High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling

S. Hug [a,b,1], A. Raue [a,c,*,1], J. Hasenauer [a], J. Bachmann [d], U. Klingmüller [d], J. Timmer [c,e,f], F.J. Theis [a,b]

[a] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany
[b] Department of Mathematics, Technische Universität München, Germany
[c] Institute for Physics, University of Freiburg, Germany
[d] Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany
[e] BIOSS Centre for Biological Signalling Studies and Freiburg Institute for Advanced Studies (FRIAS), Freiburg, Germany
[f] Department of Clinical and Experimental Medicine, Linköping University, Sweden

1. Introduction

Quantitative mathematical models can be used to describe the dynamics of cellular processes such as signal transduction. The mathematical description facilitates the understanding of complex, intertwined responses of the underlying networks of molecular reactions. The models can be employed to predict and understand features of the processes in a quantitative manner.

For building and calibrating dynamical models prior knowledge from the literature as well as quantitative experimental data is used. However, both, prior knowledge and experimental data, are limited and usually come with an associated uncertainty. These limitations and uncertainties translate to uncertainties in the mathematical model and subsequently to uncertainties of the predictions. To assess the predictive power of a model as well as its limitations, the model uncertainties have to be evaluated. As the models are often high dimensional and possess a large number of unknown parameters, this uncertainty evaluation can pose severe computational challenges. In this work we illustrate that a rigorous statistical assessment is also feasible for nonlinear high-dimensional dynamical models with over 100 parameters. For this we consider Epo-induced JAK2/STAT5 signaling, a process which has been studied extensively in recent years.

The hormone Erythropoietin (Epo) regulates erythropoesis, the production of red blood cells. Binding of Epo to its cognate receptor leads to rapid activation of JAK2 phosphorylation followed by phosphorylation of the latent transcription factor STAT5, see Fig. 1 for illustration of the model. The quantitative link between the integral STAT5 response in the nucleus and survival of erythroid progenitor cells has recently been elucidated [1]. The broad dynamical range of Epo concentrations up to 1000-fold in vivo [2] require a stringent regulatory system. In [1], it was shown that STAT5 responses are controlled by a dual feedback consisting of two inhibitory proteins, CIS and SOCS3. The two proteins adjust STAT5 phosphorylation levels over the entire range of Epo concentrations, where CIS regulates predominantly the upper concentrations range and SOCS3 the lower range. Model predictions showed that the absence (knock-out) of CIS resulted in an increase

* Corresponding author at: Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany.
E-mail addresses: andreas.raue@fdm.uni-freiburg.de (A. Raue), fabian.theis@helmholtz-muenchen.de (F.J. Theis).
[1] These authors contributed equally.

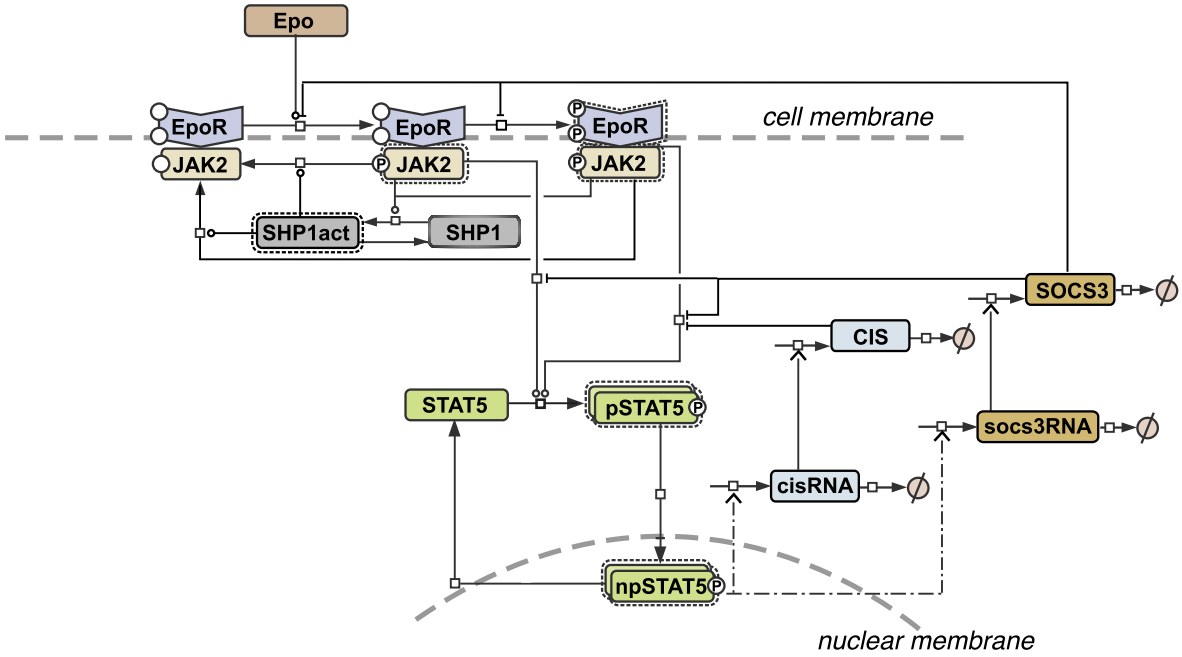

Fig. 1. Dynamical model of the Epo induced JAK2/STAT5 signal transduction pathway, adopted from [1]. The hormone Erythropoietin (Epo) binds to its membrane receptor (EpoR) and subsequently leads to receptor phosphorylation (pEpoR) and to phosphorylation of its associated Janus kinase (JAK2, pJAK2). Receptor phosphorylation is balanced by activation of a phosphatase (SHP1, SHP1act). Active EpoR/JAK2 complexes lead to phosphorylation of the Signal Transducer and Activator of Transcription (STAT5, pSTAT5) that transmits the signal to the nucleus (npSTAT5). In the nucleus, STAT5 leads to target gene expression that induces pro-survival signals and self-regulating negative feedbacks. In this case, two regulator proteins and their respective mRNAs are involved, Suppressor Of Cytokine Signaling (SOCS3) and the Cytokine-Inducible SH2-containing protein (CIS).

of STAT5 phosphorylation at low Epo concentrations, whereas the absence of SOCS3 caused an increase in the phosphorylation level at high Epo concentrations. This observation revealed division of labor by the two feedback proteins as the key property to control STAT5 responses.

In this work we analyze the Epo-induced JAK2/STAT5 signaling using a combination of maximum likelihood based parameter estimation for model calibration, identifiability analysis using the profile likelihood or profile posterior approach, and Bayesian inference using Markov chain Monte Carlo sampling (MCMC) for the translation of uncertainties to model parameters and to model predictions. Compared to the original publication [1], we carry out a refined analysis of parameter and prediction uncertainty by evaluating and interpreting posterior distributions using a Bayesian approach. The results show a second mode in the posterior that corresponds to an alternative parameterization of the model. Furthermore, the Bayesian approach reveals non-standard parameter dependency structures. The combination of high-dimensional parameter space and bimodal posterior shape represents a difficult setting for MCMC sampling. Our results indicate deficiencies of single-chain sampling schemes occurring for this high-dimensional problem with bimodal and partially flat posterior distribution. To ensure good performance for this challenging setting we instead apply a multi-chain scheme, namely Parallel Hierarchical Sampling in combination with Adaptive Metropolis Sampling, which provides improved mixing properties. Using this MCMC scheme, which is of general relevance, we can efficiently sample multimodal posterior distributions. The sampling results enable us to study the effects of the additional posterior mode on the model predictions and propose additional experiments that would allow to distinguish between both modes.

2. Methods

In this section, we present an overview of the mathematical methods applied to our biological question. First of all, this refers to modeling biological pathways as dynamical systems with ordinary differential equations. As these systems are determined by their parameters, we next present Bayesian inference as a means to inferring these parameters given the measurement data. We will then shortly introduce the profile posterior approach for identifiability analysis, which we use as a basis and complement to the Markov chain Monte Carlo (MCMC) approach. In the last section of the methods part, several of these MCMC sampling techniques which are used to sample from the complex distributions of the parameters are presented. Finally, we describe how to use the obtained samples to gain insight into the model.

2.1. Dynamical systems

Understanding cellular mechanisms has always been a key challenge of systems biology. Much effort has gone into the inference of biochemical reaction networks, such as signaling pathways, which are the main focus of this paper. All of these biochemical processes may be described by systems of biochemical reactions, where reactants are transformed into products [3]. There are various approaches for modeling the evolution of such a system over time, most common is the modeling as a dynamical system described by N ordinary differential equations (ODEs), one for each of the N modeled species, cf. e.g. [4–9]. They are characterized by a functional relationship between the current state of the system $\mathbf{x}(t) \in \mathbb{R}^N$ at time point t and its time derivative $\dot{\mathbf{x}}(t)$:

$$\dot{\mathbf{x}}(t) = f(t, \mathbf{x}(t), \mathbf{u}(t), \zeta). \tag{1}$$

This may depend on an external stimulus $\mathbf{u}(t) \in \mathbb{R}^L$ such as adding a biochemical species whose time course is not included in the model (in our case, this is e.g. stimulation of the pathway by adding Epo) as well as the dynamical parameters ζ, which are in this case e.g. the rate constants of the biochemical reactions [10]. In this approach, the dynamical parameters ζ do not depend on t and are thus constant over time. Usually, not all states of the system can actually be directly measured, so that there exists a mapping g of the

internal states $\mathbf{x}(t)$ to the external states $\mathbf{y}(t) \in \mathbb{R}^M$, sometimes also called the observables of the system:

$$\mathbf{y}(t) = g(\mathbf{x}(t), \boldsymbol{\eta}), \tag{2}$$

where $\boldsymbol{\eta}$ are scaling and offset parameters which are often due to relative measurements, e.g. when using immunoblots. In biological systems, it is most common to have $M < N$. Observations of the system at discrete time points $t_j, j = 1, \ldots, J$, i.e. the measurement data, are overlaid with noise $\boldsymbol{\epsilon}$, which we take to be independent normally distributed, $\boldsymbol{\epsilon}(t_j) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, such that we have:

