# Hierarchical optimization for the efficient parametrization of ODE models

Carolin Loos[a,1], Sabrina Krause[a,1], and Jan Hasenauer[a,2]

[a]Helmholtz Zentrum München-German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany, and Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany

(Dated: January 14, 2018)

## Abstract

Mathematical models are nowadays important tools for analyzing dynamics of cellular processes. The unknown model parameters are usually estimated from experimental data. These data often only provide information about the relative changes between conditions, hence, the observables contain scaling parameters. The unknown scaling parameters and corresponding noise parameters have to be inferred along with the dynamic parameters. The nuisance parameters often increase the dimensionality of the estimation problem substantially and cause convergence problems. In this manuscript, we propose a hierarchical optimization approach for estimating the parameters for ordinary differential equation (ODE) models from relative data. Our approach restructures the optimization problem into an inner and outer subproblem. These subproblems possess lower dimensions than the original optimization problem, and the inner problem can be solved analytically. We evaluated accuracy, robustness, and computational efficiency of the hierarchical approach by studying three signaling pathways. The proposed approach achieved better convergence than the standard approach and required a lower computation time. As the hierarchical optimization approach is widely applicable, it provides a powerful alternative to established approaches.

## 1 Introduction

Mechanistic mathematical models are used in systems biology to improve the understanding of biological processes. The mathematical models most frequently used in systems biology are probably ordinary differential equations (ODEs). ODE models are, among others, used to describe the dynamics of biochemical reaction networks (Kitano, 2002; Klipp et al., 2005; Schöberl et al., 2009) and proliferation/differentiation processes (De Boer et al., 2006). The dynamic parameters of the underlying processes, e.g., reaction rates and initial conditions, are often unknown and need to be inferred from available experimental data. The inference provides information about the plausibility of the model topology, and the inferred parameters might for instance be used to predict latent variables or the response of the process to perturbations (Molinelli et al., 2013).

The experimental data used for parameter estimation are produced by various experimental techniques. Most of these techniques provide relative data, meaning that the observation is proportional to a variable

[1]C.L. and S.K. contributed equally to this work.

[2]Lead contact, e-mail: jan.hasenauer@helmholtz-muenchen.de.

of interest, e.g., the concentration of a chemical species. This is for instance the case for Western blotting (Renart et al., 1979) and flow and mass cytometry (Herzenberg et al., 2006). If calibration curves are generated, the measured intensities can be converted to concentrations, however, in most studies this is not done due to increased resource demands.

In the literature, two methods are employed to link relative data to mathematical models: (i) evaluation of relative changes (Degasperi et al., 2017) and (ii) introduction of scaling parameters (Raue et al., 2013). In (i), relative changes between conditions are compared, and the differences between observed and simulated relative changes are minimized. While this approach is intuitive and does not alter the dimension of the fitting problem, the noise distribution is non-trivial and the residuals are not uncorrelated (Thomaseth and Radde, 2016). This is often disregarded (see, e.g., (Degasperi et al., 2017)), which yields incorrect confidence intervals. In (ii), scaling parameters are introduced to replace the calibration curves. The scaling parameters are unknown and have to be inferred along with the dynamic parameters. While this increases the dimensionality of the optimization problem (see (Bachmann et al., 2011) for an example in which the number of parameters is doubled), the noise distribution is simple and the confidence intervals consistent. To address the dimensionality increase, Weber et al. (2011) proposed an approach for estimating the conditionally optimal scaling parameters given the dynamic parameters. This approach eliminated the scaling parameters, however, it is only applicable in the special case of additive Gaussian noise with known standard deviation. Unknown noise parameters and outlier-corrupted data (Maier et al., 2017) – as found in many applications – cannot be handled.

In this study, we propose a hierachical optimization approach which generalizes the idea of Weber et al. (2011). The proposed hierarchical approach allows for arbitrary noise distributions, with known and unknown noise parameters. For Gaussian and Laplace noise, we provide analytic solutions for the inner optimization problem, which boosts the computational efficiency. To illustrate the properties of the proposed approach, we present results for two models of JAK-STAT signaling and a model of RAF/MEK/ERK signaling.

## 2 Methods

In this section, we describe the considered class of parameter estimation problems and introduce a hierarchical optimization method for estimating the parameters of ODE models from relative data under different measurement noise assumptions.

### 2.1 Mechanistic modeling of biological systems

We considered ODE models of biological processes,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t, \boldsymbol{\theta}), \boldsymbol{\theta}), \quad \mathbf{x}(t_0, \boldsymbol{\theta}) = x_0(\boldsymbol{\theta}), \tag{1}$$

in which the time- and parameter-dependent state vector $\mathbf{x}(t, \boldsymbol{\theta}) \in \mathbb{R}^{n_x}$ represents the concentrations of the species involved in the process and the vector field $f \colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$ determines how the concentrations evolve over time. The vector $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ denotes the parameters of the system, e.g., reaction rates. The initial conditions at time point $t_0$ are given by the parameter-dependent function $x_0 : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$.

Experimental data provide information about observables $\mathbf{y}(t, \theta) \in \mathbb{R}^{n_y}$. These are obtained by the

output function $\mathbf{h}\colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_y}$, which maps the states and parameters to the observables via

$$\mathbf{y}(t, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}), \boldsymbol{\theta})\,. \tag{2}$$

Due to experimental limitations the experimental data is noise corrupted,

$$\bar{y}_{i,k} = h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta}) + \varepsilon_{i,k}\,, \tag{3}$$

with $h_i$ denoting the $i$th component of the output function $\mathbf{h}$, and indices $k$ for the time point. In most applications, Gaussian noise is assumed, $\varepsilon_{i,k} \sim \mathcal{N}(0, {\sigma_{i,k}}^2)$. For outlier-corrupted data, it was shown that the assumption of Laplace noise, $\varepsilon_{i,k} \sim \text{Laplace}(0, \sigma_{i,k})$, yields more robust results (see (Maier et al., 2017) and references therein).

The measurements are collected in a dataset $\mathcal{D} = \{\bar{\mathbf{y}}_k, t_k\}_k$. The vector $\bar{\mathbf{y}}_k = (\bar{y}_{1,k}, \ldots, \bar{y}_{n_y,k})^T$ comprises the measurements for the different observables. For the general case including different experiments and conditions, we refer to the Supplementary Information, Section 1.

## 2.2 Relative experimental data

Many experimental techniques provide data which are proportional to the measured concentrations. The scaling parameters are usually incorporated in $\mathbf{h}$, defined in (2). Here, for simplicity and without loss of generality, we unplugged the scaling parameters from the function $\mathbf{h}$ and write

$$\bar{y}_{i,k} = s_{i,k} \cdot\ h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta}) + \varepsilon_{i,k}\,.$$

The scaling parameters $s_{i,k}$ and the noise parameters $\sigma_{i,k}$ are in the following combined in the matrices $\mathbf{s}$ and $\boldsymbol{\sigma}$, respectively. To distinguish the different parameter types, we refer to the parameters $\boldsymbol{\theta}$ further as dynamic parameters. In the following, we present results for the case that the scaling $s_i$ and noise parameters $\sigma_i$ are the same for each time point, but differ between observables. The general case is presented in the Supplementary Information, Section 1.

## 2.3 Formulation of parameter estimation problem from relative data

We used maximum likelihood methods, a commonly used approach to calibrate mathematical model, to estimate the parameters from experimental data. The likelihood function is given by

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = \prod_{i,k} p(\bar{y}_{i,k} | s_i \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta}), \sigma_i) \tag{4}$$

with $p$ denoting the conditional probability of $\bar{y}_{i,k}$ given the observable $y_{i,k} = s_i \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})$. This probability is for Gaussian noise

$$p(\bar{y}_{i,k} | y_{i,k}, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(\bar{y}_{i,k} - y_{i,k})^2}{2\sigma_i^2}\right)$$

with standard deviation $\sigma_i$, and for Laplace noise

$$p(\bar{y}_{i,k} | y_{i,k}, \sigma_i) = \frac{1}{2\sigma_i} \exp\left(-\frac{|\bar{y}_{i,k} - y_{i,k}|}{\sigma_i}\right)\,.$$

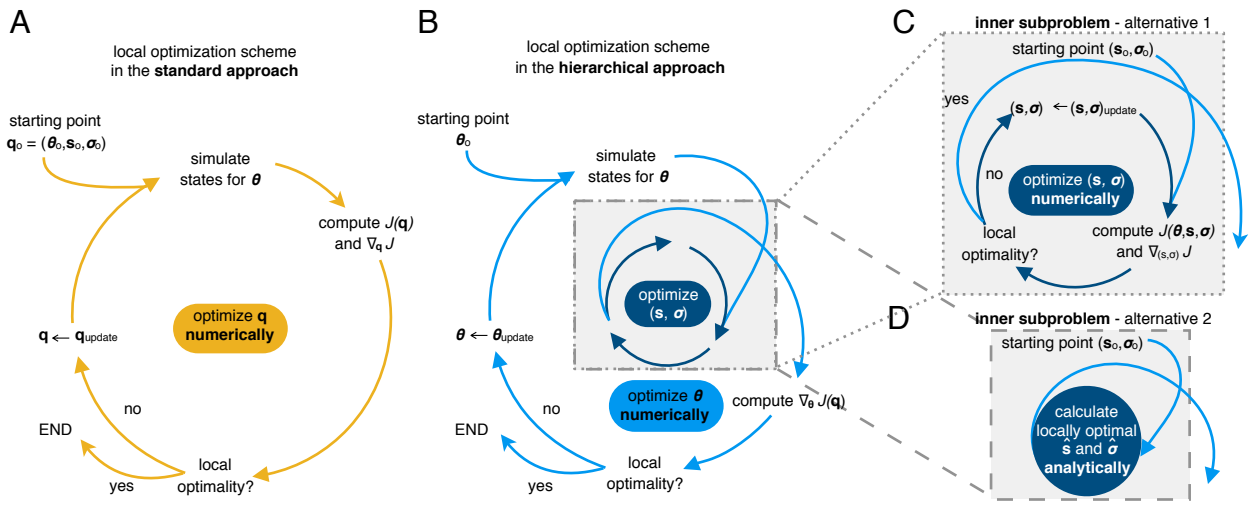



**Figure 1:** Visualization of standard and hierarchical optimization schemes. (**A**) Local optimization in the standard approach with parameters $\mathbf{q} = (\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})$. A single iteration includes the numerical simulation of the ODE model for $\boldsymbol{\theta}$, the evaluation of the objective function and its gradient, the evaluation of local optimality and stopping criteria, and the termination of the local optimization or the updating of the parameters. (**B**) Outer local optimization in the hierarchical approach with parameters $\boldsymbol{\theta}$. A single iteration includes the numerical simulation of the ODE model $\boldsymbol{\theta}$, the evaluation of the objective function and its gradient with respect to $\boldsymbol{\theta}$ using the results of the inner optimization problem, The iteration also includes the evaluation of local optimality and stopping criteria, and the termination of the local optimization or the updating of parameters. (**C,D**) Inner (local) optimization in the hierarchical approach to find the optimal scaling and noise parameter $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$ for given dynamic parameters $\boldsymbol{\theta}$. (**C**) Iterative local optimization to determine $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$. This does not require the numerical simulation of the model. (**D**) Calculating optimal parameters $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$ using analytic expressions for common noise distributions.

with scale parameter $\sigma_i$.

### 2.3.1 Standard approach to parameter estimation

For the standard approach, the dynamic parameters $\boldsymbol{\theta}$, the scaling parameters $\mathbf{s}$, and the noise parameters $\boldsymbol{\sigma}$ are estimated simultaneously. For numerical reasons, this is mostly done by minimizing the negative log-likelihood function,

$$\min_{\boldsymbol{\theta},\mathbf{s},\boldsymbol{\sigma}} J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) \quad \text{with} \quad J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = -\log \mathcal{L}(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})\,. \tag{5}$$

The parameters were combined as $\mathbf{q} = (\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})$ and the optimization problem has the dimension: number of dynamic parameters $n_\theta$ + number of scaling parameters $n_s$ + number of noise parameters $n_\sigma$. We solved it using multi-start local optimization, a method which has previously been shown to be computationally efficient. In each iteration the objective function and its gradient were computed. If the objective function for this parameters fulfills certain criteria, e.g., the norm of the gradient was below a certain threshold, the optimization was stopped, otherwise the parameter was updated and the procedure was continued (Figure 1A).

### 2.3.2 Hierarchical approach to parameter estimation

Since the optimization problem (5) often possess a large number of optimization variables and can be difficult to solve, we exploited its structure. Instead of solving simultaneously for $\boldsymbol{\theta}, \mathbf{s}$, and $\boldsymbol{\sigma}$, we considered the

hierarchical optimization problem (Figure 1B)

$$\min_{\boldsymbol{\theta}} \; J(\boldsymbol{\theta}, \hat{\mathbf{s}}(\boldsymbol{\theta}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\theta})) \tag{6}$$

$$\text{with} \quad (\hat{\mathbf{s}}(\boldsymbol{\theta}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\theta})) = \underset{\mathbf{s},\boldsymbol{\sigma}}{\operatorname{argmin}} \; J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})\,. \tag{7}$$

The inner problem (7) provides the optimal values $\hat{\mathbf{s}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\sigma}}(\boldsymbol{\theta})$ of $\mathbf{s}$ and $\boldsymbol{\sigma}$ given $\boldsymbol{\theta}$. These optimal values were used in the outer subproblem to determine the optimal value for $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}}$. It is apparent that a locally optimal point of the standard optimization problem (5) is also locally optimal for the hierarchical optimization problem (6,7), if the point is within the box constraints for the optimization.

The formulation (6) might appear more involved, however, it possesses several properties which might be advantageous:

(i) The individual dimensions of the inner and outer subproblems (6,7) are lower than the dimension of the original problem (5).

(ii) The optimization of the inner subproblem does not require the repeated numerical simulation of the ODE model.

(iii) For several noise models, e.g., Gaussian and Laplace noise, the inner subproblem can be solved analytically.

If (iii) holds, the scaling parameters $\mathbf{s}$ and also the noise parameters $\boldsymbol{\sigma}$ can be calculated directly and the amount of parameters that need to be optimized iteratively reduces to $n_\theta$ (Figure 1C,D). In the following two sections, the analytic expressions for the Gaussian and Laplace noise are derived. For this, let observable index $i$ be arbitrary but fixed.

