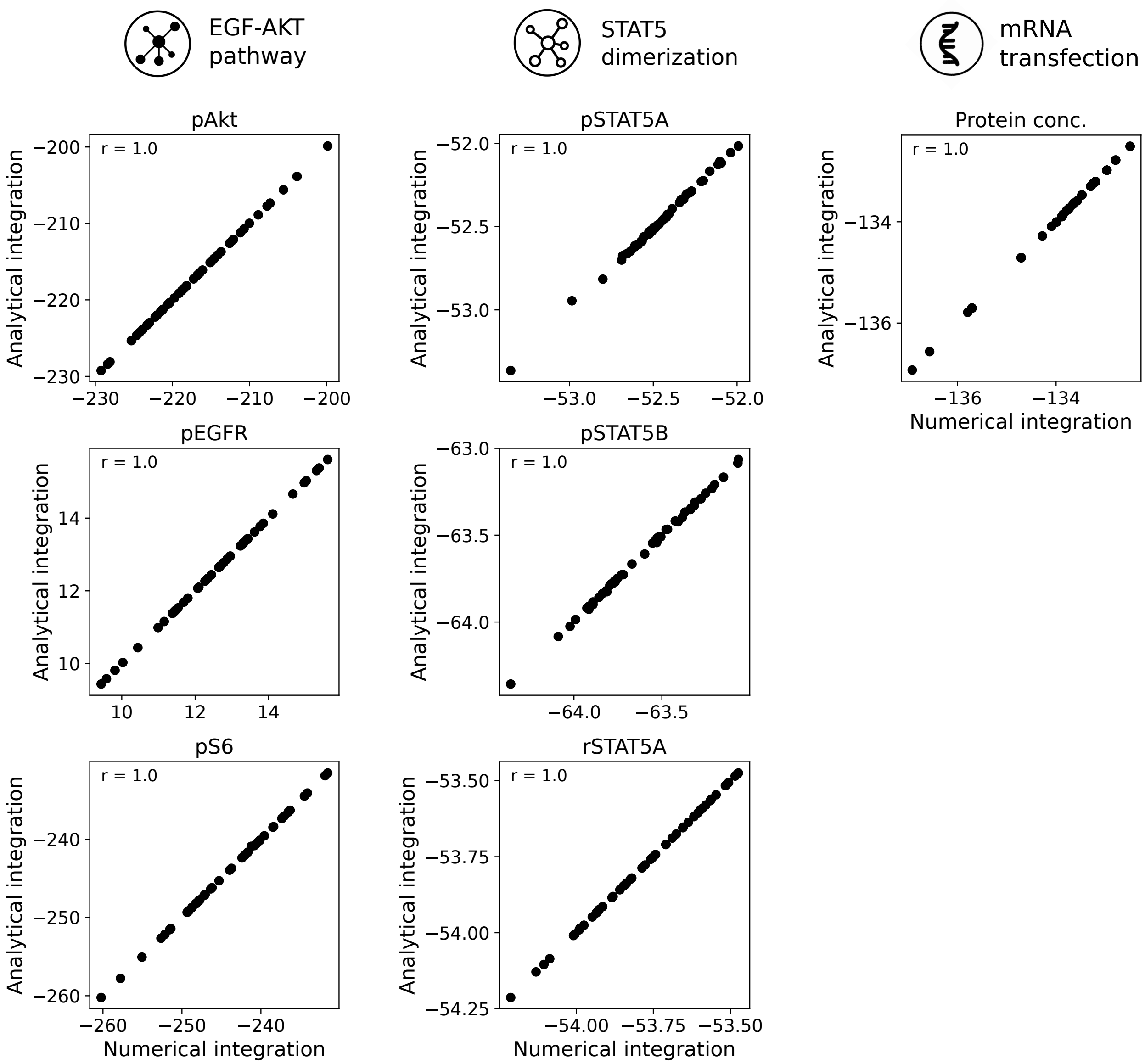

Figure S9: **Correctness of the analytical integration for model M1, M2 and M3.** Scatter plot for the agreement of analytical and numerical integration for 50 different parameter vectors. The integration results are shown for each model observable.