$$\tilde{\mathbf{y}}_j = \mathbf{y}(t_j) + \boldsymbol{\epsilon}(t_j). \tag{3}$$

Usually, the positive definite noise covariance matrix $\boldsymbol{\Sigma}$ is assumed to be diagonal, such that noise on the observables is independent, and we only need to estimate the diagonal elements $\sigma_1, \ldots, \sigma_M$ of $\boldsymbol{\Sigma}$. The noise parameters describe the deviations of the measured data from the actual dynamics of the ODEs, which are due to biological variability as well as to measurement errors. Setting $\boldsymbol{\theta} = \{\zeta, \boldsymbol{\eta}, \boldsymbol{\Sigma}, \mathbf{x}(0)\}$, where $\mathbf{x}(0)$ are the initial conditions of the system, the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^K$ completely determines the model, while the subset $\{\zeta, \boldsymbol{\eta}, \mathbf{x}(0)\}$ is sufficient to uniquely identify the dynamics of the system. We use logarithmic parameter values, mostly in order to be able to apply global optimization algorithms. Otherwise we would have to use constrained optimization, since all original parameter values are only defined on $\mathbb{R}_+$. Furthermore, also the data $\mathbf{Y}$ we base our inference on is the logarithm of the actually measured data, since it is commonly believed that biochemical reactions mostly show lognormally distributed multiplicative noise, so that the logarithm of the data shows normally distributed additive noise [11]. Neither of these two transformations changes the general dynamics of the system.

For evaluating the deviations of the measurements from the dynamics of the system, a nonlinear system of ODEs has to be solved. In most cases, finding an analytical solution is impossible, therefore the system has to be solved numerically. Some care has to be taken when choosing a solver, since many ODE systems show stiffness or other numerical issues. Here, SUNDIALS CVODEs [12] algorithm was used.

2.2. Bayesian inference

Bayesian inference is a tool for examining complex systems [3,8,13] due to its ability to combine measurement uncertainties for the experimental data with previously available prior information for the parameters. These systems are often highly nonlinear, yet they are deterministic, i.e. they are completely characterized by their parameters. Unlike physical constants, it is however not possible to compute biological parameters "once and for all". Furthermore, these parameters are not well determined due to limited knowledge [5] about e.g. the correct model structure or interaction mechanisms. The uncertainties in the parameters have to be taken into account when making any predictions or drawing any conclusions from the model because otherwise these might not be properly justified. Bayesian inference incorporates these uncertainties in a natural way and estimates all parameters of the system, whether dynamical, initial condition, scaling or noise parameter, in a joint fashion.

The Bayesian framework provides a fully probabilistic approach which builds on the so called posterior distribution [14] of a problem specific parameter space conditioned on the given experimental data. This distribution specifies a measure of belief for all possible parameter values, taking into account both how well they explain the data and how well they match the existing knowledge. This can be formalized in Bayes' theorem, which states that

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) p(\boldsymbol{\theta})}{p(\mathbf{Y})}. \tag{4}$$

Depending on the parameters $\boldsymbol{\theta}$ and the data $\mathbf{Y}$, this theorem links the four essential quantities of Bayesian inference to each other, which are:

- $p(\boldsymbol{\theta}|\mathbf{Y})$, the posterior density of the parameters and thus the probability density function of the parameters given the data,
- $\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = p(\mathbf{Y}|\boldsymbol{\theta})$, the likelihood, i.e. the conditional probability of the data $\mathbf{Y}$ given the parameter $\boldsymbol{\theta}$,
- $p(\boldsymbol{\theta})$, the **prior** of the parameter,
- $p(\mathbf{Y})$, the marginal likelihood or evidence for the data, which is of special importance when doing model selection.

The likelihood can also be seen as a cost function, as it punishes deviations of the model from the data. Higher likelihood for the parameters implies smaller deviations between the time course and the data. The likelihood as such is a straightforward consequence of the error model, so if we assume as before that the noise follows a normal distribution, then the likelihood can be written down explicitly as

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{j=1}^{J} \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{1}{2\sigma_m^2} \left(y_m(t_j) - \tilde{y}_{m,j} \right)^2 \right), \tag{5}$$

where $y_m(t_j)$ and $\tilde{y}_{m,j}$ are the m-th component of $\mathbf{y}(t_j)$ and $\tilde{\mathbf{y}}_j$, respectively. Often it is computationally easier to work with the natural logarithm of the likelihood, also called log-likelihood instead:

$$\log\left(\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})\right) = -\frac{J}{2} \sum_{m=1}^{M} \log\left(2\pi\sigma_m^2\right) - \frac{1}{2} \sum_{j=1}^{J} \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \left(y_m(t_j) - \tilde{y}_{m,j} \right)^2. \tag{6}$$

For known $\boldsymbol{\Sigma}$, maximizing the likelihood is the same as finding the least squares fit to the data. The parameter set for which the likelihood (or equivalently the log-likelihood) is maximal is called the maximum likelihood estimate (MLE) of the parameters. In the Bayesian sense, an often used point estimate is the maximum a posteriori (MAP) estimate, i.e. the parameter value for which the posterior density is maximal. However, both these point estimates might be unrepresentative, especially if the likelihood and thus the posterior distribution are multimodal.

The prior specifies our belief in a parameter vector before observing the data $\mathbf{Y}$. Often values already known from independent data in the literature can be used as prior information, for example by taking them as a mean for a normal distribution of this parameter. If the prior is normalized to be a probability density function, it is called proper, otherwise improper. Since usually parameter dependency structures are not known a priori, it is most common to set the joint prior $p(\boldsymbol{\theta})$ to the product of all the individual $p(\theta_k), k = 1, \ldots, K$.

2.3. Identifiability and profile posteriors

A related concept to Bayesian inference, yet different, is that of profile posteriors for identifiability analysis [15,16]. The profile posterior approach is analog to the profile likelihood approach [17]. Both approaches are equivalent if the prior is flat. Note that in this case, the prior is improper. In the case of non-flat priors, the comparison of profile likelihood and profile posterior can also be used to assess the information content of the data with respect to the parameters. Furthermore, this comparison reveals if identifiability is only enforced by the prior distribution. This comparison of the two profile types is thus related to a prior/posterior evaluation in a Bayesian setting.

The profile posterior allows for the calculation of confidence intervals which can be compared to the Bayesian credible intervals obtained from MCMC sampling. Also the histogram of the marginalized samples can very well be compared with the posterior profiles. Since profile posteriors are computationally less expensive, they form an optimal basis for any MCMC procedure for Bayesian inference and thus complement these for gaining thorough insight into the dynamical system at question.

Structural and practical non-identifiability can be distinguished and can have several reasons [17]. Structural non-identifiability is independent of the measurement data and instead due to some underlying fundamental redundancy in the parametrization of the model. Practical non-identifiability depends on the amount and quality of the data. Identifiability analysis is usually based on the maximum likelihood or, in our case, maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$ of the parameters. A parameter θ_k is identifiable if the confidence interval $[\rho_k^-, \rho_k^+]$ of its estimate $\hat{\theta}_k$ is finite. These confidence intervals are computed from the posterior distribution.

The basic idea of the profile posterior approach is to explore the parameter space for each parameter separately in direction of the least decrease of the posterior $p(\boldsymbol{\theta}|\mathbf{Y})$. This can be done by calculating the profile posterior

$$p_{PP}(\theta_k) = \max_{\theta_{j\neq k}}[p(\boldsymbol{\theta}|\mathbf{Y})], \tag{7}$$

re-optimizing the posterior with respect to all parameters $\theta_{j\neq k}$ except θ_k, for each value of θ_k. The presence of local optima, i.e. multiple modes in the posterior, can be detected by repeated optimization runs from different starting values. If such local optima are detected, the profile calculation has to be initiated in each of the optima. Note that the calculation of the profiles for different parameters can be performed independently and simultaneously on different computer cores. For more details on the implementation, see [17]. A generalization of this approach to model predictions by calculating prediction profiles was proposed in [18].

2.4. Markov chain Monte Carlo methods

On first glance, it is straightforward to draw conclusions from the posterior by simply applying Bayes' theorem. However, the marginal likelihood in the denominator is actually very hard to evaluate, since it represents the integral of the numerator:

$$p(\mathbf{Y}) = \int_{\mathbb{R}^K} \mathcal{L}(\boldsymbol{\theta}';\mathbf{Y})p(\boldsymbol{\theta}')d\boldsymbol{\theta}' \tag{8}$$

As this integral is taken over the whole parameter space, it is usually high-dimensional and analytically mostly intractable and also daunting to evaluate numerically. Instead, the methods of choice for inferring the posterior distribution are in many cases Markov chain Monte Carlo (MCMC) methods, since they only need the non-normalized numerator in Bayes' theorem which is the product of likelihood and prior as input. The marginal likelihood is needed explicitly for example for model selection, where it can be used for computing Bayes factors for competing models [19–25]. For the theory of Markov chains and Monte Carlo methods we refer the interested reader to the abundant literature [24,26–28], for example also the excellent books of Marin & Robert [29], or Robert & Casella [30].