### Analytic expressions for the optimal scaling and noise parameters for Gaussian noise

In this study, we evaluated the scaling and noise parameters for Gaussian noise analytically. To derive the analytic expression for the optimal parameters, we exploited that the objective function for Gaussian noise,

$$J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = \frac{1}{2} \sum_{i,k} \log(2\pi\sigma_i^2) + \left( \frac{\bar{y}_{i,k} - s_i \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})}{\sigma_i} \right)^2 .$$

is continuously differentiable, and that the gradient of $J$ at a local minimum is zero. For the inner subproblem this implies

$$\nabla_s J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})|_{\hat{\mathbf{s}},\hat{\boldsymbol{\sigma}}} = \mathbf{0} \text{ and } \nabla_\sigma J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma})|_{\hat{\mathbf{s}},\hat{\boldsymbol{\sigma}}} = \mathbf{0}.$$

These equations can be solved analytically (see Supplementary Information, Section 1), which yields

$$\hat{s}_i(\boldsymbol{\theta}) = \frac{\sum_k \bar{y}_{i,k} \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})}{\sum_k h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})^2}$$

$$\hat{\sigma}_i^2(\boldsymbol{\theta}) = \frac{1}{n_k} \sum_k \left( \bar{y}_{i,k} - \hat{s}_i(\boldsymbol{\theta}) \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta}) \right)^2$$

with number of time points $n_k$. Consistent with the structure of the hierarchical problem (6), both formulas depend only on the dynamic parameters $\boldsymbol{\theta}$.

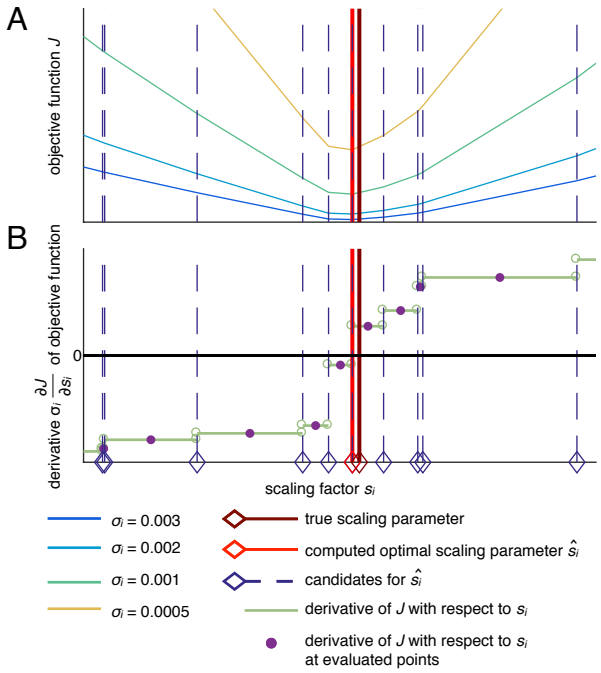



**Figure 2:** Illustration of the computation of an optimal scaling parameter $\hat{s}_i$ for Laplace noise. (**A**) Objective function $J$ for different values of $\sigma_i$, showing that the kinks indicated by the dashed lines are independent of that value. (**B**) Derivative of the objective function with respect to the scaling parameter which is not defined at the kinks. The light red and dark red lines indicate the computed scaling parameter and the true optimal scaling parameter, respectively.

In many studies (e.g., (Bachmann et al., 2011)), observation functions of the form $\log(\bar{y}_{i,k}) = \log(s_i h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})) + \epsilon_i$ are used. In the Supplementary Information, Section 2, we provide a derivation of the corresponding optimal parameters.

### Analytic expressions for the optimal scaling and noise parameters for Laplace noise

For Laplace noise the negative log-likelihood function is

$$J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = \sum_{i,k} \log(2\sigma_i) + \frac{|\bar{y}_{i,k} - s_i \cdot h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})|}{\sigma_i}. \tag{8}$$

This objective function is continuous but not continuously differentiable. In this case, a sufficient condition for a local minimum is that the right limit value of the derivative is negative and the left limit value is positive. The derivative of (8) with respect to $s_i$ can be written as

$$\frac{\partial J}{\partial s_i} = \frac{1}{\sigma_i} \cdot \sum_k \left( |h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})| \cdot \operatorname{sgn}\left( \frac{\bar{y}_{i,k}}{h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})} - s_i \right) \right),$$

As $\sigma_i$ is positive, the locations of kinks in the objective function and the corresponding jumps in the derivative are independent of $\sigma_i$ (Figure 2). Accordingly, the problem of finding $\hat{s}_i$ reduced to checking the signs of the derivative before and after the jump points $s_{i,k} = \bar{y}_{i,k}/h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})$. We sorted $s_{i,k}$ in increasing order and evaluated the derivatives at the midpoints between adjacent jumps, a procedure which is highly efficient as the ODE model does not have to be simulated. Given $\hat{s}_i$, the noise parameter $\hat{\sigma}_i$ follows from the work of

Norton (1984) as

$$\hat{\sigma}_i(\boldsymbol{\theta}) = \frac{1}{n_k} \sum_k \left( |h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})| \cdot \left| \frac{\bar{y}_{i,k}}{h_i(\mathbf{x}(t_k, \boldsymbol{\theta}), \boldsymbol{\theta})} - \hat{s}_i(\boldsymbol{\theta}) \right| \right).$$

Both derived formulas depend only on the dynamic parameters $\boldsymbol{\theta}$, in consistence with the structure of the hierarchical problem (6). In summary, we reformulated the original optimization problem (5) as a hierarchical optimization problem (6,7), and provided an analytic solution to the inner subproblem (7) for several relevant cases. Using the analytic solutions, the kinetic parameters can be inferred by solving a lower-dimensional problem.

# 3 Results

To study and compare the performance of parameter estimation from relative data using the standard approach and our hierarchical approach, we applied both to three published estimation problems.

## 3.1 Models and experimental data

The considered models describe biological signaling pathways, namely, the JAK-STAT (Swameye et al., 2003; Bachmann et al., 2011) and the RAF/MEK/ERK signaling pathway (Fiedler et al., 2016).

### 3.1.1 JAK-STAT signaling I

The first application example we considered is the model of Epo-induced JAK-STAT signaling introduced by Swameye et al. (2003) (Figure 3A). Epo yields the phosphorylation of signal transducer and activator of transcription 5 (STAT5), which dimerizes, enters the nucleus to trigger the transcription of target genes, gets dephosphorylated, and is transported to the cytoplasm. We implemented the model which describes the phosphorylated Epo receptor concentration as a time-dependent spline (Schelker et al., 2012). For further details on the model, we refer to Supplementary Information, Section 4.1.

The model parameters were estimated using immunoblotting data for the phosphorylated Epo receptor (pEpoR), phosphorylated STAT5 (pSTAT5), and the total amount of STAT5 in the cytoplasm (tSTAT5) (Figure 3B). Experimental data are available for 16 different time points. Since immunoblotting only provides relative data, the scaling parameters for the observables need to be estimated from the data. As proposed by Schelker et al. (2012), the scaling parameter for pEpoR has been fixed to avoid structural non-identifiabilities (Raue et al., 2009). This yields $n_\theta = 11$ dynamic parameters (see Supplementary Information, Section 4.1), $n_s = 2$ scaling parameters, and $n_\sigma = 3$ noise parameters.

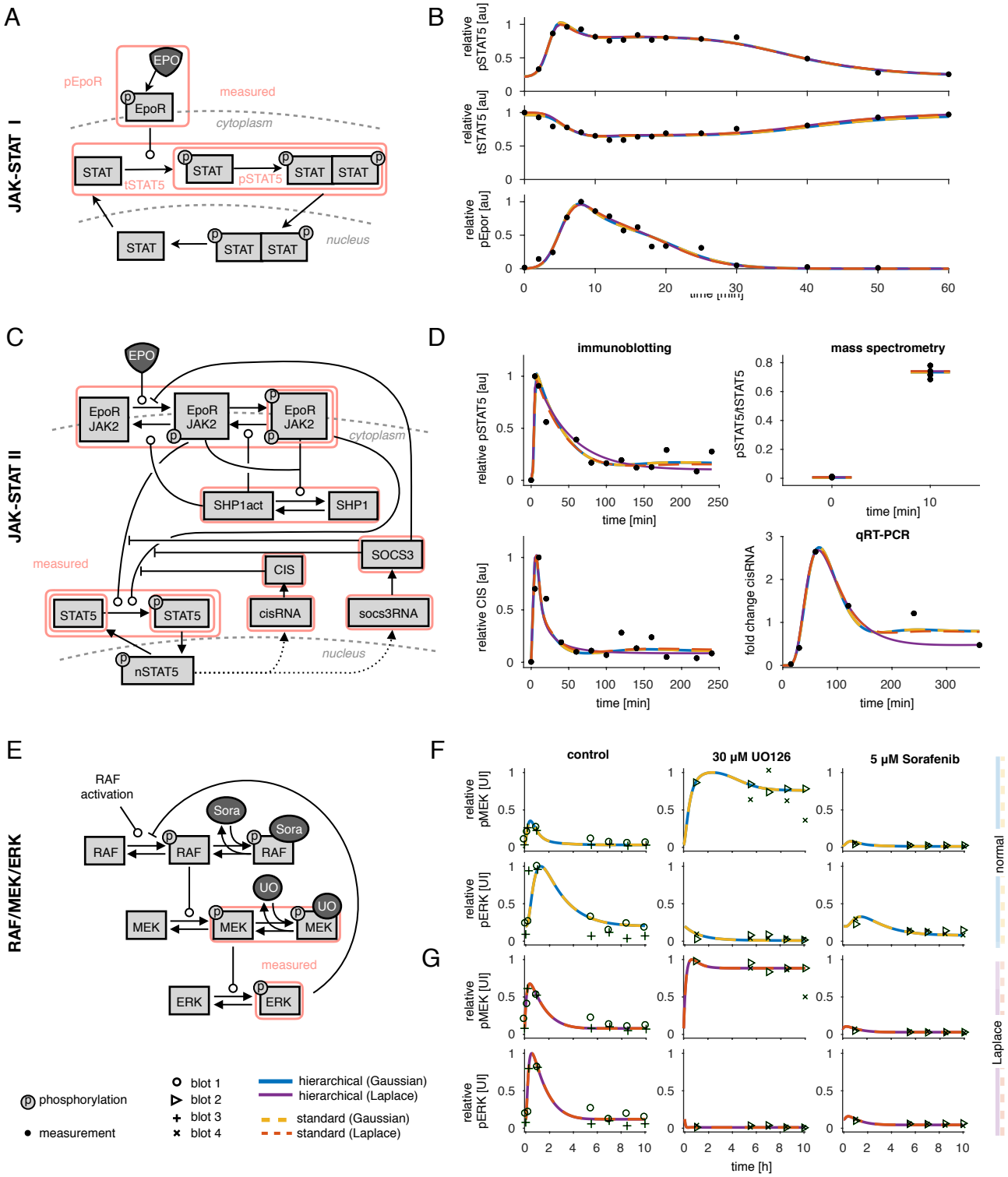



**Figure 3:** Models and experimental data. (**A,B**) JAK-STAT I. (**A**) Illustration of the model according to Swameye et al. (2003). Arrows represent biochemical reactions, and the observables of the model used are highlighted by boxes. (**B**) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization with 100 starts. The results are shown for the standard (dotted lines) and hierarchical (solid lines) approach for optimization for Gaussian and Laplace noise. (**C,D**) JAK-STAT II. (**C**) Illustration of the model according to Bachmann et al. (2011). (**D**) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization for 200 starts. 33 out of 541 data points are shown. (**E-G**) RAF/MEK/ERK. (**E**) Illustration of the model according to Fiedler et al. (2016). (**F,G**) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization for 500 starts. Different markers indicate the different blots. The data is scaled according to the estimated scaling parameters, yielding different visualizations for different parameters, as obtained with the Gaussian and the Laplace noise assumption. (**F**) Fitted trajectories for Gaussian noise for the standard (dotted line) and hierarchical (solid line) approach for optimization. (**G**) Fitted trajectories for Laplace noise.

#### 3.1.2 JAK-STAT signaling II

The second application example is the model of JAK-STAT signaling introduced by Bachmann et al. (2011). This model provides more details compared to the previous one. It includes, for instance, gene expression of cytokine-inducible SH2-containing protein (CIS) and suppressor of cytokine signaling 3 (SOCS3), and possesses more state variables and parameters (Figure 3C).

The model parameters were estimated using immunoblotting, qRT-PCR, and quantitative mass spectrometry data (Figure 3D and Supplementary Information, Figure S4). To model the observables Bachmann et al. (2011) used $n_s = 43$ scaling parameters, and $n_\sigma = 11$ noise parameters, yielding $n_\theta = 58$ remaining parameters. Some scaling and noise parameters are shared between experiments and some are shared between observables. For this model, most of the observables were compared at the $\log_{10}$ scale (see Supplementary Information, Section 4.2).

#### 3.1.3 RAF/MEK/ERK signaling

The third application example we considered is the model of RAF/MEK/ERK signaling introduced by Fiedler et al. (2016). The model describes the phosphorylation cascade and a negative feedback of phosphorylated ERK on RAF phosphorylation (Figure 3E).

Fiedler et al. (2016) collected Western blot data for HeLa cells for two observables, phosphorylated MEK, and phosphorylated ERK, with four replicates at seven time points (Figure 3F,G). Each observable and replicate was assumed to have different scaling and noise parameters, yielding 16 additional parameters (Figure 4A).

### 3.2 Evaluation of the approaches

We performed parameter estimation for the application examples using the standard and the hierarchical approach. For each example, the case of Gaussian and Laplace noise was considered. The resulting optimization problems were solved with the MATLAB toolbox PESTO (Stapor et al., 2017), using multi-start local optimization, an approach which was previously found to be computationally efficient and reliable (Raue et al., 2013). Initial points were sampled uniformly within their parameter boundaries and local optimization was performed using the interior point method implemented in the MATLAB function `fmincon.m`. Numerical simulation and forward sensitivity analysis for gradient evaluation was performed using the MATLAB toolbox AMICI (Fröhlich et al., 2017), which provides an interface to CVODES (Serban and Hindmarsh, 2005). To improve convergence and computational efficiency, $\log_{10}$-transformed parameters were used for the optimization.

#### 3.2.1 Qualitative comparison of optimization approaches for different noise distributions

As the standard and hierarchical approach should in principle be able to achieve the same fit, we first studied the agreement of trajectories for the optimal parameters. We found that they coincide for the JAK-STAT model I and the RAF/MEK/ERK model, indicating that the hierarchical approach is able to find the same optimal value as the standard approach (Figure 3B,F,G). Also the best likelihood values which were found for these two models by the two approaches coincide (Figure 4B and Supplementary Information, Figure S5). Only for the JAK-STAT model II for the case of Laplace noise, the fitted trajectories deviate (Figure 3D). Insertion of the optimum found by the hierarchical approach in the objective function of the standard approach revealed that the standard approach missed the optimal point (Supplementary

Information, Figure S3). As expected, there are differences between the results obtained with Gaussian and Laplace noise, which is visible in the trajectories and the corresponding likelihood values. Interestingly, for each model the likelihood values achieved using Laplace noise were better than for Gaussian noise (Supplementary Information, Figure S1C). This indicates that the Laplace distribution with its heavier tail is more appropriate than the Gaussian distribution for the considered estimation problems.