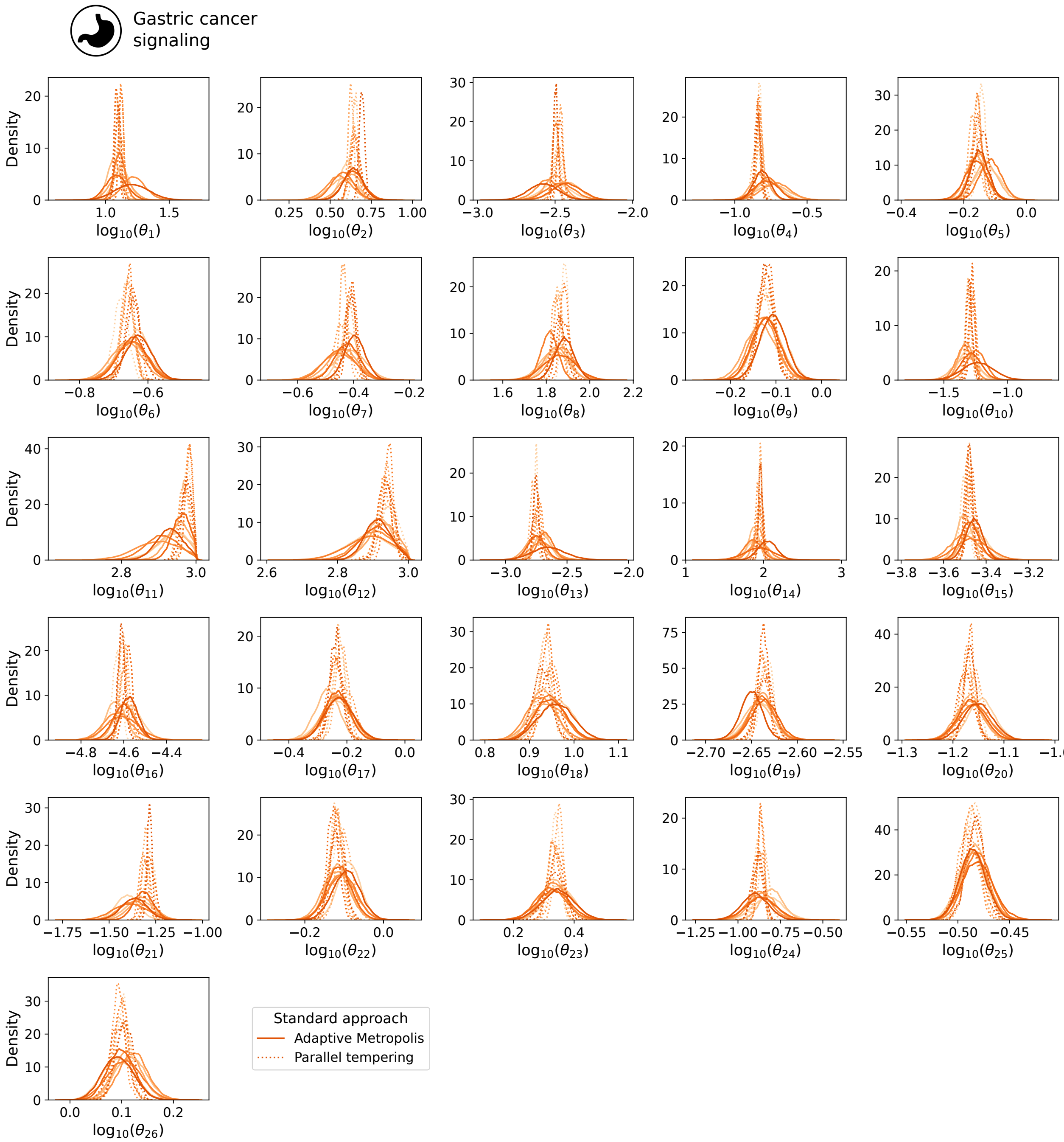

Figure S10: **Parameter marginal posterior distributions using the standard approach for model M4.** Results from two sampling algorithms (adaptive Metropolis and parallel tempering) and only the subset of model parameters are shown. The marginals were computed using a kernel density estimate.