The general idea is to approximate the posterior distribution at hand with a Markov chain $\boldsymbol{\theta}^{(i)}, 1 = 1, \ldots, I$, whose stationary distribution is the posterior distribution. The elements of the Markov chain are samples from the parameter space. We now restrict ourselves to introducing some algorithms, first of all the Metropolis–Hastings (MH) algorithm [31–33] as the basis for many more involved samplers. The MH algorithm makes use of a proposal distribution $q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(i)})$ to suggest moves of the Markov chain $\boldsymbol{\theta}^{(i)}, 1 = 1, \ldots, I$, from its current state $\boldsymbol{\theta}^{(i)}$ to a proposed next state $\boldsymbol{\theta}^{(p)}$ to draw samples from, in our case, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$. This move is accepted with the Metropolis–Hastings acceptance probability

$$\alpha\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(p)}\right) = \min\left\{1, \frac{p(\boldsymbol{\theta}^{(p)}|\mathbf{Y})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{Y})}\frac{q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(p)})}{q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(i)})}\right\}. \tag{9}$$

If $\boldsymbol{\theta}^{(p)}$ is accepted, we set the next state of the Markov chain to $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(p)}$, otherwise we set it to be the old state, $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$. The distribution of the samples created in this fashion then converges to the posterior distribution. Like the sampling distribution $p(\cdot|\mathbf{Y})$, also the proposal distribution $q(\cdot|\cdot)$ only has to be known up to a multiplicative constant, since these cancel in the Metropolis–Hastings acceptance probability. For the proposal distribution, a popular choice is a multivariate Gaussian distribution with mean at the current state of the Markov chain and to-be-determined covariance matrix $\boldsymbol{\Sigma}_q, \boldsymbol{\theta}^{(p)} \sim \mathcal{N}\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\Sigma}_q\right)$. We call the MH algorithm then a Random Walk Metropolis–Hastings. However, tuning the proposal distribution is the hardest task when implementing an MH algorithm [34,35]. If the moves of the chain are too small, it takes impractically long to explore the whole parameter space and converge to the stationary distribution. Furthermore, the samples in the Markov chain show high autocorrelation, which is a problem if independent samples are required. If the moves are however too large, usually the acceptance rate suffers since more often moves are proposed that do not fall into regions of high posterior density and the Markov chain might get "stuck" for a large number of iterations, which of course is also disadvantageous for the general autocorrelation structure of the Markov chain. In the last few years, a large number of more sophisticated MCMC algorithms has been published. Among them are more geometric methods like Riemann manifold Langevin and Hamiltonian Monte Carlo methods [36] or copula-based methods like the Copula Independence Metropolis Hastings algorithm [37]. For a more thorough overview, we refer the interested reader to the contribution by Vanlier and van Riel [38]. However, in high-dimensional systems, it is our experience that "simpler" sampling algorithms are often more efficient.

The single-chain algorithm used in this paper is a MATLAB implementation of the Adaptive Metropolis (AM) algorithm as introduced by Haario et al. [39]. It is especially suitable for sampling high-dimensional parameter spaces since its proposal function can be continuously adapted to guarantee efficient sampling of the high-dimensional space. This is a crucial factor for the convergence of the algorithm. In the AM algorithm, the proposal function is a multivariate normal distribution whose covariance matrix is updated with the information gained from the obtained samples by a recursion formula. Newly proposed samples are then accepted or rejected according to a standard Metropolis–Hastings acceptance scheme. This sampling process is strictly speaking non-Markovian as the samples depend on the past of the sampling procedure and not just on their immediate predecessor, however Haario et al. show that the algorithm has the correct ergodic properties and thus converges correctly to the posterior distribution.

We combined this method with a sampling algorithm using multiple chains [40] in order to have better mixing properties of the chain. This is necessary because otherwise one chain would have to be run for an impractically long time in order to be sure to have captured the entire mass of the posterior, in particular when considering high-dimensional parameter spaces of over 100 parameters as in our case.

Different varieties of multi-chain methods have been proposed such as parallel tempering [41], exchange Monte Carlo [42] or population-based reversible jump MCMC [43]. While these methods are also advocated for closely related problems, we applied Parallel Hierarchical Sampling (PHS) adapted from Rigat & Mira [44], see

also references therein for a more thorough introduction. In this sampling scheme, several MCMC chains are run in parallel, with one chain being the mother chain and the others being auxiliary chains. As opposed to a parallel tempering approach, these chains all target the posterior distribution and are distinguished by their starting points and proposal distributions.

In each iteration, an index is randomly chosen, and the current state of this chosen auxiliary chain is swapped with the mother chain. This swap is always accepted. All the other auxiliary chains do a regular Metropolis update step, in our special case with the Adaptive Metropolis. Each auxiliary chain runs its own Adaptive Metropolis, such that the covariance matrix for the proposal distribution is adapted to each chain individually. This ensures a good sampling performance in each chain, while at the same time having excellent mixing properties for the mother chain. We will refer to our algorithm as Adaptive Metropolis Parallel Hierarchical Sampling (AMPHS). We chose this approach since it is especially beneficial when the posterior is multimodal or more generally non-standard shaped, as it increases mixing between the modes, even though that naturally comes at higher computational costs. Furthermore, the use of several different proposal settings is especially beneficial when it is difficult or impossible to analytically derive an optimal proposal scaling. As pointed out by Rigat & Mira, a single proposal kernel may for example not be optimal for exploring a distribution with both very narrow and very wide peaks.

We started the sampling procedure for the mother chain in the MAP estimate obtained by an optimization procedure which is described in more details in Bachmann et al. [1]. For the auxiliary chains, we sampled initial values randomly from the prior, and then let an optimization algorithm run in order to start in a region with substantial posterior values. Note that this is related to the multi-start optimization strategy for the profile calculation. As pointed out by Geyer [45], thinning the chain would increase the variance, thus we use all samples from the mother chain, except for a burn-in period. This burn-in period was chosen through visual inspection of the chain and verified by the Geweke test [46]. A burn-in period is necessary even though we start in the MAP estimate, i.e. a region of high posterior density, since the MAP estimate lies at the boundary of the prior parameter space. Likewise, the Adaptive Metropolis needs quite a number of iterations for tuning the proposal distribution. The algorithm is furthermore implemented in such a way that it also adaptively tunes the acceptance probability to be within a desirable range.

From the obtained posterior samples, it is possible to derive credible intervals, the Bayesian equivalent to confidence intervals. Even more interesting, it is possible to approximate the distribution of the single parameters by marginalizing over the samples. This translates to a density for the time courses of the dynamical system by solving the ODEs for each accepted sample and then estimating the density on a grid of time points. Often a range of possible dynamics is revealed. These predictions may be experimentally verified, since they now relate directly to the modeled species and not to the parameters by themselves.

3. Results

In this section, we present our results for the high-dimensional model of Epo-induced JAK2/STAT5 signaling. Using this application we illustrate problems which can arise when studying high-dimensional models and outline potential solutions. In this case study 113 parameter have to inferred, 27 parameters of interest which are the dynamical parameters and initial conditions and 86 nuisance parameters. The parameters of interest determine the model predictions, while the nuisance parameters have to be estimated to compare model observables to the experimental data. Furthermore, we illustrate how profile posterior methods can be employed to check the convergence of the MCMC sampling scheme. In particular if the profile posterior shows multiple modes, classical convergence test such as the Geweke criterion can be misleading concerning the convergence properties of the MCMC chain. Based upon the results of the profile posterior analysis, improved MCMC sampling schemes were selected, namely the previously introduced AMPHS scheme, that are necessary for efficient sampling of the high-dimensional and nonlinear posterior in this application. We compare our results to earlier contributions. Furthermore, we first motivate the need for a multi-chain approach by showcasing the issues of a single-chain approach, before providing a detailed analysis of the multi-chain sampling results. Afterwards, we show how the samples are used for predictions that can be made from the model.

3.1. Profile posterior analysis

To evaluate the identifiability of the individual parameters we perform at first a profile posterior analysis. The results are depicted in Fig. 2 and 3. Indeed, most parameters are well determined, but there are also a few which are practically non-identifiable, e.g. CISRNATurn and SOCS3Turn. A closer inspection of the profile posteriors reveals that two parameters, namely SOCS3RNADelay and SOCS3RNATurn, do exhibit a secondary mode, see in Fig. 2. The higher mode of these is the MAP estimate found by optimization, while the secondary mode is close to the threshold that defines a 95% confidence region [17].

3.2. Single-chain sampling and its limitations

To sample the posterior distribution of the Epo-induced JAK2/STAT5 signaling pathway we first employed a single-chain method. In particular we started with Adaptive Metropolis (AM) sampling [39]. When initializing the AM at the MAP estimate, we found that the MCMC chain converged according to the commonly used Geweke test after 500.000 samples (100.000 burn-in and 400.000 retained samples).

To validate the sampling result, we started a second AM chain in the secondary mode detected using the profile posterior method, as explained above. It turned out that also this second AM run seems to converge after 500.000 samples, according to the Geweke test. However, the sample distributions of the two runs differ severely. The difference is particularly pronounced in the parameters SOCS3RNADelay and SOCS3RNATurn for which we observe the bimodality in the posterior profiles. This indicates that the individual chains indeed sufficiently sample the modes in which they were started but failed to cover the bimodality of the posterior, cf. exemplary Fig. 6 in the appendix. We further confirmed the non-convergence with the Gelman–Rubin statistics [40]. For the two aforementioned parameters, the value of $\widehat{R}$ were 2.28 and 2.65 respectively, indicating that the two chains were not sampling from the same distribution.

To unravel the source of the convergence problems, we analyzed the distribution of the MCMC samples obtained when starting the chains in the two different modes. In particular we studied whether or not the MCMC samples from the two chains are non-overlapping. This would indicate that not only the posterior profiles are bimodal but also that the corresponding modes of the posterior distribution are separated.

While this is already visible from the marginalized one-dimensional samples, we quantified the overlap of the samples with support vector machines (SVMs) [47], see C. The SVM allows us to assign the samples to the two modes in the high dimensional space and thus visualize them accordingly, which will be shown in Fig. 3.