### 3.2.2 Convergence of optimizers

As the performance of multi-start local methods depends directly on the convergence of the local optimizers, we assessed for how many starting points the local optimizer reached the best objective function value found across all runs. This was done by studying the likelihood waterfall plots (Figure 4B). We found that the proposed hierarchical approach achieved consistently a higher fraction of converged starts than the standard approach (Figure 4C). Local optimization using the hierarchical approach converged on average in 25.38% of the runs while the standard approach converged on average in 12.13% of the runs.

The application examples vary with respect to the total number of parameters and in the number of parameters which correspond to scaling or noise parameters (Figure 4A). While for the JAK-STAT model I only five parameters could be optimized analytically, for the JAK-STAT model II almost half of the parameters correspond to scaling or noise parameters. Interestingly, even when the dimension of the optimization problem was only reduced by few parameters, we observed a substantial improvement of the convergence (Figure 4C).

### 3.2.3 Computational efficiency

As computation resources are often limiting, we finally analyzed the computation time per converged start. We found that on average, the computation time per start was lower for the hierarchical approach than for the standard approach (Figure 4D). In combination with the improved convergence rate, this resulted in a substantially reduced computation time per converged start, aka a start which reach the minimal value observed across all starts (Figure 4E). Given a fixed computational budget, the hierarchical approach achieved on average 5.52 times more optimization runs which reached the best objective function values than the standard approach.

In summary, the application of our hierarchical approach to parameter estimation from relative data to the models shows consistently that our approach yields parameter values of the same quality as the standard method, while achieving better convergence and reducing the computation time substantially.

## 4 Conclusion

The statistically rigorous estimation of model parameters from relative data requires non-standard statistical models (Degasperi et al., 2017) or scaling parameters (Raue et al., 2013). Unfortunately, the former is not supported by established toolboxes and the latter increases the dimensionality of the estimation problem. In this manuscript, we introduced a hierarchical approach which avoids the increase of dimensionality and is applicable to a broad range of noise distributions. For Gaussian and Laplace noise we provided analytic expressions. The approach can be used for combinations of relative and absolute data, and for different optimization methods, including least-squares methods or global optimization methods such as particle swarm optimization (Vaz and Vicente, 2009) (see Supplementary Information, Figure S2).

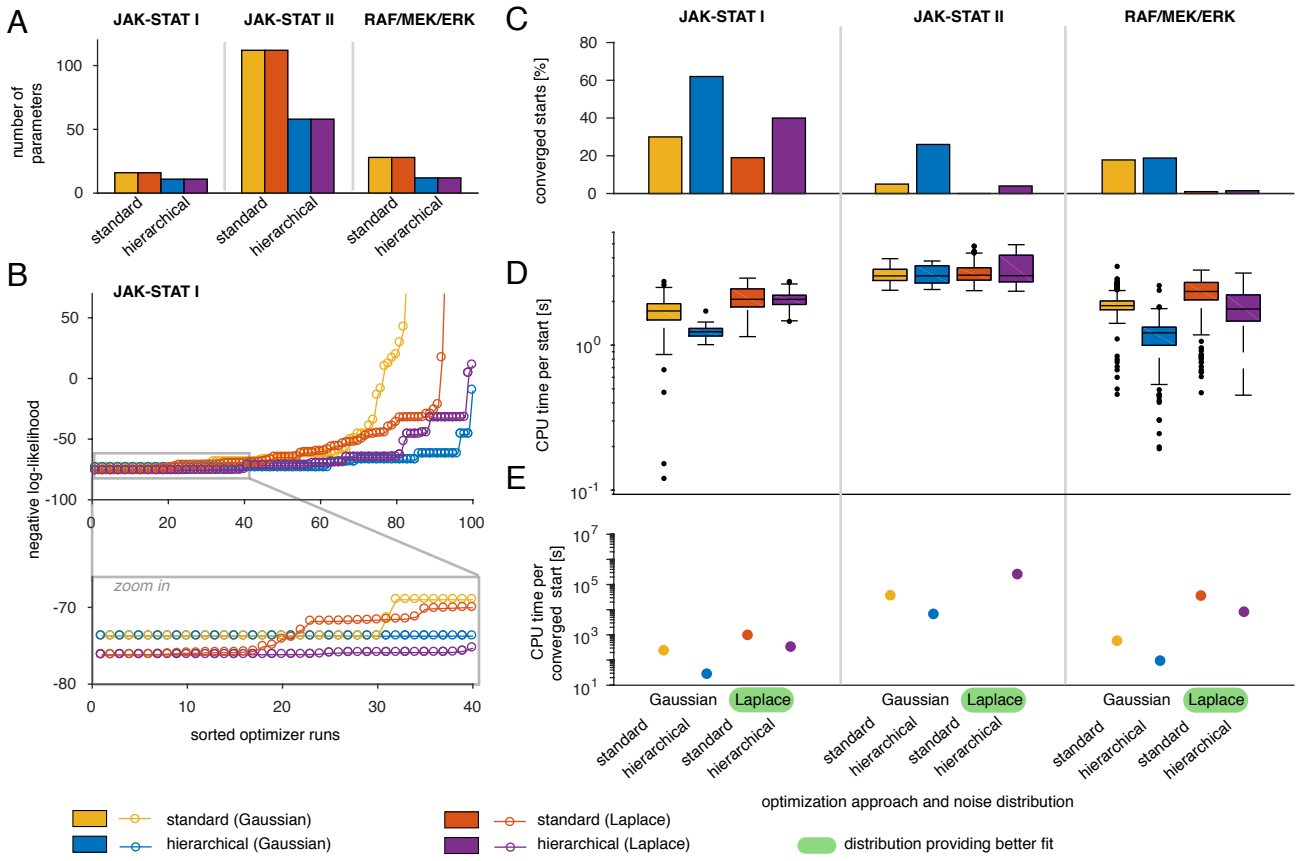



**Figure 4:** Evaluation of the standard and hierarchical approach for three application examples. (**A**) Number of parameters which need to be optimized numerically. (**B**) Likelihood waterfall plot for the JAK-STAT model I. The ascendingly sorted negative log-likelihood values are shown for both approaches (standard and hierarchical) and noise distributions (Gaussian and Laplace). (**C-E**) Comparison of the two optimization approaches and two noise distribution for the three models. The noise model with the better likelihood function is highlighted in the label. (**C**) Percentage of converged starts over all performed local optimizations. (**D**) Boxplot for the CPU time needed per start. (**E**) CPU time needed per converged start.

We evaluated the performance of our hierarchical approach and compared it to the standard approach for three models, which vary in their complexity. For all applications, we found that our hierarchical approach yielded fits of the same or better quality. In addition, convergence was improved and the computation time was shortened substantially. We demonstrated that our approach can also be used when relative and absolute data are modeled together in an experiment, and when several observables or experiments share scaling and/or noise parameters. This renders our approach applicable to a wide range of mathematical models studied in systems and computational biology. We provided a generic implementation of the objective function for the hierarchical approach for Gaussian and Laplace noise. The objective function is provided in the Supplementary Information (along with the rest of the code) and included in the MATLAB toolbox PESTO (Stapor et al., 2017).

In addition to the scaling and noise parameters, also other parameters which only contribute to the mapping from the states to the observables, could be optimized analytically. This includes offset parameters, which are used to model background intensities or unspecific binding. Extending our approach to also calculate these parameters analytically would decrease the parameters in the outer optimization even more.

We employed forward sensitivities for the calculation of the objective function gradient. However, it has been shown that for large-scale models with a high number of parameters, adjoint sensitivities can reduce the computation time needed for simulation (Fröhlich et al., 2017). Thus, a further promising approach

would be the combination of both complementary approaches for the handling of large-scale models.

To summarize, employing our hierarchical approach for optimization yielded more robust results and speed up the computation time. This renders the approach valuable for estimating parameters from relative data. The proposed approach might facilitate the handling of large-scale models, which possess many measurement parameters.

## Funding

This work has been supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 686282

## References

J. Bachmann, A. Raue, M. Schilling, M. E. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. D. Lehmann, J. Timmer, and U. Klingmüller. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, 7(1):516, July 2011

R. J. De Boer, V. V. Ganusov, D. Milutinoviò, P. D. Hodgkin, and A. S. Perelson. Estimating lymphocyte division and death rates from CFSE data. *Bull. Math. Biol.*, 68(5):1011–1031, 2006

A. Degasperi, D. Fey, and B. N. Kholodenko. Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *npj Syst Biol Appl*, 3(1):20, 2017.

A. Fiedler, S. Raeth, F. J. Theis, A. Hausser, and J. Hasenauer. Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints. *BMC Syst. Biol.*, 10(80), Aug. 2016.

F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, 13(1):e1005331, Jan. 2017.

L. A. Herzenberg, J. Tung, W. A. Moore, L. A. Herzenberg, and D. R. Parks. Interpreting flow cytometry data: A guide for the perplexed. *Nat. Immunol.*, 7(7):681–685, July 2006.

H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, Mar. 2002

E. Klipp, B. Nordlander, R. Krüger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nat. Biotechnol.*, 23(8):975–982, Aug 2005

C. Maier, C. Loos, and J. Hasenauer. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725, Mar. 2017.

E. J. Molinelli, A. Korkut, W. Wang, M. L. Miller, N. P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D. B. Solit, C. A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina, and C. Sander. Perturbation biology: Inferring signaling networks in cellular systems. *PLoS Comput. Biol.*, 9(12):e1003290, Dec. 2013.

R. M. Norton. The double exponential distribution: Using calculus to find a maximum likelihood estimator. *Am. Stat.*, 38(2):135–136, May 1984

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(25):1923–1929, May 2009

A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335, Sept. 2013.

J. Renart, J. Reiser, and G. R. Stark. 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(7):3116–3120, July 1979

M. Schelker, A. Raue, J. Timmer, and C. Kreutz. Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinformatics*, 28(18):i529–i534, 2012.

B. Schöberl, E. A. Pace, J. B. Fitzgerald, B. D. Harms, L. Xu, L. Nie, B. Linggi, A. Kalra, V. Paragas, R. Bukhalid, V. Grantcharova, N. Kohli, K. A. West, M. Leszczyniecka, M. J. Feldhaus, A. J. Kudla, and U. B. Nielsen. Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Science Signaling*, 2(77):ra31, 2009

R. Serban and A. C. Hindmarsh. CVODES: An ODE solver with sensitivity analysis capabilities. *ACM Math. Software*, 31(3):363–396, 2005

P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich, and J. Hasenauer. PESTO: Parameter EStimation TOolbox. *Bioinformatics*, btx676, 2017.

I. Swameye, T. G. Müller, J. Timmer, O. Sandra, and U. Klingmüller. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. USA*, 100 (3):1028–1033, Feb 2003. URL `http://www.pnas.org/content/100/3/1028.abstract`

C. Thomaseth and N. Radde. Normalization of western blot data affects the statistics of estimators. *IFAC-PapersOnLine*, 49(26):56–62, 2016

A. I. F. Vaz and L. N. Vicente. PSwarm: A hybrid solver for linearly constrained global derivative-free optimization. *Optim. Method. Softw.*, 24(4-5):669–685, 2009.

P. Weber, J. Hasenauer, F. Allgöwer, and N. Radde. Parameter estimation and identifiability of biological networks using relative data. In S. Bittanti, A. Cenedese, and S. Zampieri, editors, *Proc. of the 18th IFAC World Congress*, volume 18, pages 11648–11653, Milano, Italy, Aug. 2011.

Supplementary Information

# Hierarchical optimization for the efficient parametrization of ODE models

Carolin Loos [1,*], Sabrina Krause [1,*], and Jan Hasenauer [1,†]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany, and Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany,

*C.L. and S.K. contributed equally to this work.

†To whom correspondence should be addressed.

## Contents

**1 General formula for analytic scaling and noise parameters** **2**

1.1 Gaussian noise . . . 2

1.2 Laplace noise . . . 4

**2 Comparison of data and simulation at a logarithmic scale** **5**

2.1 Gaussian noise . . . 5

2.2 Laplace noise . . . 6

**3 Implementation** **7**

**4 Models and experimental data** **7**

4.1 JAK-STAT signaling I . . . 7

4.2 JAK-STAT signaling II . . . 10

4.3 RAF/MEK/ERK signaling . . . 18

# 1 General formula for analytic scaling and noise parameters

In the main manuscript, we covered experimental data sets which have different time points. Here, we provide the derivation of the expressions for the general case, in which the experimental data also comprise different replicates, experiments, and conditions, e.g., varying drug doses. We considered that the ODE system also depends on an input $\mathbf{u} \in \mathbb{R}^{n_u}$,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}), \quad \mathbf{x}(t_0, \boldsymbol{\theta}, \mathbf{u}) = x_0(\boldsymbol{\theta}, \mathbf{u}) \,, \tag{1}$$

thus, $f \colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$, which also affects the mapping to the observables

$$\mathbf{y}(t, \boldsymbol{\theta}, \mathbf{u}) = \mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}) \,. \tag{2}$$

The experimental data is then given by

$$\mathcal{D} = \left\{ \left\{ \left\{ \{ \bar{\mathbf{y}}_{k,r,c_e}, t_{k,c_e}, \mathbf{u}_{c_e} \}_k \right\}_r \right\}_{c_e \in I_e} \right\}_e \,, \tag{3}$$

including all indices for time point $k$, replicate $r$, experiment-specific condition $c_e$, and experiment $e$. The indices $I_e$ indicate which conditions correspond to a certain experiment. The measurements are mapped to the states by

$$\bar{y}_{i,k,r,c_e} = s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e}) + \varepsilon_{i,k,r,c_e} \,,$$

with $\varepsilon_{i,k,r,c_e} \sim \mathcal{N}(0, \sigma^2_{i,r,c_e})$ or $\varepsilon_{i,k,r,c_e} \sim \text{Laplace}(0, \sigma_{i,r,c_e})$, and $s_{i,r,c_e} = 1$ for absolute measurements. Also, the structure of the mapping from states to observables might be experiment-specific. The negative log-likelihood is given by

$$J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = \sum_{e,i,k,r} \sum_{c_e \in I_e} \log p(\bar{y}_{i,k,r,c_e} | s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e}), \sigma_{i,r,c_e}) \,. \tag{4}$$

In the main manuscript, we presented the analytic formulas for the case that each observable and corresponding replicate has different scaling and noise parameters, but that these parameters do not change between conditions and time points. A more general formula is provided in the following, covering, e.g., the case that replicates share the same scaling parameters, but observables do not. This can be easily generalized to also include variability between time points.