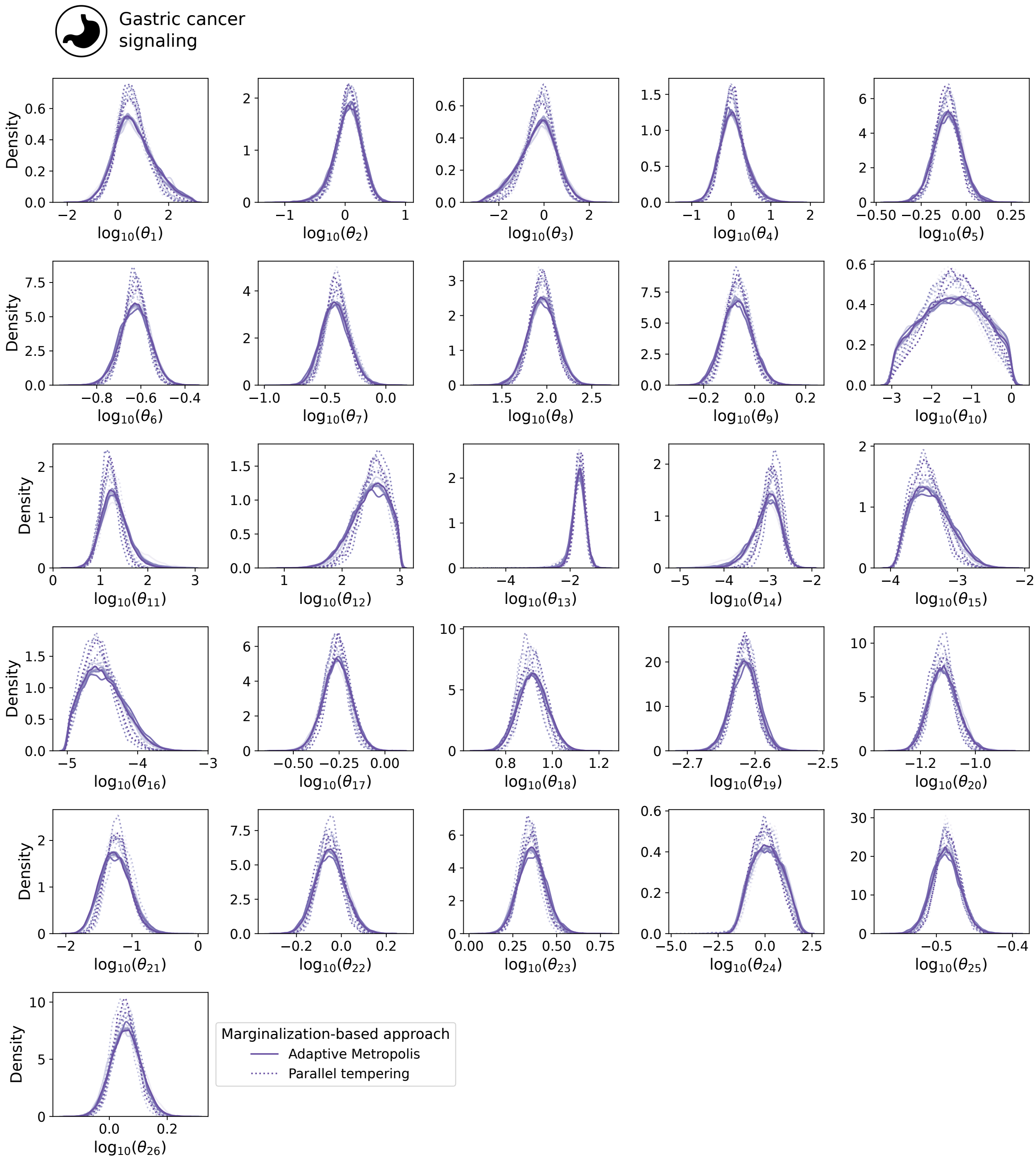

Figure S11: **Parameter marginal posterior distributions using the marginalization-based approach for model M4.** Results from two sampling algorithms (adaptive Metropolis and parallel tempering) and only the subset of model parameters are shown. The marginals were computed using a kernel density estimate.