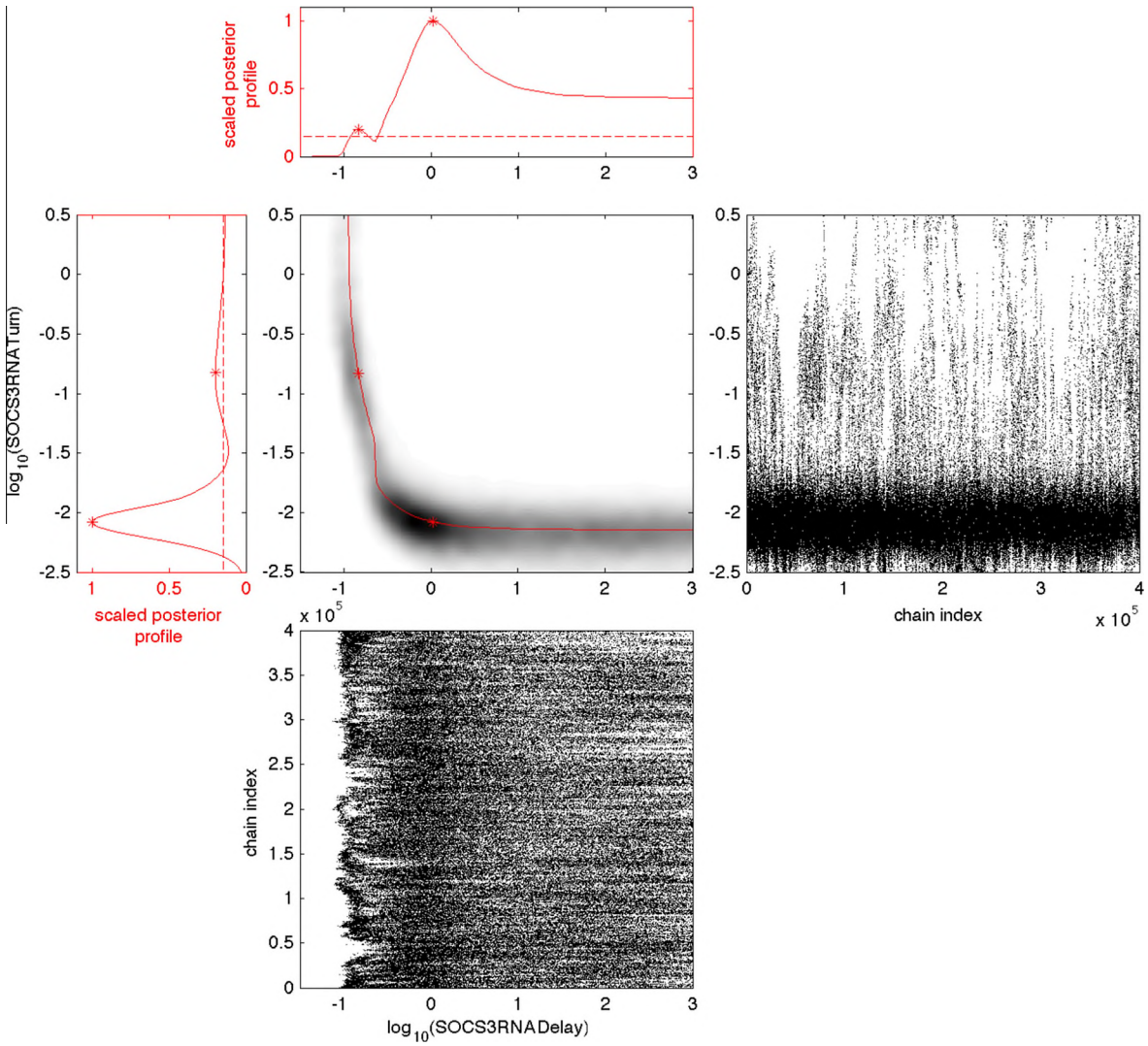

Fig. 2. MCMC chains and profiles for SOCS3RNADelay and SOCS3RNATurn. The middle panel shows posterior density over the two parameters in grey. The red line is the two-dimensional profile, the red stars the two modes. The left and top panels show the one-dimensional profiles, while the bottom and right panels show the MCMC chain for the two parameters. All panels imply a separation of the two modes in the parameter space, although they are connected by a banana-shaped ridge of high posterior density. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Our findings raise doubts concerning the sampling performances in high-dimensional parameter spaces. Although the single adaptive MCMC chains achieved a good sampling performance within the modes and although the modes are connected, the sampling of the true posterior distribution is very inefficient. To improve upon this and to ensure good mixing of the chain, we applied the Adaptive Metropolis Parallel Hierarchical Sampling.

3.3. Multi-chain sampling

For the AMPHS we used 20 auxiliary chains, each with 500.000 samples. The AMPHS was initialized with the mother chain and ten auxiliary chains in the MAP estimate, while the other ten auxiliary chains were initialized in local optima, also some who are close to the secondary mode found in the profile posterior. After visual inspection of the mother chain for the dynamical parameters, we set the burn-in period to be the first 100.000 samples, so that the further evaluation could be based on 400.000 samples. Convergence of the chain was again verified by the Geweke test. Although the test criterion is fulfilled, that alone does not ensure convergence of the sampling procedure. However, convergence is supported by the good agreement of marginal distributions and the profile posterior. Note that the Gelman–Rubin statistic is not easily applicable to the outcome of a single run of AMPHS due to the specific structure of the chains.

By analyzing the mother chain we found that mixing is much enhanced in this algorithm, as clearly the mother chain mixes very well between the two modes. Fig. 2 depicts the sampling results for the two parameters SOCS3RNADelay and SOCS3RNATurn, which were chosen here because of the bimodality expected from the profile posterior. The bottom and right panels clearly show that the chain mixes very well in the single dimensions. When looking at the samples in two dimensions, the middle panel indicates that they visit both modes, although the main mode obviously has more weight. AMPHS correctly estimates the weight assigned to each mode, see also Rigat & Mira [44] for additional examples. This can be observed from the fact that the initialization was in a weighting of 50% of the auxiliary chains near the main mode and 50% near the secondary mode. Using the SVM trained from the single chains (Section 3.2), we found that about 84% of the final samples in the mother chain belong to the main mode around the MAP estimate and 16% of samples are classified as belonging to the secondary mode. Hence, the masses of the modes seem to have a weight ratio of ca. 5 : 1. This is significantly different from the ini-

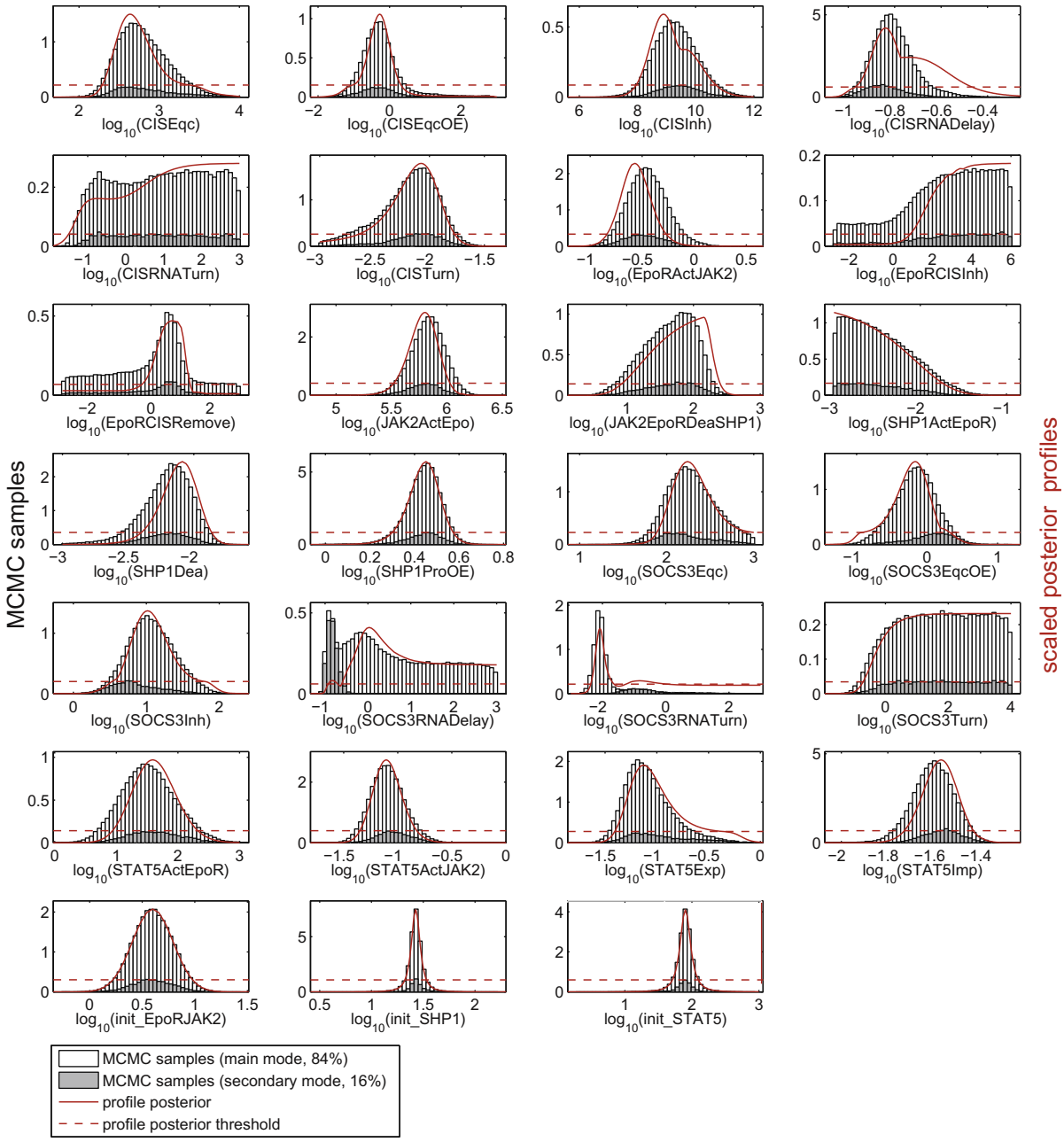

Fig. 3. MCMC samples and profile posterior for all 27 inferred dynamical parameters. Shown are the histograms of the marginalized MCMC samples, color-coded for mode membership. The height is scaled such that the area of all bars in each histogram is one. In red: profile posterior pdf, scaled so as to minimize distance to histogram.

tial weighting of 1 : 1 and thus together with the convergence of the sampling indicates correct weighting of the modes. A more systematic evaluation of the weighting of the modes depending on the initialization of the chains should be the focus of future work.