## 1.1 Gaussian noise

The general objective function under Gaussian noise is given by

$$J(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\sigma}) = \frac{1}{2} \sum_{i,r,k,e} \sum_{c_e \in I_e} \log(2\pi {\sigma_{i,r,c_e}}^2) + \left( \frac{\bar{y}_{i,k,r,c_e} - s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\sigma_{i,r,c_e}} \right)^2 .$$

To define which replicates, observables, and experiments share a scaling or noise parameter, we define

$$I_s^{i_s}, I_\sigma^{i_\sigma} \subset \mathbb{N}_+^{n_y} \times \mathbb{N}_+^{n_r} \times \mathbb{N}_+^{n_e} \,,$$

for $i_s = 1, \dots, n_s$ and $i_\sigma = 1, \dots, n_\sigma$. The number of replicates is denoted by $n_r$ and the number of experiments by $n_e$. This means, all scaling parameters $s_{i^*,r^*,c_e^*}$ for which the indices $(i^*, r^*, e^*)$ are part of the same group $I_s$ share the same scaling parameters. This yields $n_s$ different scaling parameters that are estimated from the data. For this we denote $I_s^{i_s}(i^*, r^*, c_e^*)$ the group which includes the indices $(i^*, r^*, c_e^*)$. This is analogously for the noise parameters. The derivative of the objective function with respect to a scaling parameter thus reads

$$\frac{\partial J}{\partial s_{i^*,r^*,c_e^*}} = \frac{1}{2} \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_s^{i_s}(i^*,\, r^*,\, c_e^*)}} \frac{2}{\sigma_{i,r,c_e}^2} (\bar{y}_{i,k,r,c_e} - s_{i^*,r^*,c_e^*} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})) \cdot (-h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})) \stackrel{!}{=} 0\,, \tag{5}$$

and was set to zero to obtain the analytic expression for the optimal scaling parameter. The solution does not depend on the noise parameters if $I_s^{i_s} \subset I_\sigma^{i_\sigma}\ \forall i_s$, and we solve the equation with respect to $s_{i^*,r^*,c_e^*}$ to obtain the optimal value

$$\hat{s}_{i^*,r^*,c_e^*} = \frac{\sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_s^{i_s}(i^*,\, r^*,\, c_e^*)}} \bar{y}_{i,k,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_s^{i_s}(i^*,\, r^*,\, c_e^*)}} h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})^2}\,.$$

For the noise parameters, we need

$$\frac{\partial J}{\partial \sigma_{i^*,r^*,c_e^*}^2} = \frac{1}{\sigma_{i^*,r^*,c_e^*}^2} \cdot \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} 1 - \left( \frac{\bar{y}_{i,k,r,c_e} - s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\sigma_{i^*,r^*,c_e^*}} \right)^2 \stackrel{!}{=} 0\,. \tag{6}$$

We write

$$\begin{aligned}
&\frac{1}{\sigma_{i^*,r^*,c_e^*}^2} \cdot \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} 1 - \left( \frac{\bar{y}_{i,k,r,c_e} - s_{i,r,c_e} h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\sigma_{i^*,r^*,c_e^*}} \right)^2 = 0 \\
\Leftrightarrow\quad & \sigma_{i^*,r^*,c_e^*}^2 \cdot \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} 1 = \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} (\bar{y}_{i,k,r,c_e} - s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e}))^2 \\
\Rightarrow\quad & \hat{\sigma}_{i^*,r^* c_e^*}^2 = \underbrace{\frac{1}{\sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} 1}}_{(\dagger)} \sum_{k,e} \sum_{\substack{(i,\, r,\, c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\, r^*,\, c_e^*)}} (\bar{y}_{i,k,r,c_e} - s_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e}))^2\,,
\end{aligned}$$

in which (†), the nominator, is simply the number of observations in which $\sigma_{i^*,r^*,c_e^*}$ appears. In some cases, for instance if all experiments share the same scaling parameter, we neglected the superscript $e$.

The gradient used for optimization is given by

$$\frac{\partial J}{\partial \theta} = \sum_{i,r,k,e} \sum_{c_e \in I_e} \frac{\bar{y}_{i,k,r,c_e} - \hat{s}_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\hat{\sigma}_{i,r,c_e}^2} \cdot \hat{s}_{i,r,c_e} \cdot \frac{\partial h_i^e(\mathbf{x}(t_{k,c_e}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{u}_{c_e})}{\partial \theta}\,,$$

for which $\frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial\theta}$ is obtained by forward sensitivity equations employed in AMICI, and,

$$\frac{\partial J}{\partial s}=0,\quad \frac{\partial J}{\partial \sigma}=0\,,$$

which holds due to (5) and (6). The Hessian with respect to the dynamic parameters is

$$\frac{\partial^2 J}{\partial\theta_j\partial\theta_l}=\sum_{i,r,k,e}\sum_{c_e\in I_e}\left(\frac{\hat{s}_{i,r,c_e}}{\hat{\sigma}_{i,r,c_e}}\right)^2\cdot\frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial\theta_j}\cdot\frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial\theta_l}+\underbrace{\frac{\bar{y}_{i,k,r,c_e}-\hat{s}_{i,r,c_e}\cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\hat{\sigma}^2_{i,r,c_e}}\cdot\hat{s}_{i,r,c_e}\cdot\frac{\partial^2 h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial\theta_j\partial\theta_l}}_{(*)}\,.$$

For the remaining parameter, the Hessian is zero. We implemented an approximation of the Hessian neglecting the terms $(*)$ that include higher-order sensitivities.

## 1.2 Laplace noise

For Laplace noise, the expression for the optimal scaling and noise parameters can be generalized analogously. The objective function for the general case is

$$J(\boldsymbol{\theta},\mathbf{s},\boldsymbol{\sigma})=\sum_{i,k,r,e}\sum_{c_e\in I_e}\log(2\sigma_{i,r,c_e})+\frac{|\bar{y}_{i,k,r,c_e}-s_{i,r,c_e}\cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})|}{\sigma_{i,r,c_e}}\,.$$

The derivative with respect to a scaling parameter is

$$\frac{\partial J}{\partial s_{i^*,r^*,c_e^*}}=\sum_{k,e}\sum_{\substack{(i,\,r,\,c_e)\,\in\\ I_s^{is}(i^*,\,r^*,\,c_e^*)}}\frac{1}{\sigma_{i,r,c_e}}\cdot\left(|h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})|\cdot\operatorname{sgn}\left(\frac{\bar{y}_{i,k,r,c_e}}{h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}-s_{i^*,r^*,c_e^*}\right)\right)$$

with jump points

$$\left\{\left\{s_{i,k,r,c_e}=\frac{\bar{y}_{i,k,r,c_e}}{h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}\right\}_{(i,r,c_e)\in I_s^{is}|_{(i^*,r^*,c_e^*)\in I_s^{is}}}\right\}_{k,e}\,. \tag{7}$$

These jump points are the candidates for the optimal scaling parameter and the candidate for which the sign of the derivative changes is chosen. For the optimal noise parameter we have

$$\frac{\partial J}{\partial\sigma_{i^*,r^*,c_e^*}}=\frac{1}{\sigma_{i^*,r^*,c_e^*}}\cdot\sum_{k,e}\sum_{\substack{(i,\,r,\,c_e)\,\in\\ I_\sigma^{i\sigma}(i^*,\,r^*,\,c_e^*)}}\left(1-\frac{|\bar{y}_{i,k,r,c_e}-\hat{s}_{i,r,c_e}\cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})|}{\sigma_{i^*,r^*,c_e^*}}\right)\overset{!}{=}0 \tag{8}$$

$$\hat{\sigma}_{i^*,r^*,c_e^*} = \frac{1}{\sum_{k,e} \sum_{\substack{(i,\,r,\,c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\,r^*,\,c_e^*)}} 1} \cdot \sum_{k,e} \sum_{\substack{(i,\,r,\,c_e) \in \\ I_\sigma^{i_\sigma}(i^*,\,r^*,\,c_e^*)}} \left( |h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})| \cdot \left| \frac{\bar{y}_{i,k,r,c_e}}{h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})} - \hat{s}_{i^*,r^*,c_e^*} \right| \right).$$

The gradient used for optimization is given by

$$\frac{\partial J}{\partial \theta} = -\sum_{i,r,k,e} \sum_{c_e \in I_e} \frac{\operatorname{sgn}\left(\bar{y}_{i,k,r,c_e} - \hat{s}_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})\right)}{\hat{\sigma}_{i,r,c_e}} \cdot \left( \hat{s}_{i,r,c_e} \cdot \frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial \theta} + h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e}) \cdot \frac{\partial \hat{s}_{i,r,c_e}}{\partial \theta} \right),$$

for which $\frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial \theta}$ is obtained by forward sensitivity equations employed in AMICI, and $\frac{\partial J}{\partial \sigma} = 0$, which holds due to (8).

## 2 Comparison of data and simulation at a logarithmic scale

In the main manuscript and Supplementary Information, Section 1, we provided the formulas for the comparison of data and simulation on a linear scale. However, sometimes it might be more appropriate to compare experimental data and simulation on a logarithmic scale.

### 2.1 Gaussian noise

For Gaussian noise, the objective function for the comparison on the logarithmic scale is given by

$$J(\boldsymbol{\theta},\mathbf{s},\boldsymbol{\sigma}) = \frac{1}{2} \sum_{i,k,r,e} \sum_{c_e \in I_e} \log(2\pi {\sigma_{i,r,c_e}}^2) + \left( \frac{\log(\bar{y}_{i,k,r,c_e}) - \log\left(s_{i,r,c_e} \cdot h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right)\right)}{\sigma_{i,r,c_e}} \right)^2$$

Thus, the derivative with respect to the scaling parameters is

$$\frac{\partial J}{\partial s_{i^*,r^*,c_e^*}} = \frac{1}{2} \sum_{k,e} \sum_{\substack{(i,\,r,\,c_e) \in \\ I_s^{i_s}(i^*,\,r^*,\,c_e^*)}} \frac{2\left(\log(\bar{y}_{i,k,r,c_e}) - \log(s_{i^*,r^*,c_e^*}) - \log(h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right))\right) \frac{1}{s_{i^*,r^*,c_e^*}}}{{\sigma_{i,r}^e}^2}.$$

This yields the formula for the optimal scaling parameters

$$\hat{s}_{i^*,r^*,c_e^*} = \exp\left( \frac{\sum_{k,e} \sum_{\substack{(i,\,r,\,c_e) \in \\ I_s^{i_s}(i^*,\,r^*,\,c_e^*)}} \log(\bar{y}_{i,k,r,c_e}) - \log(h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right))}{\sum_{k,e} \sum_{\substack{(i,\,r,\,c_e) \in \\ I_s^{i_s}(i^*,\,r^*,\,c_e^*)}} 1} \right) \tag{9}$$

and

$$\hat{\sigma}^2_{i^*,r^*,c_e^*} = \frac{1}{\sum\limits_{k,e} \sum\limits_{\substack{(i,\,r,\,c_e)\,\in \\ I^{i\sigma}_{\sigma}(i^*,r^*,c_e^*)}} 1} \cdot \sum_{k,e} \sum_{\substack{(i,\,r,\,c_e)\,\in \\ I^{i\sigma}_{\sigma}(i^*,r^*,c_e^*)}} \left(\log(\bar{y}_{i,k,r,c_e}) - \log(\hat{s}_{i^*,r^*,c_e^*} \cdot h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right)\right)^2 . \tag{10}$$

The gradient used for optimization is given by

$$\frac{\partial J}{\partial \theta} = \sum_{i,r,k,e} \sum_{c_e \in I_e} 2 \cdot \frac{\log(\bar{y}_{i,k,r,c_e}) - \log(\hat{s}_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e}))}{\hat{\sigma}^2_{i,r,c_e}} \cdot \frac{1}{h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})} \cdot \frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial \theta} .$$

If the data is compared at $\log_{10}$ scale, as, e.g., for the JAK-STAT signaling model proposed by Bachmann et al. (2011), the negative log-likelihood function reads

$$J(\boldsymbol{\theta},\mathbf{s},\boldsymbol{\sigma}) = \frac{1}{2} \sum_{i,k,r,e} \sum_{c_e \in I_e} \log(2\pi {\sigma_{i,r,c_e}}^2) + \left( \frac{\log_{10}(\bar{y}_{i,k,r,c_e}) - \log_{10}\left(s_{i,r,c_e} \cdot h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right)\right)}{\sigma_{i,r,c_e}} \right)^2 .$$

The optimal scaling parameters here are the same as when using the natural logarithm (9). For the optimal noise parameters the log is replaced by $\log_{10}$ in (10).

## 2.2 Laplace noise

For the Laplace distribution including the logarithmic comparison

$$J(\boldsymbol{\theta},\mathbf{s},\boldsymbol{\sigma}) = \sum_{i,k,r,e} \sum_{c_e \in I_e} \log(2\sigma_{i,r,c_e}) + \frac{\left|\log(\bar{y}_{i,k,r,c_e}) - \log\left(s_{i,r,c_e} \cdot h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right)\right)\right|}{\sigma^e_{i,r}}$$

the same procedure can be applied for the logarithmic scale as for the linear scale, with the same set of candidate scaling parameters (7) as for the linear scale. However, one has to pay attention to adapt the derivative properly, for which the change of signs is checked. The optimal noise parameters then is given by

$$\hat{\sigma}_{i^*,r^*,c_e^*} = \frac{1}{\sum\limits_{k,e} \sum\limits_{\substack{(i,\,r,\,c_e)\,\in \\ I^{i\sigma}_{\sigma}(i^*,r^*,c_e^*)}} 1} \cdot \sum_{k,e} \sum_{\substack{(i,\,r,\,c_e)\,\in \\ I^{i\sigma}_{\sigma}(i^*,r^*,c_e^*)}} \left( \left|\log\left(h_i^e\left(\mathbf{x}\left(t_{k,c_e},\boldsymbol{\theta}\right),\boldsymbol{\theta},\mathbf{u}_{c_e}\right)\right)\right| \cdot \left| \frac{\log(\bar{y}_{i,k,r,c_e})}{\log(h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e}))} - \hat{s}_{i^*,r^*,c_e^*} \right| \right).$$

The gradient used for optimization is given by

$$\frac{\partial J}{\partial \theta} = -\sum_{i,r,k,e} \sum_{c_e \in I_e} \frac{\operatorname{sgn}\left(\log(\bar{y}_{i,k,r,c_e}) - \log(\hat{s}_{i,r,c_e} \cdot h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e}))\right)}{\hat{\sigma}_{i,r,c_e}} \cdot \left( \frac{1}{h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})} \cdot \frac{\partial h_i^e(\mathbf{x}(t_{k,c_e},\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{u}_{c_e})}{\partial \theta} + \frac{1}{\hat{s}_{i,r,c_e}} \cdot \frac{\partial \hat{s}_{i,r,c_e}}{\partial \theta} \right) .$$

# 3 Implementation

We implemented the log-likelihood function and the analytic calculation of the scaling and noise parameters in easy-to-use MATLAB functions. The log-likelihood function is provided in `loglikelihoodHierarchical.m`, which provides the log-likelihood value, the gradient of the log-likelihood function with respect to the dynamic parameters, and in the case of Gaussian noise also an approximation to the Hessian by neglecting second-order derivatives. The functions and examples are incorporated in the toolbox PESTO (Stapor et al., 2017)and can be found on GitHub: `http://github.com/ICB-DCM/PESTO`. The simulated observables, their sensitivities, the experimental data, and the specification of measurement noise, scale of comparison between simulation and data, and shared parameters needs to be supplied by the user.