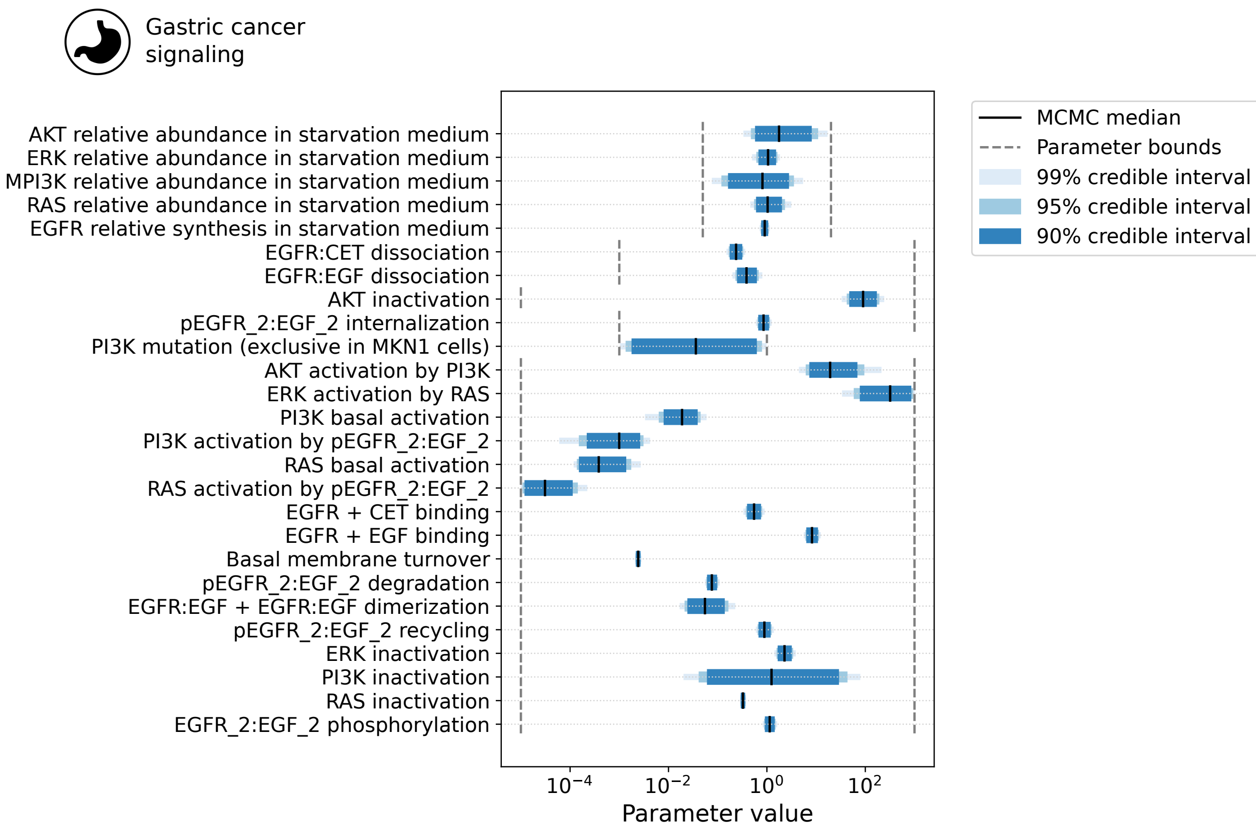

Figure S12: **Credible intervals for the model parameters of model M4.** The credible intervals were extracted from the MCMC samples obtained with the marginalization-based approach. The credible levels 90%, 95% and 99% are shown. Parameter bounds used for sampling are indicated in black dashed lines. Only the subset of model parameters are shown.

Table S1: **Overview of the marginalization-based approach applied to different observable combinations under unknown additive and multiplicative Gaussian measurement noise.** Observation parameters considered are scaling factors (s) and offsets (b). The noise is denoted as precision $\lambda := 1/\sigma^2$. For multiplicative noise, the logarithm of the scaling factor ($s_{\log}$) is used. Unknown/estimated observation parameters are denoted by ✓, otherwise the fixed numerical value is shown. Further details for each case are in the *Supplementary Material*.