To evaluate the AMPHS we considered all chains, even though for all further analysis, we only use the samples from the mother chain. Each of the auxiliary chains has at the end an acceptance rate of about 6.5%, the mother chain has per definition an acceptance rate of 100%, since the swap with an auxiliary chain is always accepted. While 6.5% might sound suboptimal, we believe that for the AMPHS sampling scheme the acceptance rate is adequate, since the swaps with the mother chain perturb the adaption of the covariance matrix in the auxiliary chains. This is the price that has to be payed for the excellent mixing in the mother chain.

3.4. Comparison of MCMC results and profile posterior

To compare the profile posterior and the AMPHS sampling results, Fig. 3 shows the histograms of the individual parameters against the corresponding profile posteriors. For most of the parameters, we find an excellent agreement between the shape of the profile posterior and the marginalized samples, cf. exemplary parameters CISEqc or CISEqcOE. However, especially for SOCS3RNADelay, we see a much more pronounced bimodality in the samples than in the profile posterior alone. The same, though not as clearly, hold true for SOCS3RNATurn. The difference between sampling result and posterior profile arises from the fact that the height of the modes – as determined by the posterior profile (maximization) – does not necessarily correspond to the mass

of the mode (marginalization) – as determined by the MCMC sample. Interestingly, for this example, the ratio between the masses of the modes is rather similar to the ratio of the maximum posterior probability density in the individual modes, which is also roughly 5 : 1.

When taking a closer look at Fig. 2, again of the two parameters SOCS3RNADelay and SOCS3RNATurn against each other, one can see that the region of high posterior density is not one with two clear modes with deep valleys in between, but rather a banana-shaped ridge with one global and one local maximum. This highly non-elliptical shape also explains the failure of the single chain AM runs to switch between the two modes adequately fast and often. Obviously, an elliptically-shaped normal distribution is not ideally suited as a proposal distribution for inferring a bimodal banana-shaped distribution. However, in combination with the AMPHS scheme, it is sufficiently efficient.

3.5. Prediction of inhibitory effects

In Bachmann et al. [1], it was already shown that SOCS3 and CIS act as a dual negative feedback on the level of nuclear phosphorylated STAT5, thus providing regulation over a broad range of Epo concentrations. The effect of SOCS3 is more pronounced for high Epo levels, while CIS primarily works as a negative feedback at low Epo levels. In addition to confidence intervals estimated from the profile posterior (see in [1]) the obtained MCMC samples now also allow computing the posterior density of the prediction, see Fig. 4.

Regarding the previously observed modes in the parameter posterior distribution we found that the differences between the corresponding predictions for pSTAT5 are minor, cf. also Fig. 7 in the appendix. This can be explained by the fact that the two modes mainly differ in the SOCS3RNADelay and SOCS3RNATurn which primarily influence SOCS3. As the effect of SOCS3 on pSTAT5 is indirect, the bimodality has negligible effect on the level of pSTAT5 predictions. The results confirm the role of the dual negative feedback. However, the advantage of the sampling-based approach only becomes fully apparent when considering SOCS3 itself.

When we analyzed the predicted dynamics for SOCS3 we found that these indeed depend on the mode, as shown in Fig. 5. For the parameters in the main mode, the model predicts that after stimulation SOCS3 goes directly to a steady state. In contrast, for the parameters in the secondary mode we observe an overshoot. This overshoot is caused by the increased delay in the SOCS3 RNA export from the nucleus, SOCS3RNADelay. Based on this prediction it would be sufficient to have a better resolved measurement of SOCS3 between 0 and 100 min to distinguish between the two modes.

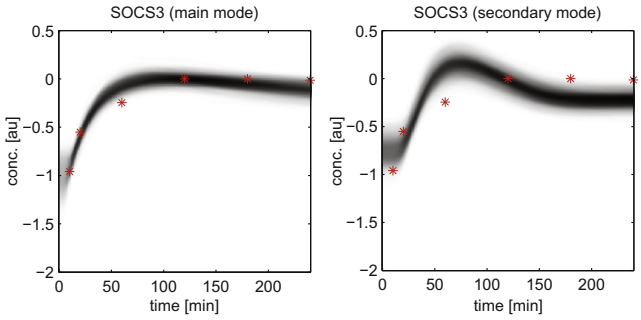

Fig. 5. Differences in SOCS3 dynamics. Experimental data for SOCS3 (red stars) and density of the trajectories corresponding to the MCMC samples for main mode (left) and secondary mode (right). For the parameters in the main mode, the model predicts that after stimulation SOCS3 goes directly to a steady state. In contrast, for the parameters in the secondary mode we observe an overshoot. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

In summary, the obtained MCMC samples allow for a detailed evaluation of the model. Furthermore, new experiments that allow to further characterize the model and improve the explanatory power were designed.

4. Discussion

Statistical inference for high dimensional problems is a challenging issue. In this paper, we provide a proof of concept that Bayesian inference in high-dimensional dynamical systems is feasible. MCMC sampling of over 100 parameters is nevertheless a challenging task. Special care has to be taken when checking and verifying the results.

We studied two different approaches to such problems, the profile posterior approach and Bayesian Markov chain Monte Carlo sampling, both having their own strengths and weaknesses. For the profile posterior approach repeated runs of optimization are necessary. Optimization can pose a computational bottleneck in high dimensional parameter space. In particular, if several modes are present in the posterior, multiple runs of profile calculation can become necessary. In the example considered here, optimization is working efficiently and calculation of the profiles takes about ten to twenty minutes per parameter on a normal desktop computer. MCMC sampling of over 100 parameters is a challenging task as well. Special care has to be taken when checking and verifying the results. We have shown that single-chain algorithms can run into severe

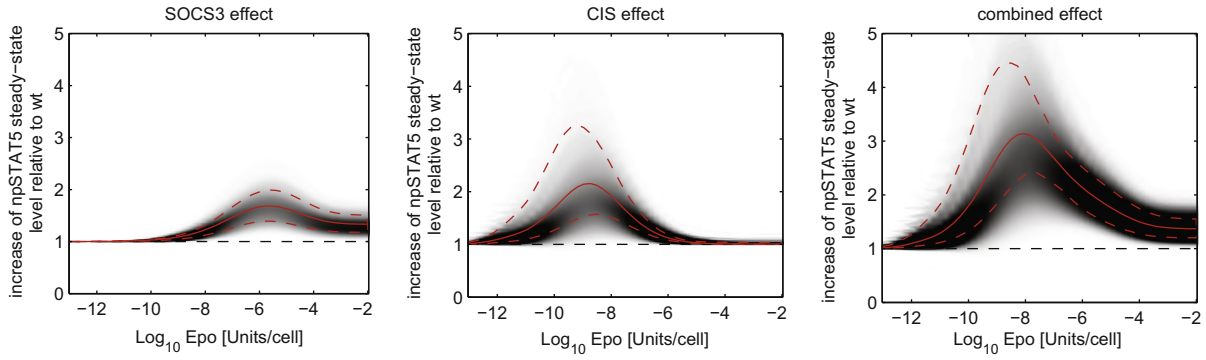

Fig. 4. Uncertainty in prediction of the cellular response. Simulation of the steady-state level of phosphorylated STAT5 in the nucleus, with only one transcriptional negative regulator, CIS or SOCS3, being present and their combined effect. The increase of pSTAT5 steady-state levels was calculated relative to wild-type cells (black dashed line) in steady state. Grey shading indicates the density calculated from the posterior samples, the red line represents the solution belonging to the MAP estimate, dashed red lines indicate 95% confidence bands for the prediction taken from [1]. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

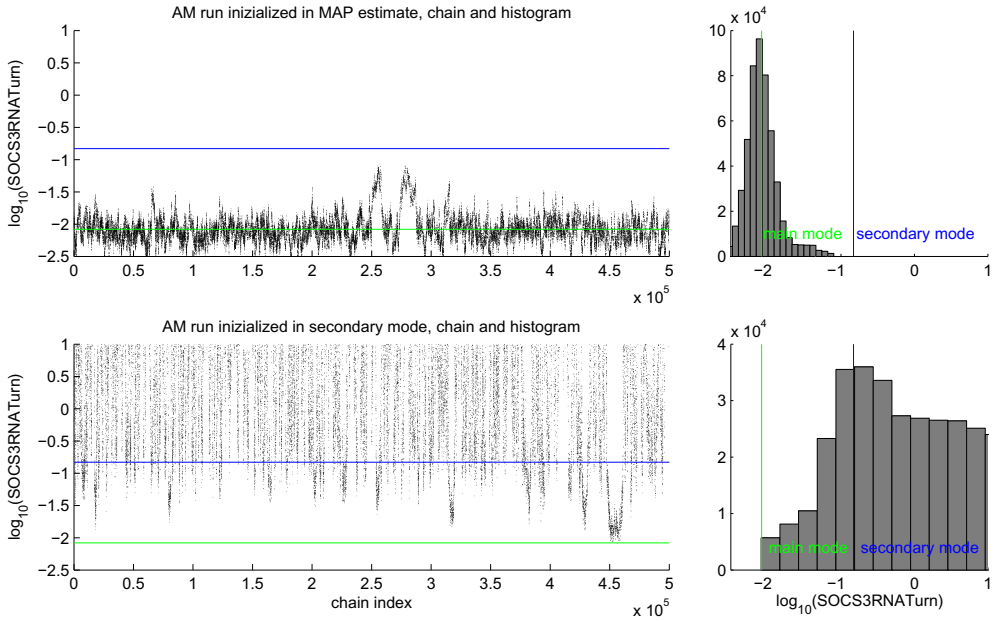

Fig. 6. Results of single-chain AM runs. The top row displays the chain and histogram for the AM run started in the MAP estimate of exemplary parameter SOC3RNATurn, while the bottom row shows the AM run started in the secondary mode. Both chains show nice mixing, however both the chains and the histograms reveal the totally different marginal distributions.