For our analysis, we employed the toolbox AMICI (Fröhlich et al., 2017) for the simulation of the system and the simulation of the sensitivities, and the toolbox PESTO (Stapor et al., 2017) for the estimation of the parameters.

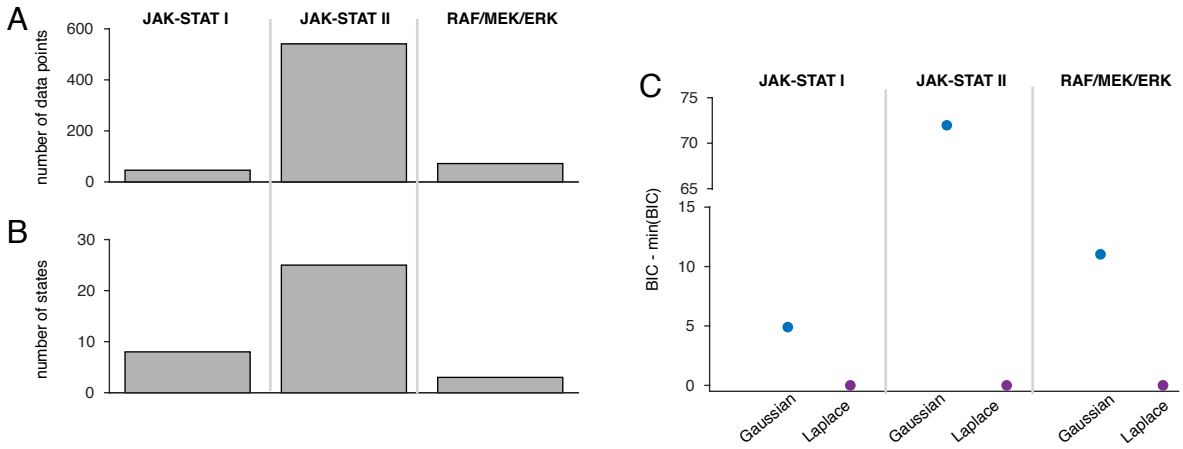
# 4 Models and experimental data

In the following, we provide the details of the mathematical models. The considered models vary in their number of parameters (Figure 6A), number of data points that are used to calibrate the models (Figure S1A), and number of states of the underlying ODE system (Figure S1B).

## 4.1 JAK-STAT signaling I

For the first model, we used the model introduced by Schelker et al. (2012), which is defined by the ODE system

$$
\begin{aligned}
\frac{\partial[\mathrm{STAT}]}{\partial t} &= \frac{1}{\Omega_{\mathrm{cyt}}}\left(\Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT5}]\,p_4 - \Omega_{\mathrm{cyt}}\,[\mathrm{STAT}]\,p_1\,g\right)\\
\frac{\partial[\mathrm{pSTAT}]}{\partial t} &= -2\,p_2\,[\mathrm{pSTAT}]^2 - [\mathrm{STAT}]\,p_1\,g\\
\frac{\partial[\mathrm{pSTAT\_pSTAT}]}{\partial t} &= p_2\,[\mathrm{pSTAT}]^2 - p_3\,[\mathrm{pSTAT\_pSTAT}]\\
\frac{\partial[\mathrm{nSTAT1}]}{\partial t} &= -\frac{p_4}{\Omega_{\mathrm{nuc}}}\left(\Omega_{\mathrm{cyt}}\,[\mathrm{STAT}] - \Omega_{\mathrm{cyt}}\,[\mathrm{STAT}]_0 + 2\,\Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT1}]\right.\\
&\quad + \Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT2}] + \Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT3}] + \Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT4}]\\
&\quad \left. + \Omega_{\mathrm{nuc}}\,[\mathrm{nSTAT5}] + \Omega_{\mathrm{cyt}}\,[\mathrm{pSTAT}] + 2\,\Omega_{\mathrm{cyt}}\,[\mathrm{pSTAT\_pSTAT}]\right)\\
\frac{\partial[\mathrm{nSTAT2}]}{\partial t} &= p_4\,([\mathrm{nSTAT1}] - [\mathrm{nSTAT2}])\\
\frac{\partial[\mathrm{nSTAT3}]}{\partial t} &= p_4\,([\mathrm{nSTAT2}] - [\mathrm{nSTAT3}])\\
\frac{\partial[\mathrm{nSTAT4}]}{\partial t} &= p_4\,([\mathrm{nSTAT3}] - [\mathrm{nSTAT4}])\\
\frac{\partial[\mathrm{nSTAT5}]}{\partial t} &= p_4\,([\mathrm{nSTAT4}] - [\mathrm{nSTAT5}]),
\end{aligned}
$$




Supplementary Figure S1: Comparison of the three models. (A) Number of experimental data points used to calibrate the models. (B) Number of states $n_x$. (C) Comparison of Gaussian and Laplace noise for the three models based on the Bayesian Information Criterion (BIC) (Raftery, 1999), which rewards high likelihood values and penalizes high number of parameters.

with kinetic parameters $p_1, ..., p_4$. The brackets indicate the concentrations of the corresponding species. The initial conditions are given by

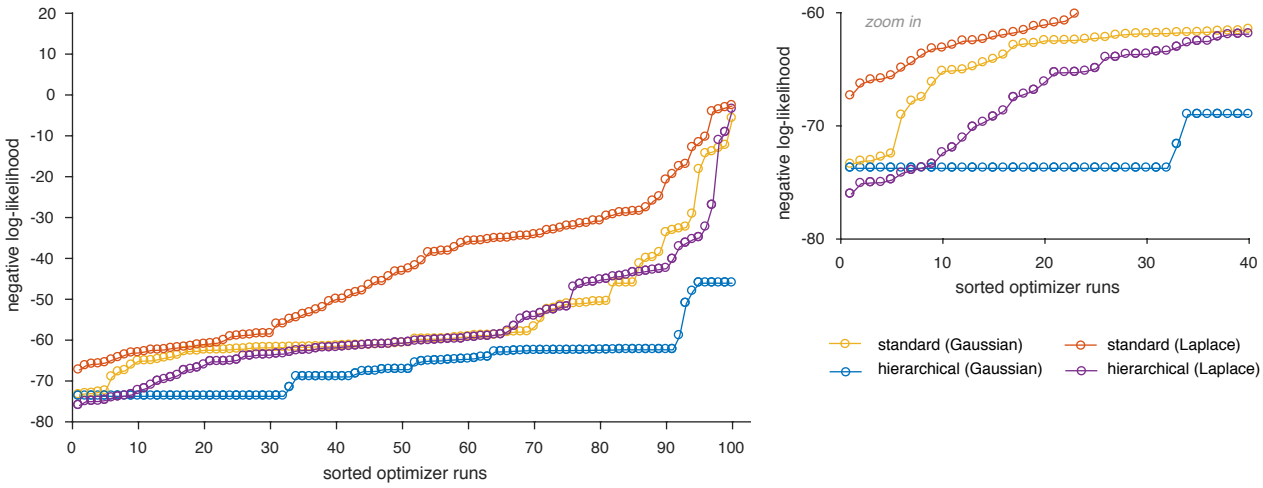
$$\mathbf{x}(t_0) = (1, [\text{pSTAT}]_0, [\text{pSTAT\_pSTAT}]_0, [\text{nSTAT1}]_0, [\text{nSTAT2}]_0, [\text{nSTAT3}]_0, [\text{nSTAT4}]_0, [\text{nSTAT5}]_0)^T ,$$

for which the initial condition for STAT is set to 1 in order to remove structural non-identifiabilities (Schelker et al., 2012). The states $\text{nSTAT1}, \ldots, \text{nSTAT5}$ are intermediate steps, resulting from a linear chain approximation to model the delay of STAT binding to the DNA in the nucleus. The volumes of the cytoplasm and nucleus are denoted by $\Omega_{\text{cyt}} = 1.4\,\text{pl}$ and $\Omega_{\text{nuc}} = 0.45\,\text{pl}$, respectively (Raue et al., 2009).

The observables are defined by $y_1$ for total concentration of phosphorylated STAT in the cytoplasm (pSTAT), $y_2$ for the total concentration of STAT in the cytoplasm (tSTAT), and $y_3$ for the phosphorylated Epo receptors (pEpoR) (see Figure 3A in the main manuscript). They are linked to the states of the system via

$$\begin{aligned} y_1 &= s_1 \left(o_1 + [\text{pSTAT}] + 2[\text{pSTAT\_pSTAT}]\right) \\ y_2 &= s_2 \left(o_2 + [\text{STAT}] + [\text{pSTAT}] + 2[\text{pSTAT\_pSTAT}]\right) \\ y_3 &= g \,. \end{aligned}$$

The concentration of Epo receptors is modeled as time-dependent cubic spline function $g$ with parameters $sp_1, \ldots, sp_5$, which are also estimated from the data. The parameters $o_1$ and $o_2$ define the offsets needed to model the background noise. The model comprises the parameters $\mathbf{q} = (p_1, p_2, p_3, p_4, sp_1, sp_2, sp_3, sp_4, sp_5, o_1, o_2, s_1, s_2, \sigma_1, \sigma_2, \sigma_3)^T$, for which $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4, sp_1, sp_2, sp_3, sp_4, sp_5, o_1, o_2)$ was optimized in the outer optimization problem of the hierarchical approach. The scaling parameters $\mathbf{s} = (s_1, s_2)$ and noise parameters $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ for observables $y_1$, $y_2$, and $y_3$, respectively, were optimized in the inner optimization problem. The subscript for these parameters indicates the observable. We neglected indices $r, e$, and $c_e$, since only one experiment, replicate, and condition is considered. The parameter boundaries for the optimization are given




Supplementary Figure S2: Likelihood waterfall plot for JAK-STAT signaling I using particle swarm optimization.

by

$$\log_{10}(\mathbf{q})_{\mathrm{lb}} = (-5, -3, -5, -3, -5, -5, -5, -5, -6, -5, -5, -5, -5, -5, -5, -5)^T$$

for the lower bound and

$$\log_{10}(\mathbf{q})_{\mathrm{ub}} = (3, 6, 3, 6, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3)^T .$$

for the upper bound (Maier et al., 2017). We performed 100 optimizations, starting from randomly drawn parameter values. The starting points for the dynamic parameters were the same for both optimization approaches.

The comparison between the two noise assumptions revealed that the Laplace noise is more appropriate. However, the difference in BIC values was below 10, which indicates that the improvement was not substantial (Figure S1C) (Kass and Raftery, 1995).

To evaluate the possibility of using the hierarchical optimization also within global optimization, we repeated the analysis using an particle swarm algorithm (Vaz and Vicente, 2009). This method does not need gradient information and has been shown to outperform other global optimization methods (Vaz and Vicente, 2009). The waterfall plots are shown in Figure S2. Interestingly, only the hierarchical optimization for the Gaussian noise was able to find the same optimum as the deterministic optimization. For the other settings the convergence suffered. However, as for the optimization with `fmincon`, the hierarchical approach was superior to the standard approach and the Laplace noise fitted the data better than the Gaussian noise.

## 4.2 JAK-STAT signaling II

The ODE system for JAK-STAT signaling model II is given by (Bachmann et al., 2011)

$$
\begin{aligned}
\frac{\partial[\text{EpoRJAK2}]}{\partial t} &= [\text{EpoRpJAK2}]\,\frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}[\text{SHP1Act}] \\
&\quad + \frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}\,[\text{SHP1Act}]\,([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]) \\
&\quad - \frac{[\text{Epo}]\cdot[\text{EpoRJAK2}]\cdot\text{JAK2ActEpo}}{[\text{SOCS3}]\cdot\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} \\
\frac{\partial[\text{EpoRpJAK2}]}{\partial t} &= \frac{[\text{Epo}]\,[\text{EpoRJAK2}]\,\text{JAK2ActEpo}}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} - \frac{[\text{EpoRpJAK2}]\text{EpoRActJAK2}}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} \\
&\quad - \frac{3\cdot[\text{EpoRpJAK2}]\cdot\text{EpoRActJAK2}}{(\text{EpoRCISInh}\cdot[\text{EpoRJAK2}_{\text{CIS}}]+1)\cdot([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1)} \\
&\quad - [\text{EpoRpJAK2}]\,\frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}\,[\text{SHP1Act}] \\
\frac{\partial[\text{p1EpoRpJAK2}]}{\partial t} &= \frac{[\text{EpoRpJAK2}]\cdot\text{EpoRActJAK2}}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} - \frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}[\text{SHP1Act}][\text{p1EpoRpJAK2}] \\
&\quad - \frac{3\cdot\text{EpoRActJAK2}\cdot[\text{p1EpoRpJAK2}]}{(\text{EpoRCISInh}\cdot[\text{EpoRJAK2}_{\text{CIS}}]+1)\cdot([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1)} \\
\frac{\partial[\text{p2EpoRpJAK2}]}{\partial t} &= \frac{3\cdot[\text{EpoRpJAK2}]\cdot\text{EpoRActJAK2}}{(\text{EpoRCISInh}\cdot[\text{EpoRJAK2}_{\text{CIS}}]+1)\cdot([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1)} \\
&\quad - \frac{\text{EpoRActJAK2}\cdot[\text{p2EpoRpJAK2}]}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} - \frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}[\text{SHP1Act}][\text{p2EpoRpJAK2}] \\
\frac{\partial[\text{p12EpoRpJAK2}]}{\partial t} &= \frac{\text{EpoRActJAK2}\cdot[\text{p2EpoRpJAK2}]}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} - \frac{\text{JAK2EpoRDeaSHP1}}{\text{init}_{\text{SHP1}}}\cdot[\text{SHP1Act}]\cdot[\text{p12EpoRpJAK2}] \\
&\quad + \frac{3\cdot\text{EpoRActJAK2}\cdot[\text{p1EpoRpJAK2}]}{(\text{EpoRCISInh}\cdot[\text{EpoRJAK2}_{\text{CIS}}]+1)([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1)} \\
\frac{\partial[\text{EpoRJAK2}_{\text{CIS}}]}{\partial t} &= -[\text{EpoRJAK2}_{\text{CIS}}]\cdot\frac{\text{EpoRCISRemove}}{\text{init}_{\text{EpoRJAK2}}}\,([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}]) \\
\frac{\partial[\text{SHP1}]}{\partial t} &= \text{SHP1Dea}[\text{SHP1Act}] - [\text{SHP1}]\cdot\frac{\text{SHP1ActEpoR}}{\text{init}_{\text{EpoRJAK2}}} \\
&\quad \cdot\big([\text{EpoRpJAK2}] + [\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]\big) \\
\frac{\partial[\text{SHP1Act}]}{\partial t} &= [\text{SHP1}]\cdot\frac{\text{SHP1ActEpoR}}{\text{init}_{\text{EpoRJAK2}}}\cdot\big([\text{EpoRpJAK2}] \\
&\quad + [\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]\big) - \text{SHP1Dea}\cdot[\text{SHP1Act}] \\
\frac{\partial[\text{STAT5}]}{\partial t} &= \frac{\text{STAT5Exp}\cdot[\text{npSTAT5}]\cdot 0.275}{0.4} \\
&\quad - \frac{[\text{STAT5}]\cdot\frac{\text{STAT5ActJAK2}}{\text{init}_{\text{EpoRJAK2}}}}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1} \\
&\quad \cdot([\text{EpoRpJAK2} + [\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]) \\
&\quad - [\text{STAT5}]\frac{\text{STAT5ActEpoR}}{\text{init}^2_{\text{EpoRJAK2}}}\,\frac{([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}])^2}{([\text{CIS}]\frac{\text{CISInh}}{\text{CISEqc}}+1)([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}}+1)}
\end{aligned}
$$