	s	b	λ	Prior distribution	Analytical solution	Conditional sampling
Additive	✓	✓	✓	$p(s, b, \lambda \mid \nu, \tau, \mu, \kappa, \alpha, \beta) =$ $= \mathcal{N}(s \mid \nu, (\lambda\tau)^{-1}) \cdot \mathcal{N}(b \mid \mu, (\lambda\kappa)^{-1}) \cdot \Gamma(\lambda \mid \alpha, \beta)$ $= \sqrt{\frac{\lambda\tau}{2\pi}} \exp\left(-\frac{\lambda\tau(s-\nu)^2}{2}\right) \cdot \sqrt{\frac{\lambda\kappa}{2\pi}} \exp\left(-\frac{\kappa\lambda(b-\mu)^2}{2}\right)$ $\cdot \frac{\lambda^{\alpha-1}\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta\lambda)$	$p(\mathcal{D} \mid \theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa\tau}{(\kappa+n_t)\left(\tau+\sum_{i=1}^{n_t} h_i^2\right)-\left(\sum_{i=1}^{n_t} h_i\right)^2}}$ with $C := \beta + \frac{1}{2}\left(\kappa\mu^2 + \tau\nu^2 + \sum_{i=1}^{n_t} \bar{y}_i^2 - \frac{\left(\kappa\mu+\sum_{i=1}^{n_t} \bar{y}_i\right)^2}{\kappa+n_t}\right.$ $\left. - \frac{\left(\left(\kappa\mu+\sum_{i=1}^{n_t} \bar{y}_i\right)\left(\sum_{i=1}^{n_t} h_i\right)-(\kappa+n_t)\left(\tau\nu+\sum_{i=1}^{n_t} h_i\bar{y}_i\right)\right)^2}{(\kappa+n_t)\left((\kappa+n_t)\left(\tau+\sum_{i=1}^{n_t} h_i^2\right)-\left(\sum_{i=1}^{n_t} h_i\right)^2\right)}\right)$	$\lambda \propto \Gamma\left(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C\right)$ $b \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\left(\sum_{i=1}^{n_t} \bar{y}_i - h_i\right)}{\kappa+n_t}, \lambda' = \lambda(\kappa+n_t)\right)$ $s \propto \mathcal{N}\left(\mu' = \frac{(\kappa+n_t)\left(\tau\nu+\sum_{i=1}^{n_t} h_i\bar{y}_i\right)-\left(\kappa\mu+\sum_{i=1}^{n_t} \bar{y}_i\right)\left(\sum_{i=1}^{n_t} h_i\right)}{(\kappa+n_t)\left(\tau+\sum_{i=1}^{n_t} h_i^2\right)-\left(\sum_{i=1}^{n_t} h_i\right)^2},\right.$ $\left.\lambda' = \lambda\left(\tau + \sum_{i=1}^{n_t} h_i^2 - \frac{\left(\sum_{i=1}^{n_t} h_i\right)^2}{(\kappa+n_t)}\right)\right)$
	✓	0	✓	$p(s, \lambda \mid \mu, \kappa, \alpha, \beta) =$ $= \mathcal{N}(s \mid \mu, (\lambda\kappa)^{-1}) \cdot \Gamma(\lambda \mid \alpha, \beta)$ $= \frac{\beta^\alpha\sqrt{\kappa}}{\Gamma(\alpha)\sqrt{2\pi}} \lambda^{\alpha-1/2} \exp\left(-\beta\lambda - \frac{\kappa\lambda(s-\mu)^2}{2}\right)$	$p(\mathcal{D} \mid \theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa}{\kappa+\sum_{i=1}^{n_t} {h_i}^2}}$ with $C := \beta + \frac{1}{2}\left(\kappa\mu^2 + \sum_{i=1}^{n} \bar{y}_i^2 - \frac{\left(\kappa\mu+\sum_{i=1}^{n} \bar{y}_i h_i\right)^2}{\kappa+\sum_{i=1}^{n_t} {h_i}^2}\right)$	$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C)$ $s \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\sum_{i=1}^{n_t} \bar{y}_i h_i}{\kappa+\sum_{i=1}^{n_t} h_i^2}, \lambda' = \lambda\left(\kappa + \sum_{i=1}^{n_t} h_i^2\right)\right)$
	1	✓	✓	$p(b, \lambda \mid \mu, \kappa, \alpha, \beta) =$ $= \mathcal{N}(b \mid \mu, (\lambda\kappa)^{-1}) \cdot \Gamma(\lambda \mid \alpha, \beta)$ $= \frac{\beta^\alpha\sqrt{\kappa}}{\Gamma(\alpha)\sqrt{2\pi}} \lambda^{\alpha-1/2} \exp\left(-\beta\lambda - \frac{\kappa\lambda(b-\mu)^2}{2}\right)$	$p(\mathcal{D} \mid \theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa}{\kappa+n_t}}$ with $C := \beta + \frac{1}{2}\left(\kappa\mu^2 + \sum_{i=1}^{n_t} (\bar{y}_i - h_i)^2 - \frac{\left(\kappa\mu+\sum_{i=1}^{n_t} \bar{y}_i - h_i\right)^2}{\kappa+n_t}\right)$	$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C)$ $b \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\left(\sum_{i=1}^{n_t} \bar{y}_i - h_i\right)}{\kappa+n_t}, \lambda' = \lambda(\kappa+n_t)\right)$
	1	0	✓	$p(\lambda \mid \alpha, \beta) = \Gamma(\lambda \mid \alpha, \beta) =$ $= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$	$p(\mathcal{D} \mid \theta) = \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right)$ with $C := \beta + \frac{1}{2}\sum_{i=1}^{n_t} (\bar{y}_i - h_i)^2$	$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C)$
Multiplicative	✓	0	✓	$p(s_{\log}, \lambda \mid \mu, \kappa, \alpha, \beta) =$ $= \mathcal{N}(s_{\log} \mid \mu, (\lambda\kappa)^{-1}) \cdot \Gamma(\lambda \mid \alpha, \beta)$ $= \frac{\beta^\alpha\sqrt{\kappa}}{\Gamma(\alpha)\sqrt{2\pi}} \lambda^{\alpha-1/2} \exp\left(-\beta\lambda - \frac{\kappa\lambda(s_{\log}-\mu)^2}{2}\right)$	$p(\mathcal{D} \mid \theta) = \left(\prod_{i=1}^{n_t} \frac{1}{\bar{y}_i}\right) \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right) \cdot \sqrt{\frac{\kappa}{\kappa+n_t}}$ with $C := \beta + \frac{1}{2}\left(\kappa\mu^2 + \sum_{i=1}^{n_t} (\log(\bar{y}_i/h_i))^2 - \frac{\left(\kappa\mu+\sum_{i=1}^{n_t} \log(\bar{y}_i/h_i)\right)^2}{\kappa+n_t}\right)$	$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C)$ $s_{\log} \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\sum_{i=1}^{n_t} \log(\bar{y}_i/h_i)}{\kappa+n_t}, \lambda' = \lambda(\kappa+n_t)\right)$
	1	0	✓	$p(\lambda \mid \alpha, \beta) = \Gamma(\lambda \mid \alpha, \beta) =$ $= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$	$p(\mathcal{D} \mid \theta) = \left(\prod_{i=1}^{n_t} \frac{1}{\bar{y}_i}\right) \frac{(\beta/C)^\alpha}{\Gamma(\alpha)(2\pi C)^{n_t/2}} \cdot \Gamma\left(\alpha + \frac{n_t}{2}\right)$ with $C := \beta + \frac{1}{2}\sum_{i=1}^{n_t} (\log(\bar{y}_i) - \log(h_i))^2$	$\lambda \propto \mathrm{Gamma}(\alpha' = \alpha + \frac{n_t}{2}, \beta' = C)$