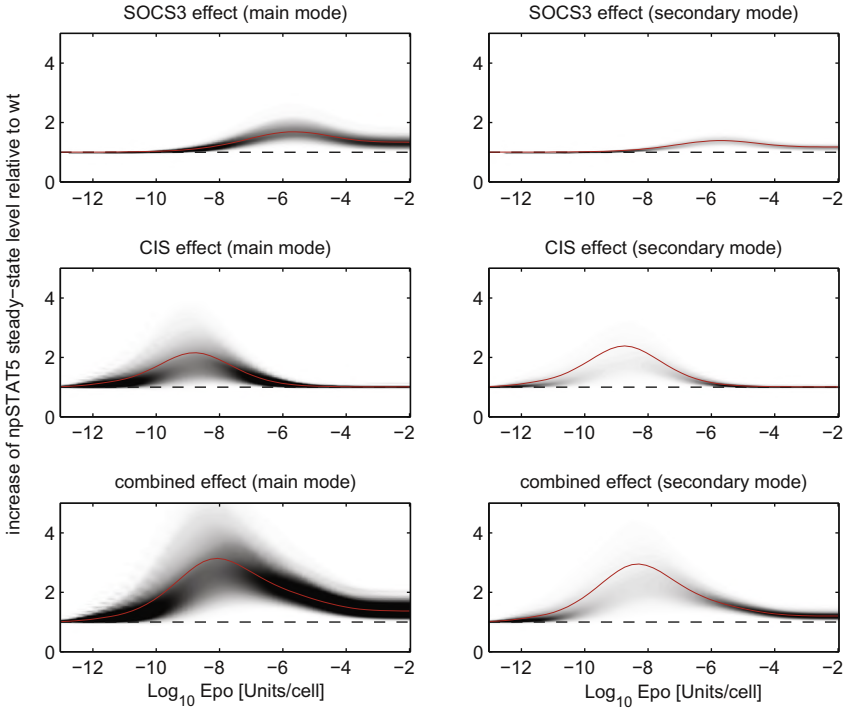

Fig. 7. Uncertainty in prediction of the cellular response. Simulation of the steady-state level of phosphorylated STAT5 in the nucleus, with only one transcriptional negative regulator, CIS or SOCS3, being present and their combined effect. The increase of pSTAT5 steady-state levels was calculated relative to wild-type cells (black dashed line) in steady state. Grey shading indicates the density calculated from the posterior samples, the red line represents the solution belonging to the MAP estimate, dashed red lines indicate 95% confidence bands for the prediction taken from [1]. Left column based on main mode samples, right column based on secondary mode samples. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

problems in high-dimensional systems, which are furthermore not easily diagnosed from the MCMC run alone. In the single-chain case, the Geweke test could not detect that the chains get locked in local modes of the posterior. If single chains are run repeatedly from different starting points, the Gelman–Rubin statistics can detect non-convergence. However, this relies on a representative set of starting points that have to be determined beforehand. For the multi-chain approach, the selection of representative starting points is important to ensure convergence to the posterior distribution, i.e. correct weighting of the posterior modes, in acceptable time. Once reliable results of MCMC sampling are obtained, uncertainties can easily be projected on any model prediction including the high-dimensional correlation structure. In line with results obtained for smaller applications [15,16,19], we advocate the combination of MCMC sampling

with the profile posterior approach to ensure the robustness and reliability of the results.

When breaking down the complexity of the high-dimensional parameter space both our approaches apply different strategies that can lead to different results: marginalization for MCMC sampling and maximization for the profile calculation. In the asymptotic case, i.e. if a sufficient amount of experimental data is available, the results of both approaches agree. This is the case for many parameters considered in this example. However, in the finite case, the results of both approaches can be quite different, as was observed for some parameters in this example. To which extent one or the other approach is preferable remains to be assessed systematically. Results obtained by [16] indicate that marginalization should be used with care in the presence of non-identifiabilities.

Computing the profile posterior is computationally faster and more robust than MCMC sampling. The results of the profile posterior computation, e.g., knowledge about local optima and other structure in the posterior, can then be employed to improve and check MCMC sampling. Nevertheless, even with this additional knowledge, efficient MCMC sampling of the posterior distribution remains an issue. For the JAK2/STAT5 signaling model considered here, we found that single-chain MCMC runs do not mix well. Therefore, we introduced an Adaptive Metropolis Parallel Hierarchical Sampler which can employ knowledge about local maxima in the posterior distribution. Instead of using one chain which starts, e.g., in the MAP estimate, several MCMC chains are started in the individual modes of the posterior distribution with different transition kernels. Combined with the swapping of the states of the chains, this provided better mixing properties. However, even with the AMPHS some care has to be taken, for example when initializing the chains. In our experience, drawing the starting values for the auxiliary chains from the prior is inefficient. Possibly due to the high dimensionality of the parameter space the chains take impractically long to reach region of high posterior density. To increase performance the chains can be initialized in local optima found in the profile posteriors.

In high-dimensional systems it is important to have excellent mixing in the Markov chain to ensure exploration of the whole parameter space. Multi-chain approaches are clearly superior to single-chain sampling schemes from the Metropolis–Hastings family with respect to the mixing. Alternatively, other sampling schemes that provide good mixing could be employed, for example an Independence Walk, however it is unclear what kind of problems one could run into when using these.

To check the convergence of MCMC schemes in high-dimensional parameter spaces or multimodal problems in general we used a combination of approaches. During our studies we found that the complementation of classical convergence criteria [48] with information about the modes is beneficial. This information can be obtained using e.g. multi-start optimization algorithms. If a multi-chain sampler is initialized distributed across all detected modes and the convergence diagnostics indicates convergence, it is more likely that all modes of the posterior distribution have been adequately sampled, compared to the single-chain approach. As multi-start optimization is often used to determine the MAP estimate [1], all necessary information is readily available and does not require additional computational effort.

The sampling of high-dimensional posterior distributions is methodologically and computationally challenging. One might ask whether the additional effort compared to the computation of profile posteriors is justified. The amount of data collected during the analysis is often enormous. To handle these data and to actually allow for additional insight exploration-based visualization methods [49] as well as automated tools, such as (un) supervised learning approaches [47] are necessary. Using the sampling results we could show that the posterior distribution is bimodal and that the two modes correspond to alternative parameterization of the model. Either the turnover rates of SOCS3 RNA can be high and the RNA export in the cytosol low, or vice versa. By inspection of the predictions corresponding to the individual modes it is possible to verify one of the models experimentally. The sampling results were used for the prediction of the inhibitory effect of SOCS3 and CIS for different Epo levels, as well as for the dynamics of SOCS3. The two modes of the posterior clearly manifest in the SOCS3 dynamics. This is due to the fact that the two parameters showing the bimodality most prominently, namely SOCS3RNADelay and SOCS3RNATurn, are directly linked to SOCS3.

The methods presented in this paper are applicable to many dynamical systems, not only from systems biology. Especially in nonlinear and high-dimensional systems relying on a single method or convergence test can be misleading. Combining several approaches should yield a robust procedure for similar inference problems.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) [Virtual Liver (Grant No. 0315766), LungSys II (Grant No. 0316042G), SysMBO (Grant No. 0315494A)], the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (SBCancer DKFZ I.2, V.2 and CoReNe HMGU), the Excellence Initiative of the German Federal and State Governments (EXC 294) and the European Union within the ERC grant "LatentCauses".

We wish to thank the organizers of the PEDS-I and -II workshop, especially Shota Gugushvili, for a nice and fruitful meeting and for organizing this special issue. Furthermore, we thank Natal van Riel and Joep Vanlier for feedback and stimulating discussions and for organizing the NCSB workshop. Moreover, we are grateful to Christiane Fuchs for thorough proofreading of the manuscript.

Appendix A. ODE system and experimental data

The ODE system for the JAK2/STAT5 is described by 29 dynamical variables and was solved for these parameters by using the CVODES solver with an absolute and relative accuracy of 10^{-8}. The ODE equations and the solver were compiled into C-executable files for MATLAB. Experimental data is available for 24 different experimental conditions. As these evaluations of the ODE systems are independent, they could be parallelized for numerical efficiency. After all suitable transformations etc. described in more details in Bachmann et al. [1], 115 unknown parameters remain, of these two more could be fixed to a scale. All in all, like for the analysis in [1], 113 parameters are sampled by our approach. The experimental data for the dynamics of the system consists of 541 data points which are the basis of our inference.

Appendix B. Prior and sampling setup

For one of the parameters, the absolute concentration of the EpoR_JAK complex, we found a literature value that could be included as prior information for the concentration scale of the receptor complex into the sampling. For all other parameters, uniform priors in logarithmic parameter space were used. The range of these priors was determined already for the optimization done in Bachmann et al. [1]. It is especially important for those of the parameters that are non-identifiable. The prior in this case prevents the sampler from going to infinity in these parameters which would be detrimental for the complete exploration of the parameter space.

For the sampling of the JAK2/STAT5 pathway system via the AMPHS sampling scheme, we chose to use 20 auxiliary chains for 500.000 samples each. We do not thin the Markov chains, but save all generated samples. Still, the computational cost for such a large system is quite heavy, the sampling run took about three days on a standard AMD Opteron 2.4 GHz multicore machine using 5 cores. Higher degrees of parallelization are possible and would reduce the run time.