$$\frac{\partial[\text{pSTAT5}]}{\partial t} = \frac{[\text{STAT5}]\frac{\text{STAT5ActJAK2}}{\text{init}_{\text{EpoRJAK2}}} \cdot ([\text{EpoRpJAK2}] + [\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}])}{[\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}} + 1}$$
$$- \text{STAT5Imp} \cdot [\text{pSTAT5}] + [\text{STAT5}]\frac{\text{STAT5ActEpoR}}{\text{init}^2_{\text{iEpoRJAK2}}} \cdot \frac{([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}])^2}{([\text{CIS}]\frac{\text{CISInh}}{\text{CISEqc}} + 1) \cdot ([\text{SOCS3}]\frac{\text{SOCS3Inh}}{\text{SOCS3Eqc}} + 1)}$$

$$\frac{\partial[\text{npSTAT5}]}{\partial t} = \frac{\text{STAT5Imp} \cdot 0.4 \cdot [\text{pSTAT5}]}{0.275} - \text{STAT5Exp} \cdot [\text{npSTAT5}]$$

$$\frac{\partial[\text{CISnRNA1}]}{\partial t} = -[\text{CISnRNA1}] \cdot \text{CISRNADelay} - \frac{1}{\text{init}_{\text{STAT5}}} \cdot \text{CISRNATurn} \cdot [\text{npSTAT5}] \cdot (\text{ActD} - 1)$$

$$\frac{\partial[\text{CISnRNA2}]}{\partial t} = [\text{CISnRNA1}] \cdot \text{CISRNADelay} - [\text{CISnRNA2}] \cdot \text{CISRNADelay}$$

$$\frac{\partial[\text{CISnRNA3}]}{\partial t} = [\text{CISnRNA2}] \cdot \text{CISRNADelay} - [\text{CISnRNA3}] \cdot \text{CISRNADelay}$$

$$\frac{\partial[\text{CISnRNA4}]}{\partial t} = [\text{CISnRNA3}] \cdot \text{CISRNADelay} - [\text{CISnRNA4}] \cdot \text{CISRNADelay}$$

$$\frac{\partial[\text{CISnRNA5}]}{\partial t} = [\text{CISnRNA4}] \cdot \text{CISRNADelay} - [\text{CISnRNA5}] \cdot \text{CISRNADelay}$$

$$\frac{\partial[\text{CISRNA}]}{\partial t} = \frac{[\text{CISnRNA5}] \cdot \text{CISRNADelay} \cdot 0.275}{0.4} - [\text{CISRNA}] \cdot \text{CISRNATurn}$$

$$\frac{\partial[\text{CIS}]}{\partial t} = [\text{CISRNA}] \cdot \text{CISEqc} \cdot \text{CISTurn} - [\text{CIS}] \cdot \text{CISTurn} + \text{CISoe} \cdot \text{CISTurn} \cdot \text{CISEqcOE} \cdot \text{CISEqc}$$

$$\frac{\partial[\text{SOCS3nRNA1}]}{\partial t} = -[\text{SOCS3nRNA1}] \cdot \text{SOCS3RNADelay} - \frac{1}{\text{init}_{\text{STAT5}}} \cdot \text{SOCS3RNATurn} \cdot [\text{npSTAT5}] \cdot (\text{ActD} - 1)$$

$$\frac{\partial[\text{SOCS3nRNA2}]}{\partial t} = [\text{SOCS3nRNA1}] \cdot \text{SOCS3RNADelay} - [\text{SOCS3nRNA2}] \cdot \text{SOCS3RNADelay}$$

$$\frac{\partial[\text{SOCS3nRNA3}]}{\partial t} = [\text{SOCS3nRNA2}] \cdot \text{SOCS3RNADelay} - [\text{SOCS3nRNA3}] \cdot \text{SOCS3RNADelay}$$

$$\frac{\partial[\text{SOCS3nRNA4}]}{\partial t} = [\text{SOCS3nRNA3}] \cdot \text{SOCS3RNADelay} - [\text{SOCS3nRNA4}] \cdot \text{SOCS3RNADelay}$$

$$\frac{\partial[\text{SOCS3nRNA5}]}{\partial t} = [\text{SOCS3nRNA4}] \cdot \text{SOCS3RNADelay} - [\text{SOCS3nRNA5}] \cdot \text{SOCS3RNADelay}$$

$$\frac{\partial[\text{SOCS3RNA}]}{\partial t} = \frac{[\text{SOCSnRNA5}] \cdot \text{SOCSRNADelay} \cdot 0.275}{0.4} - [\text{SOCSRNA}] \cdot \text{SOCSRNATurn}$$

$$\frac{\partial[\text{SOCS3}]}{\partial t} = [\text{SOCS3RNA}] \cdot \text{SOCS3Eqc} \cdot \text{SOCS3Turn} - [\text{SOCS3}] \cdot \text{SOCS3Turn}$$
$$+ \text{SOCS3oe} \cdot \text{SOCS3Turn} \cdot \text{SOCS3EqcOE} \cdot \text{SOCS3Eqc}\,,$$

with condition-specific initial conditions (see Table S1) denoted by $x_{i,c_e}(0)$ for observable index $i$ under condition indexed by $c_e$:

$$x_{1,c_e}(0) = \text{init}_{\text{EpoRJAK2}}, x_{9,c_e}(0) = \text{init}_{\text{STAT5}}, x_{i,c_e}(0) = 0, i = \{2,3,4,5,8,10,11,12,13,14,15,16,17,19,20,21,22,23,24\}, \forall c_e$$
$$x_{i,c_e}(0) = 0, i = \{6,18,25\}, c_e = \{1,2,3,4,5,6,15,\ldots,36\}$$
$$x_{6,c_e}(0) = u_{c_e,2}, x_{18,c_e}(0) = u_{c_e,2} \cdot (\text{CISEqc} \cdot \text{CISEqcOE}), c_e = \{7,8,9,10\}$$
$$x_{7,c_e}(0) = \text{init}_{\text{SHP1}}, c_e = \{1,2,3,4,5,6,7,8,9,10,11,12,15,\ldots,36\}$$
$$x_{7,c_e}(0) = (1 + u_{c_e,4} \cdot \text{SHP1ProOE}) \cdot \text{init}_{\text{SHP1}}, c_e = \{13,14\}$$
$$x_{6,c_e}(0) = 0, x_{18,c_e}(0) = 0, x_{25,c_e}(0) = u_{c_e,3} \cdot (\text{SOCS3Eqc} \cdot \text{SOCS3EqcOE}), c_e = \{11,12\}$$

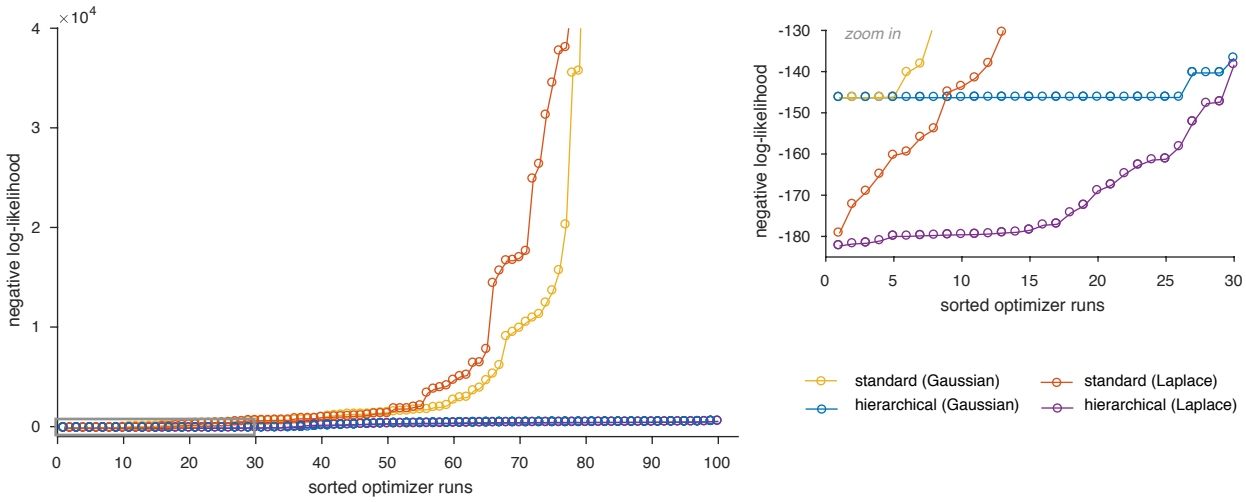
The observables are given by

$$
\begin{aligned}
y_1 = \mathrm{pJAK2_{au}} &= s_{1,c_e} \cdot \Big(o_{1,c_e} + \frac{2}{\mathrm{init_{EpoRJAK2}}} \cdot ([\mathrm{EpoRpJAK2}] + [\mathrm{p1EpoRpJAK2}] + \\
&\qquad [\mathrm{p2EpoRpJAK2}] + [\mathrm{p12EpoRpJAK2}])\Big) \\
y_2 = \mathrm{pEpoR_{au}} &= s_{2,c_e} \cdot \left(o_{2,c_e} + \frac{16}{\mathrm{init_{EpoRJAK2}}} \cdot ([\mathrm{p1EpoRpJAK2}] + [\mathrm{p2EpoRpJAK2}] + [\mathrm{p12EpoRpJAK2}])\right) \\
y_3 = \mathrm{CIS_{au}} &= s_{3,c_e} \cdot \left(o_{3,c_e} + \frac{[\mathrm{CIS}]}{\mathrm{CISEqc}}\right) \\
y_4 = \mathrm{SOCS3_{au}} &= s_{4,c_e} \cdot \left(o_{4,c_e} + \frac{[\mathrm{SOCS3}]}{\mathrm{SOCS3Eqc}}\right) \\
y_5 = \mathrm{tSTAT5_{au}} &= s_{5,c_e} \cdot \left(\frac{1}{\mathrm{init_{STAT5}}}([\mathrm{STAT5}] + [\mathrm{pSTAT5}])\right) \\
y_6 = \mathrm{pSTAT5_{au}} &= s_{6,c_e} \cdot \left(o_{6,c_e} + \frac{1}{\mathrm{init_{STAT5}}}[\mathrm{pSTAT5}]\right) \\
y_7 = \mathrm{STAT5_{abs}} &= [\mathrm{STAT5}] \\
y_8 = \mathrm{SHP1_{abs}} &= [\mathrm{SHP1}] + [\mathrm{SHP1Act}] \\
y_9 = \mathrm{CIS_{abs}} &= [\mathrm{CIS}] \\
y_{10} = \mathrm{SOCS3_{abs}} &= [\mathrm{SOCS3}] \\
y_{11} = \mathrm{pSTAT5B_{rel}} &= o_{11} + 100\frac{[\mathrm{pSTAT5}]}{[\mathrm{pSTAT5}] + [\mathrm{STAT5}]} \\
y_{12} = \mathrm{SOCS3RNA_{foldA}} &= 1 + s_{12} \cdot [\mathrm{SOCS3RNA}] \\
y_{13} = \mathrm{SOCS3RNA_{foldB}} &= 1 + s_{13} \cdot [\mathrm{SOCS3RNA}] \\
y_{14} = \mathrm{SOCS3RNA_{foldC}} &= 1 + s_{14} \cdot [\mathrm{SOCS3RNA}] \\
y_{15} = \mathrm{CISRNA_{foldA}} &= 1 + s_{15} \cdot [\mathrm{CISRNA}] \\
y_{16} = \mathrm{CISRNA_{foldB}} &= 1 + s_{16} \cdot [\mathrm{CISRNA}] \\
y_{17} = \mathrm{CISRNA_{foldC}} &= 1 + s_{17} \cdot [\mathrm{CISRNA}] \\
y_{18} = \mathrm{tSHP1_{au}} &= s_{18} \cdot \left(\frac{1}{\mathrm{init_{SHP1}}}([\mathrm{SHP1}] + [\mathrm{SHP1Act}])(1 + (\mathrm{SHP1oe} \cdot \mathrm{SHP1ProOE}))\right) \\
y_{19} = \mathrm{CIS_{au1}} &= s_{19} \cdot \frac{[\mathrm{CIS}]}{\mathrm{CISEqc}} \\
y_{20} = \mathrm{CIS_{au2}} &= s_{20} \cdot \frac{[\mathrm{CIS}]}{\mathrm{CISEqc}}\,.
\end{aligned}
$$

The parameters $\boldsymbol{\theta}$ are

$$
\begin{aligned}
\boldsymbol{\theta} = (&\mathrm{CISEqc}, \mathrm{CISEqcOE}, \mathrm{CISInh}, \mathrm{CISRNADelay}, \mathrm{CISRNATurn}, \mathrm{CISTurn}, \mathrm{EpoRActJAK2}, \mathrm{EpoRCISInh}, \\
&\mathrm{EpoRCISRemove}, \mathrm{JAK2ActEpo}, \mathrm{JAK2EpoRDeaSHP1}, \mathrm{SHP1ActEpoR}, \mathrm{SHP1Dea}, \mathrm{SHP1ProOE}, \\
&\mathrm{SOCS3Eqc}, \mathrm{SOCS3EqcOE}, \mathrm{SOCS3Inh}, \mathrm{SOCS3RNADelay}, \mathrm{SOCS3RNATurn}, \mathrm{SOCS3Turn}, \\
&\mathrm{STAT5ActEpoR}, \mathrm{STAT5ActJAK2}, \mathrm{STAT5Exp}, \mathrm{STAT5Imp}, \mathrm{init_{EpoRJAK2}}, \mathrm{init_{SHP1}}, \mathrm{init_{STAT5}} \\
&o_{1,1}, o_{1,4}, o_{1,6}, o_{1,7}, o_{1,11}, o_{1,13}, o_{1,15}, o_{1,20}, o_{2,1}, o_{2,4}, o_{2,6}, o_{2,7}, o_{2,9}, o_{2,11}, o_{2,13}, o_{2,15}, o_{2,20}, o_{3,1}, o_{3,4}, o_{3,7}, o_{3,11}, o_{3,13}, \\
&o_{4,1}, o_{4,7}, o_{4,11}, o_{6,1}, o_{6,2}, o_{6,4}, o_{6,7}, o_{6,11}, o_{6,13})^T
\end{aligned}
$$

with $n_\theta = 58$. For experiment SHP1oe ($e = 9$), the parameter $\mathrm{init_{SHP1}}$ was replaced by $\mathrm{init_{SHP1}} \cdot (1 + (\mathrm{SHP1oe} \cdot \mathrm{SHP1ProOE}))$ in the model equations. For the notation of the offset, scaling, and noise parameters, we neglected the index $r$, since these parameters are shared for the replicates. The first subscript indicates the observable, and the second