Table S2: **Overview of the marginalization-based approach applied to different observable combinations under experimentally measured additive and multiplicative Gaussian measurement noise.** Observation parameters considered are scaling factors (s) and offsets (b). The experimentally measured noise is denoted as precision $\bar{\lambda} := 1/\bar{\sigma}^2$. For multiplicative noise, the logarithm of the scaling factor ($s_{\log}$) is used. Unknown/estimated observation parameters are denoted by ✓, otherwise the fixed numerical value is shown. Further details for each case are in the *Supplementary Material*.

	s	***b***	**Prior distribution**	**Analytical solution**	**Conditional sampling**
Additive	✓	✓	$p(s, b \mid \nu, \tau, \mu, \kappa) =$ $= \mathcal{N}(s \mid \nu, \tau^{-1}) \cdot \mathcal{N}(b \mid \mu, \kappa^{-1})$ $= \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left(-\frac{\tau}{2}(s-\nu)^2\right)$ $\cdot \sqrt{\frac{\kappa}{2\pi}} \cdot \exp\left(-\frac{\kappa}{2}(b-\mu)^2\right)$	$p(\mathcal{D} \mid \theta) = \left(\frac{\bar{\lambda}}{2\pi}\right)^{n_t/2} \sqrt{\frac{\kappa\tau}{\bar{\lambda}\left(\left(\kappa+\bar{\lambda}n_t\right)\left(\frac{\tau}{\bar{\lambda}}+\sum_{i=1}^{n_t} h_i^2\right)-\bar{\lambda}\left(\sum_{i=1}^{n_t} h_i\right)^2\right)}}$ $\cdot \exp\left(-\frac{1}{2}\left(\kappa\mu^2 + \tau\nu^2 + \bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i^2 - \frac{\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i\right)^2}{\kappa+\bar{\lambda}n_t}\right.\right.$ $\left.\left. - \frac{\left(\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i\right)\left(\bar{\lambda}\sum_{i=1}^{n_t} h_i\right)-\left(\kappa+\bar{\lambda}n_t\right)\left(\tau\nu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i h_i\right)\right)^2}{\left(\kappa+\bar{\lambda}n_t\right)\left(\left(\kappa+\bar{\lambda}n_t\right)\left(\tau+\bar{\lambda}\sum_{i=1}^{n_t} h_i^2\right)-\left(\bar{\lambda}\sum_{i=1}^{n_t} h_i\right)^2\right)}\right)\right)$	$b \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t}(\bar{y}_i-h_i)}{\kappa+\bar{\lambda}n_t}, \lambda' = \kappa + \bar{\lambda}n_t\right)$ $s \propto \mathcal{N}\left(\mu' = \frac{\left(\kappa+\bar{\lambda}n_t\right)\left(\tau\nu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i h_i\right)-\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i\right)\left(\bar{\lambda}\sum_{i=1}^{n_t} h_i\right)}{\left(\kappa+\bar{\lambda}n_t\right)\left(\tau+\bar{\lambda}\sum_{i=1}^{n_t} h_i^2\right)-\left(\bar{\lambda}\sum_{i=1}^{n_t} h_i\right)^2},\right.$ $\left.\lambda' = \tau + \bar{\lambda}\sum_{i=1}^{n_t} h_i^2 - \frac{\left(\bar{\lambda}\sum_{i=1}^{n_t} h_i\right)^2}{\kappa+\bar{\lambda}n_t}\right)$
	✓	0	$p(s \mid \mu, \kappa) = \mathcal{N}(s \mid \mu, \kappa^{-1}) =$ $= \sqrt{\frac{\kappa}{2\pi}} \cdot \exp\left(-\frac{\kappa}{2}(s-\mu)^2\right)$	$p(\mathcal{D} \mid \theta) = \left(\frac{\bar{\lambda}}{2\pi}\right)^{n_t/2} \sqrt{\frac{\kappa}{\kappa+\bar{\lambda}\sum_{i=1}^{n_t} h_i^2}}$ $\cdot \exp\left(-\frac{1}{2}\left(\kappa\mu^2 + \bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i^2 - \frac{\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i h_i\right)^2}{\kappa+\bar{\lambda}\sum_{i=1}^{n_t} h_i^2}\right)\right)$	$s \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t} \bar{y}_i h_i}{\kappa+\bar{\lambda}\sum_{i=1}^{n_t} h_i^2}, \lambda' = \kappa + \bar{\lambda}\sum_{i=1}^{n_t} h_i^2\right)$
	1	✓	$p(b \mid \mu, \kappa) = \mathcal{N}(b \mid \mu, \kappa^{-1}) =$ $= \sqrt{\frac{\kappa}{2\pi}} \cdot \exp\left(-\frac{\kappa}{2}(b-\mu)^2\right)$	$p(\mathcal{D} \mid \theta) = \left(\frac{\bar{\lambda}}{2\pi}\right)^{n_t/2} \sqrt{\frac{\kappa}{\kappa+\bar{\lambda}n_t}}$ $\cdot \exp\left(-\frac{1}{2}\left(\kappa\mu^2 + \bar{\lambda}\sum_{i=1}^{n_t} (\bar{y}_i - h_i)^2 - \frac{\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t}(\bar{y}_i-h_i)\right)^2}{\kappa+\bar{\lambda}n_t}\right)\right)$	$b \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t}(\bar{y}_i-h_i)}{\kappa+\bar{\lambda}n_t}, \lambda' = \kappa + \bar{\lambda}n_t\right)$
Multiplicative	✓	0	$p(s_{\log} \mid \mu, \kappa) = \mathcal{N}(s_{\log} \mid \mu, \kappa^{-1}) =$ $= \sqrt{\frac{\kappa}{2\pi}} \cdot \exp\left(-\frac{\kappa}{2}(s_{\log}-\mu)^2\right)$	$p(\mathcal{D} \mid \theta) = \left(\frac{\bar{\lambda}}{2\pi}\right)^{n_t/2} \sqrt{\frac{\kappa}{\kappa+\bar{\lambda}n_t}} \left(\prod_{i=1}^{n_t} \frac{1}{\bar{y}_i}\right)$ $\cdot \exp\left(-\frac{1}{2}\left(\kappa\mu^2 + \bar{\lambda}\sum_{i=1}^{n_t} (\log(\bar{y}_i/h_i))^2 - \frac{\left(\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t}\log(\bar{y}_i/h_i)\right)^2}{\kappa+\bar{\lambda}n_t}\right)\right)$	$s_{\log} \propto \mathcal{N}\left(\mu' = \frac{\kappa\mu+\bar{\lambda}\sum_{i=1}^{n_t}\log(\bar{y}_i/h_i)}{\kappa+\bar{\lambda}n_t}, \lambda' = \kappa + \bar{\lambda}n_t\right)$