Furthermore, the AMPHS requires the specification of a starting covariance matrix. We found it sufficient to take an identity matrix in each auxiliary chain. Alternatively, one could run a short chain initialized with an identity matrix and then calculate an initial covariance matrix from these prerun samples. Since it is a special advantage of the Parallel Hierarchical sampling scheme that different proposals can be used in each chain, we chose different scaling factors for the identity matrix ranging from 10^{-6} to 10^{-9}.

Appendix C. Support vector machine

For a rigorous assessment of the bimodality of the posterior, we used a two-step procedure based on support vector machines. In the first step a SVM is trained with the first 250.000 MCMC samples of both individual AM chains. In the second step the classification performance of the trained SVM is evaluated on the remaining 250.000 samples of each chain. The percentage of misclassification provides an estimate of the overlap of the two chains as an optimal support vector classification would only misclassify the points in the overlapping region. To obtain the best possible estimate for the overlap we evaluated the classification performance for different kernel functions, i.e. linear, polynomial, and radial basis function, and a variety of different kernel parameters and penalty parameters. This comparison has been carried out using the `LIBSVM` toolbox for MATLAB [50].

Our analysis revealed that already a linear SVM achieves a rather good classification performance with a true-positive rate of 94.5% and a false-positive rate of 4.7%. The improvement observed when using nonlinear SVMs was negligible. Furthermore, we found that the classification performance achieved using all parameters is indistinguishable from the case when using only the dynamical parameters. Even the mere use of SOCS3RNADelay and SOCS3RNATurn for the classification ensures a true-positive rate of 94.0% and a false-positive rate of 5.9%. Altogether this shows that the overlap of the two MCMC chains is approximately 5% and that the samples seem to be separated well using only two parameter dimension.

Appendix D. Additional figures

Fig. 6.

Appendix E. Equations of dynamical model

The rate equations of the reactions are

$$v_1 = \frac{[\mathrm{Epo}]\cdot[\mathrm{EpoRJAK2}]\cdot \mathrm{JAK2ActEpo}}{[\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1}$$

$$v_2 = [\mathrm{EpoRpJAK2}]\cdot \mathrm{JAK2EpoRDeaSHP1}\cdot[\mathrm{SHP1Act}]$$

$$v_3 = \frac{[\mathrm{EpoRpJAK2}]\cdot \mathrm{EpoRActJAK2}}{[\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1}$$

$$v_4 = \frac{3\cdot[\mathrm{EpoRpJAK2}]\cdot \mathrm{EpoRActJAK2}}{(\mathrm{EpoRCISInh}\cdot[\mathrm{EpoRJAK2\_CIS}]+1)\cdot([\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1)}$$

$$v_5 = \frac{3\cdot \mathrm{EpoRActJAK2}\cdot[\mathrm{p1EpoRpJAK2}]}{(\mathrm{EpoRCISInh}\cdot[\mathrm{EpoRJAK2\_CIS}]+1)\cdot([\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1)}$$

$$v_6 = \frac{\mathrm{EpoRActJAK2}\cdot[\mathrm{p2EpoRpJAK2}]}{[\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1}$$

$$v_7 = \mathrm{JAK2EpoRDeaSHP1}\cdot[\mathrm{SHP1Act}]\cdot[\mathrm{p1EpoRpJAK2}]$$

$$v_8 = \mathrm{JAK2EpoRDeaSHP1}\cdot[\mathrm{SHP1Act}]\cdot[\mathrm{p2EpoRpJAK2}]$$

$$v_9 = \mathrm{JAK2EpoRDeaSHP1}\cdot[\mathrm{SHP1Act}]\cdot[\mathrm{p12EpoRpJAK2}]$$

$$v_{10} = [\mathrm{EpoRJAK2\_CIS}]\cdot \mathrm{EpoRCISRemove}\cdot([\mathrm{p12EpoRpJAK2}]+[\mathrm{p1EpoRpJAK2}])$$

$$v_{11} = [\mathrm{SHP1}]\cdot \mathrm{SHP1ActEpoR}\cdot([\mathrm{EpoRpJAK2}]+[\mathrm{p12EpoRpJAK2}]+[\mathrm{p1EpoRpJAK2}]+[\mathrm{p2EpoRpJAK2}])$$

$$v_{12} = \mathrm{SHP1Dea}\cdot[\mathrm{SHP1Act}]$$

$$v_{13} = \frac{[\mathrm{STAT5}]\cdot \mathrm{STAT5ActJAK2}}{[\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1}\cdot([\mathrm{EpoRpJAK2}]+[\mathrm{p12EpoRpJAK2}]+[\mathrm{p1EpoRpJAK2}]+[\mathrm{p2EpoRpJAK2}])$$

$$v_{14} = \frac{[\mathrm{STAT5}]\cdot \mathrm{STAT5ActEpoR}\cdot([\mathrm{p12EpoRpJAK2}]+[\mathrm{p1EpoRpJAK2}])^2}{([\mathrm{CIS}]\cdot \mathrm{CISInh}+1)\cdot([\mathrm{SOCS3}]\cdot \mathrm{SOCS3Inh}+1)}$$

$$v_{15} = \mathrm{STAT5Imp}\cdot[\mathrm{pSTAT5}]$$

$$v_{16} = \mathrm{STAT5Exp}\cdot[\mathrm{npSTAT5}]$$

$$v_{17} = -\mathrm{CISRNAEqc}\cdot \mathrm{CISRNATurn}\cdot[\mathrm{npSTAT5}]$$

$$v_{18} = [\mathrm{CISnRNA1}]\cdot \mathrm{CISRNADelay}$$

$$v_{19} = [\mathrm{CISnRNA2}]\cdot \mathrm{CISRNADelay}$$

$$v_{20} = [\mathrm{CISnRNA3}]\cdot \mathrm{CISRNADelay}$$

$$v_{21} = [\mathrm{CISnRNA4}]\cdot \mathrm{CISRNADelay}$$

$$v_{22} = [\mathrm{CISnRNA5}]\cdot \mathrm{CISRNADelay}$$

$$v_{23} = [\mathrm{CISRNA}]\cdot \mathrm{CISRNATurn}$$

$$v_{24} = [\mathrm{CISRNA}]\cdot \mathrm{CISEqc}\cdot \mathrm{CISTurn}$$

$$v_{25} = [\mathrm{CIS}]\cdot \mathrm{CISTurn}$$

$$v_{26} = -\mathrm{SOCS3RNAEqc}\cdot \mathrm{SOCS3RNATurn}\cdot[\mathrm{npSTAT5}]$$

$$v_{27} = [\mathrm{SOCS3nRNA1}]\cdot \mathrm{SOCS3RNADelay}$$

$$v_{28} = [\mathrm{SOCS3nRNA2}]\cdot \mathrm{SOCS3RNADelay}$$

$$v_{29} = [\mathrm{SOCS3nRNA3}]\cdot \mathrm{SOCS3RNADelay}$$

$$v_{30} = [\mathrm{SOCS3nRNA4}]\cdot \mathrm{SOCS3RNADelay}$$

$$v_{31} = [\mathrm{SOCS3nRNA5}]\cdot \mathrm{SOCS3RNADelay}$$

$$v_{32} = [\mathrm{SOCS3RNA}]\cdot \mathrm{SOCS3RNATurn}$$

$$v_{33} = [\mathrm{SOCS3RNA}]\cdot \mathrm{SOCS3Eqc}\cdot \mathrm{SOCS3Turn}$$

$$v_{34} = [\mathrm{SOCS3}]\cdot \mathrm{SOCS3Turn}$$

Reactions v_{18} to v_{22} and v_{27} to v_{31} account for a delay that summarize the processing steps of the mRNA by a linear chain of reactions [51] with common rate constant CISRNADelay and SOCS3RNADelay, respectively. The ODE systems is composed out of the rate equations by

$$\mathrm{d}[\mathrm{EpoRJAK2}]/\mathrm{d}t = -v_1 + v_2 + v_7 + v_8 + v_9$$

$$\mathrm{d}[\mathrm{EpoRpJAK2}]/\mathrm{d}t = +v_1 - v_2 - v_3 - v_4$$

$$\mathrm{d}[\mathrm{p1EpoRpJAK2}]/\mathrm{d}t = +v_3 - v_5 - v_7$$

$$\mathrm{d}[\mathrm{p1EpoRpJAK2}]/\mathrm{d}t = +v_4 - v_6 - v_8$$

$$\mathrm{d}[\mathrm{p12EpoRpJAK2}]/\mathrm{d}t = +v_5 + v_6 - v_9$$

$$\mathrm{d}[\mathrm{EpoRJAK2\_CIS}]/\mathrm{d}t = -v_{10}$$

$$\mathrm{d}[\mathrm{SHP1}]/\mathrm{d}t = -v_{11} + v_{12}$$

$$\mathrm{d}[\mathrm{SHP1Act}]/\mathrm{d}t = +v_{11} - v_{12}$$

$$\mathrm{d}[\mathrm{STAT5}]/\mathrm{d}t = -v_{13} - v_{14} + v_{16}\cdot\frac{0.275}{0.4}$$

$$\mathrm{d}[\mathrm{pSTAT5}]/\mathrm{d}t = +v_{13} + v_{14} - v_{15}$$

$$\mathrm{d}[\mathrm{npSTAT5}]/\mathrm{d}t = +v_{15}\cdot\frac{0.4}{0.275} - v_{16}$$