Supplementary Figure S3: Likelihood waterfall for the JAK-STAT signaling model II.

the condition. However, all conditions belonging to the same experiment share the scaling and offset parameters and thus the parameters are only listed for the first condition of each experiment. The experiments and corresponding condition indices are summarizes in Table S1. For simplicity, we note the scaling parameters as vector $\mathbf{s}$ which contains only the unique parameters $s_{i,c_e}$ which need to be estimated from the data. Thus, it is

$$\mathbf{s} = (s_{1,1}, s_{1,4}, s_{1,6}, s_{1,7}, s_{1,11}, s_{1,15}, s_{1,20}, s_{2,1}, s_{2,4}, s_{2,5}, s_{2,7}, s_{2,9}, s_{2,11}, s_{2,13}, s_{2,15}, s_{2,20}, s_{3,1}, s_{3,4}, s_{3,7}, s_{3,11}, s_{3,13}, s_{4,1}, s_{4,7}, s_{4,11}, s_{5,1}, s_{5,4}, s_{5,13}, s_{6,1}, s_{6,4}, s_{6,7}, s_{6,11}, s_{6,13}, s_{6,26}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20})^T$$

with $n_s = 42$. The noise parameters do not differ between experiments or replicates, thus, neglecting the subscripts for the experiment-specific condition index $c_e$ and for the replicate index $r$, the noise parameters, which need to be estimated from the data are given by

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_3, \sigma_4, \sigma_5, \sigma_7, \sigma_8, \sigma_9, \sigma_{10}, \sigma_{11}, \sigma_{12}, \sigma_{18})^T$$

with $n_\sigma = 11$. Some observables have the same noise parameters:

$$\begin{aligned}
\sigma_1 &= \sigma_2 \\
\sigma_3 &= \sigma_{19} = \sigma_{20}\,, \\
\sigma_5 &= \sigma_6\,, \\
\sigma_{12} &= \sigma_{13} = \sigma_{14} = \sigma_{15} = \sigma_{16} = \sigma_{17}\,.
\end{aligned}$$

A minor modification from the model proposed by Bachmann et al. (2011) is that the parameterization for the noise of $\text{pSTAT5B}_{\text{au}}$ did not include an additional parameter for the SOCS3oe experiment, and that the observables for RNA were fitted in linear space. The observable $\text{pSTAT5B}_{\text{rel}}$ was also fitted on a linear scale, while the other observables were compared at a $\log_{10}$ scale (as done by Bachmann et al. (2011)). In our setting, the offset parameters were also multiplied with the scaling parameters, which yielded different optimal values for the offset parameters compared to those found by Bachmann et al. (2011). We performed 100 multi-starts for each optimization approach and noise assumption. The

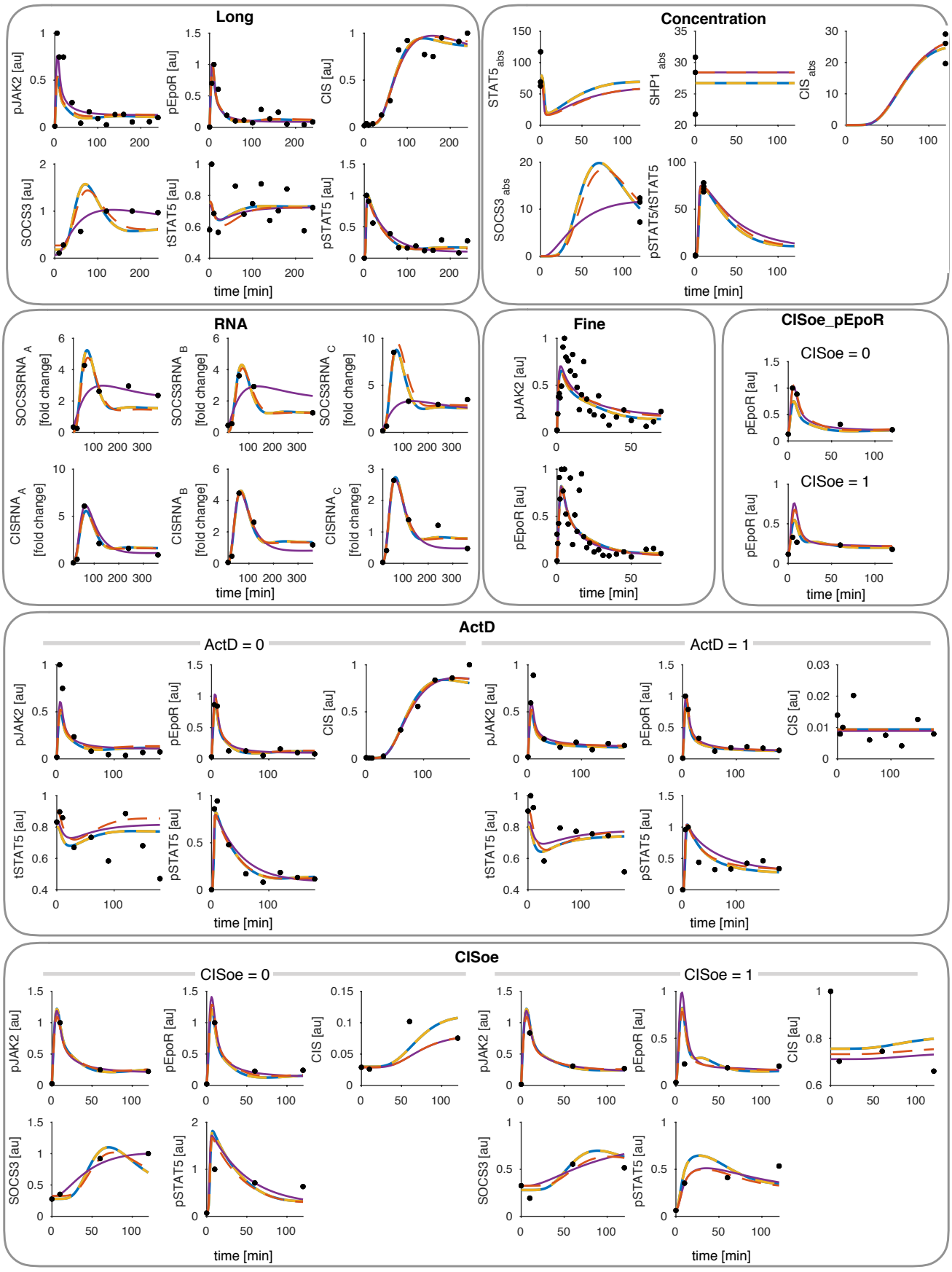
Supplementary Table S1: Overview for the experimental data of JAK-STAT signaling model II.

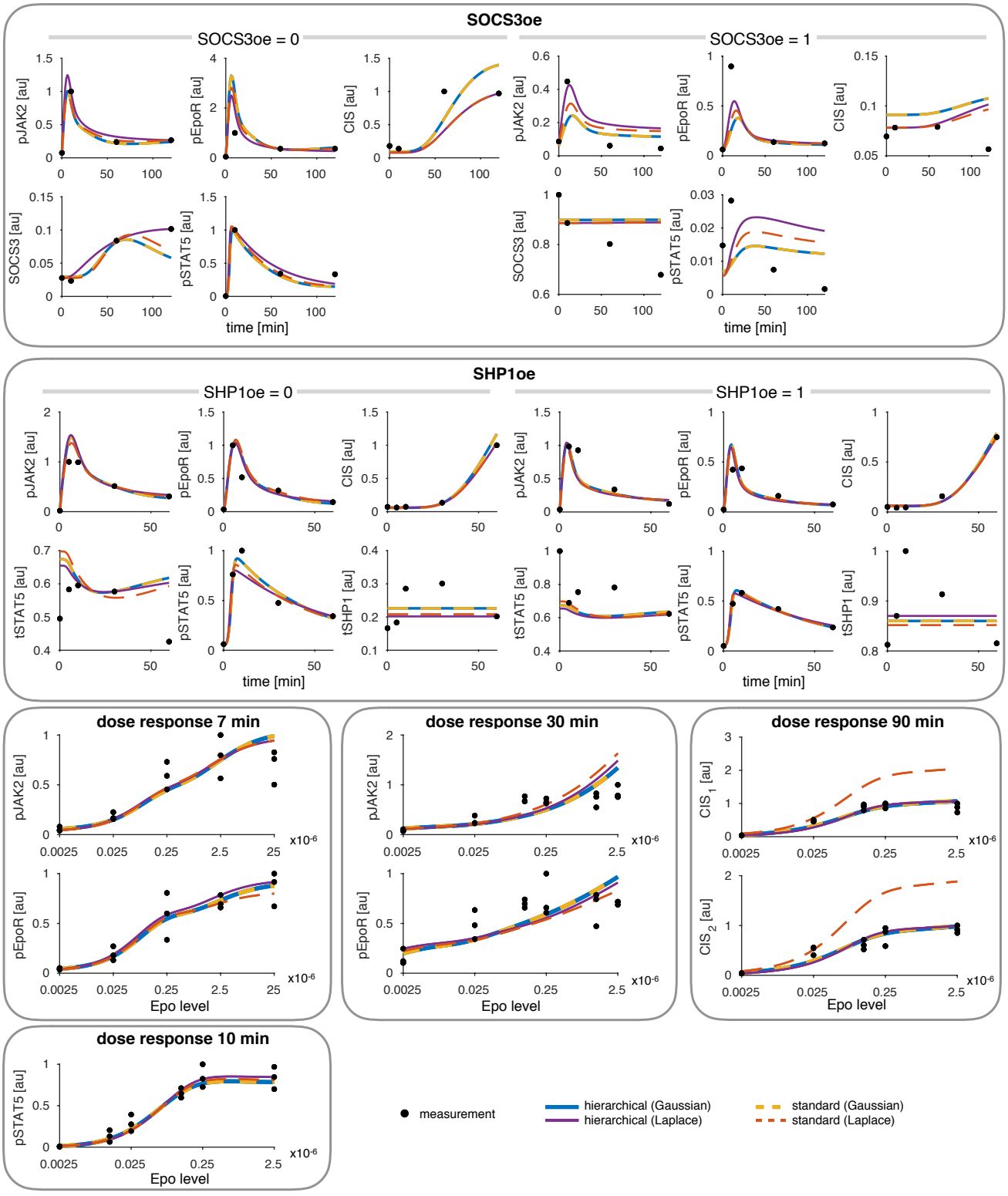
| name | experiment index $e$ | condition index | condition $\mathbf{u}$ | | | | |
|---|---|---|---|---|---|---|---|
| | | | ActD | CISoe | SOCS3oe | SHP1oe | [Epo]/$10^{-6}$ |
| Long | 1 | 1 | 0 | 0 | 0 | 0 | 0.125 |
| Concentration | 2 | 2 | 0 | 0 | 0 | 0 | 0.125 |
| RNA | 3 | 3 | 0 | 0 | 0 | 0 | 0.125 |
| ActD | 4 | 4 | 0 | 0 | 0 | 0 | 0.125 |
| | | 5 | 1 | 0 | 0 | 0 | 0.125 |
| Fine | 5 | 6 | 0 | 0 | 0 | 0 | 1.25 |
| CISoe | 6 | 7 | 0 | 0 | 0 | 0 | 0.125 |
| | | 8 | 0 | 1 | 0 | 0 | 0.125 |
| CISoe_pEpoR | 7 | 9 | 0 | 0 | 0 | 0 | 0.125 |
| | | 10 | 0 | 1 | 0 | 0 | 0.125 |
| SOCS3oe | 8 | 11 | 0 | 0 | 0 | 0 | 0.125 |
| | | 12 | 0 | 0 | 1 | 0 | 0.125 |
| SHP1oe | 9 | 13 | 0 | 0 | 0 | 0 | 0.125 |
| | | 14 | 0 | 0 | 0 | 1 | 0.125 |
| dose response 7 min | 10 | 15 | 0 | 0 | 0 | 0 | 0.0025 |
| | | 16 | 0 | 0 | 0 | 0 | 0.025 |
| | | 17 | 0 | 0 | 0 | 0 | 0.25 |
| | | 18 | 0 | 0 | 0 | 0 | 2.5 |
| | | 19 | 0 | 0 | 0 | 0 | 25 |
| dose response 30 min | 11 | 20 | 0 | 0 | 0 | 0 | 0.0025 |
| | | 21 | 0 | 0 | 0 | 0 | 0.025 |
| | | 22 | 0 | 0 | 0 | 0 | 0.125 |
| | | 23 | 0 | 0 | 0 | 0 | 0.25 |
| | | 24 | 0 | 0 | 0 | 0 | 1.25 |
| | | 25 | 0 | 0 | 0 | 0 | 2.5 |
| dose response 10 min | 12 | 26 | 0 | 0 | 0 | 0 | 0.0025 |
| | | 27 | 0 | 0 | 0 | 0 | 0.0125 |
| | | 28 | 0 | 0 | 0 | 0 | 0.025 |
| | | 29 | 0 | 0 | 0 | 0 | 0.125 |
| | | 30 | 0 | 0 | 0 | 0 | 0.25 |
| | | 31 | 0 | 0 | 0 | 0 | 2.5 |
| dose response 90 min | 13 | 32 | 0 | 0 | 0 | 0 | 0.0025 |
| | | 33 | 0 | 0 | 0 | 0 | 0.1025 |
| | | 34 | 0 | 0 | 0 | 0 | 0.125 |
| | | 35 | 0 | 0 | 0 | 0 | 0.25 |
| | | 36 | 0 | 0 | 0 | 0 | 2.5 |

parameter boundaries are $\log_{10}(\boldsymbol{\theta})_{\text{lb}} = -3$ and $\log_{10}(\boldsymbol{\theta})_{\text{ub}} = 3$, except for

$$\log_{10}(\text{CISEqc}, \text{CISInh}, \text{EpoRActJAK2}, \text{EpoRCISInh}, \text{JAK2ActEpo}, \text{JAK2EpoRDeaSHP1}, \text{SOCS3Turn})_{\text{ub}} = (4, 12, 5, 6, 9, 4, 4)^T$$
$$\log_{10}(o_{i,c_e})_{\text{lb}} = -5 \quad \forall i, c_e$$
$$\log_{10}(o_{i,c_e})_{\text{ub}} = 3 \quad \forall i, c_e$$
$$\log_{10}(\mathbf{s})_{\text{lb}} = (-3, \ldots, -3)^T$$
$$\log_{10}(\mathbf{s})_{\text{ub}} = (3, \ldots, 3)^T$$
$$\log_{10}(\boldsymbol{\sigma})_{\text{lb}} = (-3, \ldots, -3)^T$$
$$\log_{10}(\boldsymbol{\sigma})_{\text{ub}} = (3, \ldots, 3)^T .$$

The fitted experimental data for the whole data set are shown in Figure S4. The comparison of Gaussian and Laplace noise showed that Laplace noise yielded a substantially improved fit of the data (Figure S1C).