References

[1] Kitano, H. Systems biology: A brief overview. *Science* **295**, 1662–1664 (2002).

[2] Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. *Systems biology in practice* (Wiley-VCH, Weinheim, 2005).

[3] Schöberl, B. *et al.* Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Science Signaling* **2**, ra31 (2009).

[4] Fey, D. *et al.* Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.* **8** (2015).

[5] Hass, H. *et al.* Predicting ligand-dependent tumors from multi-dimensional signaling features. *npj Syst Biol Appl* **3**, 27 (2017).

[6] Giordano, G. *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med* **26**, 855–860 (2020).

[7] Zhao, S. & Chen, H. Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant Biol* **8**, 11–19 (2020).

[8] Renart, J., Reiser, J. & Stark, G. R. Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: A method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. USA* **76**, 3116–3120 (1979).

[9] Sanderson, M. J., Smith, I., Parker, I. & Bootman, M. D. Fluorescence Microscopy. *Cold Spring Harb. Protoc.* **2014** (2014). URL `https://doi.org/10.1101/pdb.top071795`.

[10] Blasi, T. *et al.* Combinatorial histone acetylation patterns are generated by motif-specific reactions. *Cell Systems* **2**, 49–58 (2016).

[11] Kreutz, C. *et al.* An error model for protein quantification. *Bioinformatics* **23**, 2747–2753 (2007).

[12] Raue, A. *et al.* Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* **8**, e74335 (2013).

[13] Degasperi, A., Fey, D. & Kholodenko, B. N. Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *npj Syst Biol Appl* **3**, 20 (2017).

[14] Weber, P., Hasenauer, J., Allgöwer, F. & Radde, N. Parameter estimation and identifiability of biological networks using relative data. In Bittanti, S., Cenedese, A. & Zampieri, S. (eds.) *Proc. of the 18th IFAC World Congress*, vol. 18, 11648–11653 (Milano, Italy, 2011).

[15] Xu, T.-R. *et al.* Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.* **3**, ra20 (2010).

[16] Raue, A., Kreutz, C., Theis, F. J. & Timmer, J. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philos T Roy Soc A* **371** (2013).

[17] Hug, S. *et al.* High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Math. Biosci.* **246**, 293–304 (2013).

[18] Haario, H., Saksman, E. & Tamminen, J. An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242 (2001).

[19] Graham, M. M. & Storkey, A. J. Continuously tempered hamiltonian monte carlo. In *UAI* (2017).

[20] Hoffman, M. D. & Gelman, A. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623 (2014).

[21] Lacki, M. K. & Miasojedow, B. State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Stat. Comput.* **26**, 951–964 (2015).

[22] Bellman, R. E. *Adaptive Control Processes* (Princeton University Press, 1961). URL `https://doi.org/10.1515/9781400874668`.