$$\mathrm{d}[\mathrm{CISnRNA1}]/\mathrm{d}t = +v_{17} - v_{18}$$

$$
\begin{aligned}
\mathrm{d}[\mathrm{CISnRNA2}]/\mathrm{d}t &= +v_{18} - v_{19}\\
\mathrm{d}[\mathrm{CISnRNA3}]/\mathrm{d}t &= +v_{19} - v_{20}\\
\mathrm{d}[\mathrm{CISnRNA4}]/\mathrm{d}t &= +v_{20} - v_{21}\\
\mathrm{d}[\mathrm{CISnRNA5}]/\mathrm{d}t &= +v_{21} - v_{22}\\
\mathrm{d}[\mathrm{CISRNA}]/\mathrm{d}t &= +v_{22}\cdot\frac{0.275}{0.4} - v_{23}\\
\mathrm{d}[\mathrm{CIS}]/\mathrm{d}t &= +v_{24} - v_{25}\\
\mathrm{d}[\mathrm{SOCS3nRNA1}]/\mathrm{d}t &= +v_{26} - v_{27}\\
\mathrm{d}[\mathrm{SOCS3nRNA2}]/\mathrm{d}t &= +v_{27} - v_{28}\\
\mathrm{d}[\mathrm{SOCS3nRNA3}]/\mathrm{d}t &= +v_{28} - v_{29}\\
\mathrm{d}[\mathrm{SOCS3nRNA4}]/\mathrm{d}t &= +v_{29} - v_{30}\\
\mathrm{d}[\mathrm{SOCS3nRNA5}]/\mathrm{d}t &= +v_{30} - v_{31}\\
\mathrm{d}[\mathrm{SOCS3RNA}]/\mathrm{d}t &= +v_{31}\cdot\frac{0.275}{0.4} - v_{32}\\
\mathrm{d}[\mathrm{SOCS3}]/\mathrm{d}t &= +v_{33} - v_{34}.
\end{aligned}
$$

The volume factors vol_cyt = 0.4 pl and vol_nuc = 0.275 pl account for transitions between different compartments and are determined experimentally. The species npSTAT5, ClSnRNA1–5 and SOCS3nRNA1–5 are located in the nuclear compartment, the remaining species in the cytoplasmatic compartment.

The initial condition are set to zero except for

$$
\begin{aligned}
[\mathrm{EpoRJAK2}](0) &= \mathrm{init\_EpoRJAK2}\\
[\mathrm{SHP1}](0) &= \mathrm{init\_SHP1}\\
[\mathrm{STAT5}](0) &= \mathrm{init\_STAT5}.
\end{aligned}
$$

References

[1] J. Bachmann, A. Raue, M. Schilling, M. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. Lehmann, J. Timmer, U. Klingmüller, Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range, Molecular Systems Biology 7 (2011).

[2] V. Becker, M. Schilling, J. Bachmann, U. Baumann, A. Raue, T. Maiwald, J. Timmer, U. Klingmüller, Covering a broad dynamic range: information processing at the erythropoietin receptor, Science 328 (2010) 1404.

[3] D. Wilkinson, Stochastic Modelling for Systems Biology, Chapman & Hall/CRC, 2006.

[4] T.-R. Xu, V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, G.S. Baillie, D. Ketley, A.J. Dunlop, G. Milligan, M.D. Houslay, W. Kolch, Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species, Science Signaling 3 (2010) 1.

[5] M. Girolami, Bayesian inference for differential equations, Theoretical Computer Science 408 (2008) 4.

[6] V. Raia, M. Schilling, M. Böhm, B. Hahn, A. Kowarsch, A. Raue, C. Sticht, S. Bohl, M. Saile, P. Möller, N. Gretz, J. Timmer, F. Theis, W. Lehmann, P. Lichter, U. Klingmüller, Dynamic mathematical modeling of IL13-induced signaling in hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets, Cancer Research 71 (2011) 1.

[7] S. Bohl, Dynamic modeling of signal processing for IL-6-induced STAT3 signal transduction in primary mouse hepatocytes, Ph.D. thesis, Ruperto-Carola University of Heidelberg, Germany, 2009.

[8] N. Lawrence, M. Girolami, M. Rattray, G. Sanguinetti, Leaning and Inference in Computational Systems Biology, MIT Press, 2010.

[9] J. Timmer, T. Müller, I. Swameye, O. Sandra, U. Klingmüller, Modeling the nonlinear dynamics of cellular signal transduction, International Journal of Bifurcation and Chaos 14 (2004) 2069.

[10] E. Coddington, N. Levinson, Theory of Ordinary Differential Equations, Tata McGraw-Hill, 1972.

[11] C. Kreutz, M.M.B. Rodriguez, T. Maiwald, M. Seidl, H.E. Blum, L. Mohr, J. Timmer, An error model for protein quantification, Bioinformatics 23 (2007) 2747.

[12] R. Serban, A. Hindmarsh, Cvodes: the sensitivity-enabled ode solver in Sundials, in: Proceedings of IDETC/CIE, vol. 24.

[13] K. Brown, J. Sethna, Statistical mechanical approaches to models with many poorly known parameters, Physical Review E 68 (2003) 021904.

[14] J. Bernardo, A. Smith, M. Berliner, Bayesian Theory, vol. 62, Wiley, 1994.

[15] J. Vanlier, C. Tiemann, P. Hilbers, N. van Riel, An integrated strategy for prediction uncertainty analysis, Bioinformatics 28 (2012) 1130.

[16] A. Raue, C. Kreutz, F. Theis, J. Timmer, Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability, Philosophical Transacions of the Royal Society A 371 (2013).

[17] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics 25 (2009) 1923.

[18] C. Kreutz, A. Raue, J. Timmer, Likelihood based observability analysis and confidence intervals for predictions of dynamic models, BMC Systems Biology 6 (2012) 120.

[19] D. Schmidl, S. Hug, W. Li, M. Greiter, F. Theis, Bayesian model selection validates a biokinetic model for zirconium processing in humans, BMC Systems Biology 6 (2012) 95.

[20] B. Calderhead, M. Girolami, Estimating Bayes factors via thermodynamic integration and population MCMC, Computational Statistics & Data Analysis 53 (2009) 4028.

[21] R. Kass, A. Raftery, Bayes factors, Journal of the American Statistical Association (1995) 773.

[22] N. Lartillot, H. Philippe, Computing Bayes factors using thermodynamic integration, Systematic biology 55 (2006) 195.

[23] N. Friel, A. Pettitt, Marginal likelihood estimation via power posteriors, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (2008) 589.

[24] R. Neal, Probabilistic inference using Markov chain Monte Carlo methods, Technical Report CRG-TR-93-1, University of Toronto, Department of Computer Science, 1993.

[25] D. Gamerman, H. Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, vol. 68, Chapman & Hall/CRC, 2006.

[26] R. Kass, B. Carlin, A. Gelman, R. Neal, Markov chain Monte Carlo in practice: A roundtable discussion, American Statistician 52 (1998) 93.

[27] S. Brooks, Markov chain Monte Carlo method and its application, Journal of the Royal Statistical Society: Series D (The Statistician) 47 (1998) 69.

[28] J. Liu, Monte Carlo Strategies in Scientific Computing, Springer-Verlag, 2008.

[29] J. Marin, C. Robert, Bayesian Core: A Practical Approach to Computational Bayesian Statistics, Springer Verlag, 2007.

[30] C. Robert, G. Casella, C. Robert, Monte Carlo Statistical Methods, vol. 2, Springer, New York, 1999.

[31] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, The Journal of Chemical Physics 21 (1953) 1087.

[32] W. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97.

[33] I. Beichl, F. Sullivan, The Metropolis algorithm, Computing in Science & Engineering 2 (2000) 65.

[34] G. Roberts, A. Gelman, W. Gilks, Weak convergence and optimal scaling of random walk Metropolis algorithms, The Annals of Applied Probability 7 (1997) 110.

[35] D. Schmidl, C. Czado, S. Hug, F.J. Theis, A vine-copula based adaptive MCMC approach for efficient inference of dynamical systems, Bayesian Analysis 8 (2013) 1.

[36] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2011) 123.

[37] D. Schmidl, Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods, Ph.D. thesis, Technische Universität München, 2012.

[38] J. Vanlier, C. Tiemann, P. Hilbers, N. van Riel, Parameter uncertainty in biochemical models described by ordinary differential equations, Mathematical Biosciences, 2013.

[39] H. Haario, E. Saksman, J. Tamminen, An adaptive Metropolis algorithm, Bernoulli 7 (2001) 223.

[40] A. Gelman, D. Rubin, Inference from iterative simulation using multiple sequences, Statistical Science 7 (1992) 457.

[41] R. Neal, Sampling from multimodal distributions using tempered transitions, Statistics and Computing 6 (1996) 353.

[42] K. Hukushima, K. Nemoto, Exchange monte carlo method and application to spin glass simulations, Journal of the Physical Society of Japan 65 (1996) 1604.

[43] A. Jasra, D.A. Stephens, C.C. Holmes, Population-based reversible jump Markov chain Monte Carlo, Biometrika 94 (2007) 787.

[44] F. Rigat, A. Mira, Parallel hierarchical sampling: a general-purpose interacting Markov chains Monte Carlo algorithm, Computational Statistics & Data Analysis 56 (2012) 1450.

[45] C. Geyer, Practical Markov Chain Monte Carlo, Statistical Science (1992).

[46] J. Geweke, Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, Federal Reserve Bank of Minneapolis, Research Department, 1991.

[47] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[48] S.P. Brooks, G.O. Roberts, Assessing convergence of Markov chain Monte Carlo algorithms, Statistical Computation 8 (1998) 319.

[49] C. Vehlow, J. Hasenauer, A. Kramer, J. Heinrich, N. Radde, F. Allgöwer, D. Weiskopf, Uncertainty-aware visual analysis of biochemical reaction networks, in: IEEE Symposium on Biological Data Visualization, 2012, pp. 91–98.

[50] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, IEEE/ACM Transactions om Intelligent System Technology 2 (2011) 1.

[51] N. MacDonald, Time delay in simple chemostat models, Biotechnology and Bioengineering 18 (1976) 805.