Supplementary Figure S4: Experimental data for JAK-STAT signaling model II. Boxes indicate different experiments. The lines highlight the different models (Gaussian and Laplace noise) and optimization approaches (standard and hierarchical).

### 4.3 RAF/MEK/ERK signaling

The ODE system for the RAF/MEK/ERK signaling model is given by

$$\begin{aligned}\frac{dx_1}{\partial t} &= k_{1,\max}(t)\frac{K_1}{K_1+[\text{pERK}]}(1-x_1)-k_2x_1\\ \frac{dx_2}{\partial t} &= \frac{k_3[\text{Raf}]_0K_2x_1}{K_2+[\text{sora}]}(1-x_2)-k_4x_2\\ \frac{dx_3}{\partial t} &= \frac{k_5[\text{MEK}]_0K_3x_2}{K_3+[\text{UO126}]}(1-x_3)-k_6x_3\end{aligned}$$

with states $x_1 = [\text{pRaf}]/[\text{Raf}]_0$, $x_2 = [\text{pMEK}]/[\text{MEK}]_0$, and $x_3 = [\text{pERK}]/[\text{ERK}]_0$, and

$$k_{1,\max}(t) = k_{1,0} + k_{1,1}\left(1-\exp\left(-\frac{t}{\tau_1}\right)\right)\exp\left(-\frac{t}{\tau_2}\right)$$

(see (Fiedler et al., 2016) for more details). The initial conditions were assumed to be the steady states reached without stimulation and for $k_{1,\max} = k_{1,0}$. Defining $\tilde{K}_1 = K_1/[\text{ERK}]_0$, $\tilde{k}_3 = k_3[\text{Raf}]_0$ and $\tilde{k}_5 = k_5[\text{MEK}]_0$, we obtain

$$\begin{aligned}x_1(0) = \Bigg(&\tilde{K}_1k_{1,0} + \Bigg(\tilde{K}_1^2k_{1,0}^2 + \frac{2\tilde{K}_1^2k_6k_{1,0}^2}{\tilde{k}_5} + \frac{\tilde{K}_1^2k_6^2k_{1,0}^2}{\tilde{k}_5^2} + \frac{\tilde{K}_1^2k_4^2k_6^2\left(k_{1,0}+k_2\right)^2}{(\tilde{k}_3\tilde{k}_5)^2} + \\ &\frac{2\tilde{K}_1^2k_4k_6^2k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5^2} + \frac{2\tilde{K}_1^2k_4k_6k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5} + \frac{4\tilde{K}_1k_2k_4k_6k_{1,0}}{\tilde{k}_3\tilde{k}_5}\Bigg)^{\frac{1}{2}} + \frac{\tilde{K}_1k_6k_{1,0}}{\tilde{k}_5} - \\ &\frac{\tilde{K}_1k_4k_6\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5}\Bigg) \Big/ \left(2\left(k_2+\tilde{K}_1k_{1,0}+\tilde{K}_1k_2+\frac{\tilde{K}_1k_2k_6}{\tilde{k}_5}+\frac{\tilde{K}_1k_6k_{1,0}}{\tilde{k}_5}\right)\right)\end{aligned}$$

$$\begin{aligned}x_2(0) = \Bigg(&\Bigg(\tilde{K}_1^2k_{1,0}^2 + \frac{2\tilde{K}_1^2k_6k_{1,0}^2}{\tilde{k}_5} + \frac{\tilde{K}_1^2k_6^2k_{1,0}^2}{\tilde{k}_5^2} + \frac{\tilde{K}_1^2k_4^2k_6^2\left(k_{1,0}+k_2\right)^2}{(\tilde{k}_3\tilde{k}_5)^2} + \\ &\frac{2\tilde{K}_1^2k_4k_6^2k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5^2} + \frac{2\tilde{K}_1^2k_4k_6k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5} + \frac{4\tilde{K}_1k_2k_4k_6k_{1,0}}{\tilde{k}_3\tilde{k}_5}\Bigg)^{\frac{1}{2}} + \\ &\tilde{K}_1k_{1,0} + \frac{\tilde{K}_1k_6k_{1,0}}{\tilde{k}_5} - \frac{\tilde{K}_1k_2k_4k_6}{\tilde{k}_3\tilde{k}_5} - \frac{\tilde{K}_1k_4k_6k_{1,0}}{\tilde{k}_3\tilde{k}_5}\Bigg) \Big/ \\ &\Bigg(\Bigg(\tilde{K}_1^2k_{1,0}^2 + \frac{2\tilde{K}_1^2k_6k_{1,0}^2}{\tilde{k}_5} + \frac{\tilde{K}_1^2k_6^2k_{1,0}^2}{\tilde{k}_5^2} + \frac{\tilde{K}_1^2k_4^2k_6^2\left(k_{1,0}+k_2\right)^2}{(\tilde{k}_3\tilde{k}_5)^2} + \frac{2\tilde{K}_1^2k_4k_6^2k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5^2} + \\ &\frac{2\tilde{K}_1^2k_4k_6k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5} + \frac{4\tilde{K}_1k_2k_4k_6k_{1,0}}{\tilde{k}_3\tilde{k}_5}\Bigg)^{\frac{1}{2}} + \\ &\tilde{K}_1k_{1,0} + \frac{\tilde{K}_1k_6k_{1,0}}{\tilde{k}_5} + \frac{k_2k_4}{1\tilde{k}_3}\left(2\tilde{K}_1 + \frac{\tilde{K}_1k_6}{\tilde{k}_5}+2\right) + \frac{\tilde{K}_1k_4k_{1,0}}{\tilde{k}_3}\left(\frac{k_6}{\tilde{k}_5}+2\right)\Bigg)\end{aligned}$$

$$
\begin{aligned}
x_3(0) = \Bigg( \Bigg( & \tilde{K}_1^2 \left(k_{1,0}\right)^2 + \frac{2\tilde{K}_1^2 k_6 k_{1,0}^2}{\tilde{k}_5} + \frac{\tilde{K}_1^2 k_6^2 k_{1,0}^2}{\tilde{k}_5^2} + \frac{\tilde{K}_1^2 k_4^2 k_6^2 \left(k_{1,0}+k_2\right)^2}{(\tilde{k}_3\tilde{k}_5)^2} + \\
& \frac{2\tilde{K}_1^2 k_4 k_6^2 k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5^2} + \frac{2\tilde{K}_1^2 k_4 k_6 k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5} \\
& + \frac{4\tilde{K}_1 k_2 k_4 k_6 k_{1,0}}{\tilde{k}_3\tilde{k}_5} \Bigg)^{\frac{1}{2}} + \tilde{K}_1 k_{1,0} + \frac{\tilde{K}_1 k_6 k_{1,0}}{\tilde{k}_5} - \frac{\tilde{K}_1 k_2 k_4 k_6}{1\tilde{k}_3\tilde{k}_5} \\
& - \frac{\tilde{K}_1 k_4 k_6 k_{1,0}}{\tilde{k}_3\tilde{k}_5} \Bigg) \Big/ \Bigg( \left(\frac{k_6}{\tilde{k}_5}+1\right) \Bigg( \tilde{K}_1^2 k_{1,0}^2 + \frac{2\tilde{K}_1^2 k_6 k_{1,0}^2}{\tilde{k}_5} + \frac{\tilde{K}_1^2 k_6^2 k_{1,0}^2}{\tilde{k}_5^2} + \\
& \frac{\tilde{K}_1^2 k_4^2 k_6^2 \left(k_{1,0}+k_2\right)^2}{(\tilde{k}_3\tilde{k}_5)^2} + \frac{2\tilde{K}_1^2 k_4 k_6^2 k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5^2} + \\
& \frac{2\tilde{K}_1^2 k_4 k_6 k_{1,0}\left(k_{1,0}+k_2\right)}{\tilde{k}_3\tilde{k}_5} + \frac{4\tilde{K}_1 k_2 k_4 k_6 k_{1,0}}{\tilde{k}_3\tilde{k}_5} \Bigg)^{\frac{1}{2}} + \tilde{K}_1 k_{1,0}\left(\frac{k_6}{\tilde{k}_5}+1\right)^2 \\
& + \frac{k_2 k_4 k_6}{\tilde{k}_3\tilde{k}_5}\left(\tilde{K}_1 + \frac{\tilde{K}_1 k_6}{\tilde{k}_5} + 2\right) + \frac{\tilde{K}_1 k_6 k_6 k_{1,0}}{\tilde{k}_3\tilde{k}_5}\left(\frac{k_6}{\tilde{k}_5}+1\right)\Bigg).
\end{aligned}
$$

The observables are given by

$$
\begin{aligned}
y_{1,r} &= s_{1,r}[\text{pMEK}] \\
y_{2,r} &= s_{2,r}[\text{pERK}]\,,
\end{aligned}
$$

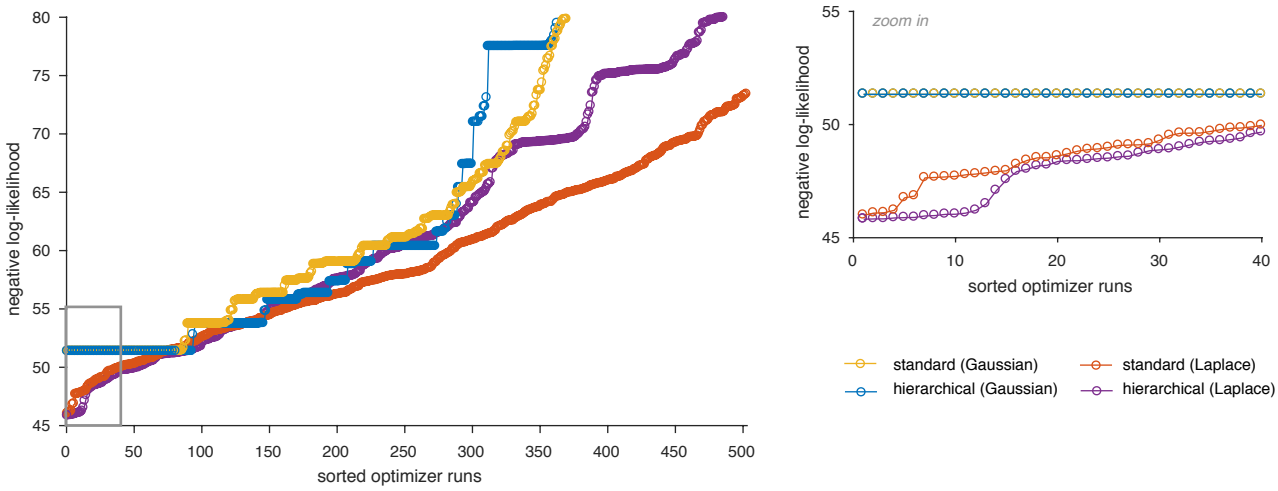
for replicates $r = 1, \ldots, 4$. The indices for conditions and experiments are neglected, since the scaling and noise parameters do not differ for these. The input $\mathbf{u}$ describes the concentrations [sora] and [UO126] and the three different conditions are $u_1 = (0,0)^T$, $u_2 = (0,30)^T$, and $u_3 = (5,0)^T$. The parameters, which are estimated from the data, are

$$
\begin{aligned}
\mathbf{q} = \Bigg( & \frac{k_{1,0}}{k_{1,1}}, k_{1,1}, \tau_1, \tau_2, \frac{K_1}{[\text{ERK}]_0}, k_2, K_2, k_3[\text{Raf}]_0, K_3, k_4, k_5[\text{MEK}]_0, k_6, \\
& s_{1,2}, s_{1,3}, s_{1,4}, s_{2,1}, s_{2,2}, s_{2,3}, s_{2,4}, \sigma_{1,2}, \sigma_{1,3}, \sigma_{1,4}, \sigma_{2,1}, \sigma_{2,2}, \sigma_{2,3}, \sigma_{2,4} \Bigg)^T.
\end{aligned}
$$

with specific scaling and noise parameters for replicates and observables. The parameters boundaries for the optimization are

$$
\begin{aligned}
\log_{10}(\mathbf{q})_{\text{lb}} &= (-7, \ldots, -7)^T \\
\log_{10}(\mathbf{q})_{\text{ub}} &= (5, \ldots, 5)^T\,.
\end{aligned}
$$

We performed 500 multi-starts to obtain the optimal parameters for both distributions. The comparison between the two noise assumptions for the measurement noise showed that the Laplace noise yielded a substantially better fit (Figure S1C).




Supplementary Figure S5: Likelihood waterfall plot for RAF/MEK/ERK signaling.

# References

Bachmann, J., Raue, A., Schilling, M., Böhm, M. E., Kreutz, C., Kaschek, D., Busch, H., Gretz, N., Lehmann, W. D., Timmer, J., and Klingmüller, U. (2011). Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, 7(1):516.

Fiedler, A., Raeth, S., Theis, F. J., Hausser, A., and Hasenauer, J. (2016). Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints. *BMC Syst. Biol.*, 10(80).

Fröhlich, F., Kaltenbacher, B., Theis, F. J., and Hasenauer, J. (2017). Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, 13(1):e1005331.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, 90(430):773–795.

Maier, C., Loos, C., and Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725.

Raftery, A. E. (1999). Bayes factors and BIC. *Socio. Meth. Res.*, 27(3):411–417.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(25):1923–1929.

Schelker, M., Raue, A., Timmer, J., and Kreutz, C. (2012). Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinformatics*, 28(18):i529–i534.

Stapor, P., Weindl, D., Ballnus, B., Hug, S., Loos, C., Fiedler, A., Krause, S., Hross, S., Fröhlich, F., and Hasenauer, J. (2017). PESTO: Parameter EStimation TOolbox. *Bioinformatics*, btx676.

Vaz, A. I. F. and Vicente, L. N. (2009). PSwarm: A hybrid solver for linearly constrained global derivative-free optimization. *Optim. Method. Softw.*, 24(4-5):669–685.