[23] Taylor, A. N. & Kitching, T. D. Analytic methods for cosmological likelihoods. *Monthly Notices of the Royal Astronomical Society* **408**, 865–875 (2010). URL `https://doi.org/10.1111/j.1365-2966.2010.17201.x`.

[24] Loos, C., Krause, S. & Hasenauer, J. Hierarchical optimization for the efficient parametrization of ODE models. *Bioinf.* **34**, 4266–4273 (2018).

[25] Schmiester, L., Schälte, Y., Fröhlich, F., Hasenauer, J. & Weindl, D. Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics* **36**, 594–602 (2019). URL `https://doi.org/10.1093/bioinformatics/btz581`. `https://academic.oup.com/bioinformatics/article-pdf/36/2/594/31962762/btz581.pdf`.

[26] Bachmann, J. *et al.* Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.* **7**, 516 (2011).

[27] Raimúndez, E. *et al.* Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines. *PLoS Comput. Biol.* **16**, e1007147 (2020).

[28] Maier, C., Loos, C. & Hasenauer, J. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics* **33**, 718–725 (2017).

[29] Fujita, K. A. *et al.* Decoupling of receptor and downstream signals in the Akt pathway by its low-pass filter characteristics. *Sci Signal* **3**, ra56 (2010).

[30] Boehm, M. E. *et al.* Identification of isoform-specific dynamics in phosphorylation-dependent STAT5 dimerization by quantitative mass spectrometry and mathematical modeling. *Journal of Proteome Research* **13**, 5685–5694 (2014).

[31] Leonhardt, C. *et al.* Single-cell mRNA transfection studies: Delivery, kinetics and statistics by numbers. *Nanomedicine: Nanotechnology, Biology, and Medicine* **10**, 679–688 (2014).

[32] Hasenauer, J., Hasenauer, C., Hucho, T. & Theis, F. J. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.* **10**, e1003686 (2014).

[33] Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J. M., Smith, A. F. M., Dawid, A. P. & Berger, J. O. (eds.) *Bayesian Statistics*, vol. 4, 169–193 (University Press Oxford, 1992).

[34] Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points **113**, 9978–9985 (2000). URL `https://doi.org/10.1063/1.1323224`.

[35] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018). URL `https://doi.org/10.21105/joss.00861`.

[36] Girolami, M. & Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B* **73**, 123–214 (2011).

[37] Paszke, A. *et al.* Automatic differentiation in PyTorch (2017).

[38] Villaverde, A. F., Raimúndez, E., Hasenauer, J. & Banga, J. R. A comparison of methods for quantifying prediction uncertainty in systems biology. *IFAC-PapersOnLine* (2019).

[39] Sisson, S. A., Fan, Y. & Tanaka, M. M. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.* **104**, 1760–1765 (2007).

[40] Toni, T., Ozaki, Y.-i., Kirk, P., Kuroda, S. & Stumpf, M. P. H. Elucidating the in vivo phosphorylation dynamics of the ERK MAP kinase using quantitative proteomics data and Bayesian model selection. *Mol. Biosyst.* **8**, 1921–1929 (2012).

[41] Schälte, Y. & Hasenauer, J. Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. *Bioinformatics* **36**, i551–i559 (2020).

[42] Villaverde, A. F., Froehlich, F., Weindl, D., Hasenauer, J. & Banga, J. R. Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics* bty736 (2018).

[43] Fröhlich, F. & Sorger, P. K. Fides: Reliable trust-region optimization for parameter estimation of ordinary differential equation models. *PLoS Comput. Biol.* **18**, e1010322 (2022). URL `https://doi.org/10.1371/journal.pcbi.1010322`.

[44] Hass, H. *et al.* Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics* **35**, 3073–3082 (2019). `https://academic.oup.com/bioinformatics/article-pdf/35/17/3073/29591749/btz020.pdf`.

[45] Kreutz, C. New concepts for evaluating the performance of computational methods. *IFAC-PapersOnLine* **49**, 63–70 (2016).

[46] Ballnus, B. *et al.* Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst Biol* **11**, 63 (2017).

[47] Vousden, W., Farr, W. & Mandel, I. Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Mon. Not. R. Astron. Soc.* **455**, 1919–1937 (2016).

[48] Miasojedow, B., Moulines, E. & Vihola, M. An adaptive parallel tempering algorithm. *J. Comput. Graph. Stat.* **22**, 649–664 (2013).

[49] Schälte, Y. *et al.* pyPESTO - Parameter EStimation TOolbox for python (2021). URL `https://zenodo.org/record/2553546`.

[50] Haario, H., Laine, M., Mira, A. & Saksman, E. DRAM: Efficient adaptive MCMC. *Stat. Comp.* **16**, 339–354 (2006).

[51] Schmiester, L. *et al.* PEtab—interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology* **17**, 1–10 (2021).

[52] Fröhlich, F. *et al.* AMICI: High-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics* **btab227** (2021). Btab227, `https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btab227/37371345/btab227.pdf`.