Bioinformatics, 37(23), 2021, 4493–4500
doi: 10.1093/bioinformatics/btab512
Advance Access Publication Date: 14 July 2021
Original Paper

Systems biology

Efficient gradient-based parameter estimation for dynamic models using qualitative data

Leonard Schmiester [1,2], Daniel Weindl [1] and Jan Hasenauer [1,2,3,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg 85764, Germany, [2]Center for Mathematics, Technische Universität München, Garching 85748, Germany and [3]Faculty of Mathematics and Natural Sciences, University of Bonn, Bonn 53113, Germany

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier
Received on February 6, 2021; revised on July 2, 2021; editorial decision on July 5, 2021; accepted on July 8, 2021

Abstract

Motivation: Unknown parameters of dynamical models are commonly estimated from experimental data. However, while various efficient optimization and uncertainty analysis methods have been proposed for quantitative data, methods for qualitative data are rare and suffer from bad scaling and convergence.

Results: Here, we propose an efficient and reliable framework for estimating the parameters of ordinary differential equation models from qualitative data. In this framework, we derive a semi-analytical algorithm for gradient calculation of the optimal scaling method developed for qualitative data. This enables the use of efficient gradient-based optimization algorithms. We demonstrate that the use of gradient information improves performance of optimization and uncertainty quantification on several application examples. On average, we achieve a speedup of more than one order of magnitude compared to gradient-free optimization. In addition, in some examples, the gradient-based approach yields substantially improved objective function values and quality of the fits. Accordingly, the proposed framework substantially improves the parameterization of models from qualitative data.

Availability and implementation: The proposed approach is implemented in the open-source Python Parameter EStimation TOolbox (pyPESTO). pyPESTO is available at https://github.com/ICB-DCM/pyPESTO. All application examples and code to reproduce this study are available at https://doi.org/10.5281/zenodo.4507613.

Contact: jan.hasenauer@uni-bonn.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Systems biology models based on ordinary differential equations (ODEs) have enabled a profound understanding of many biological processes. Application areas include the study of signal transduction, gene regulation and metabolism (see, e.g. Bachmann *et al.*, 2011; Fröhlich *et al.*, 2018; Intosalmi *et al.*, 2016; Ozbudak *et al.*, 2004). The ODE models employed in these and other applications commonly comprise parameters, such as reaction rate constants or initial concentrations of biochemical species, which cannot be measured directly and therefore have to be inferred from experimental data (Mitra and Hlavacek, 2019). This is achieved by optimizing the agreement of the model simulation with experimental data, e.g. by minimizing the sum of squared distances or by maximizing a likelihood function. Various optimization methods have been developed to solve parameter estimation problems. This includes multi-start local optimization methods, global optimization methods and hybrid optimization methods (see Villaverde *et al.*, 2018 for detailed a discussion). Several empirical studies (Raue *et al.*, 2013; Schälte *et al.*, 2018; Villaverde *et al.*, 2018) have shown that optimization methods which use the gradient of the objective function with respect to the parameters tend to be more efficient than gradient-free optimization methods. Yet, while gradient calculation for objective functions using quantitative data is well established (Fröhlich *et al.*, 2017; Raue *et al.*, 2013; Sengupta *et al.*, 2014), respective tools for qualitative data are missing.

A spectrum of experimental setups and techniques provide qualitative observations, meaning that no exact quantitative relation to the concentration e.g. of biochemical species is available (Pargett and Umulis, 2013). Examples include imaging data for certain stainings (Brooks *et al.*, 2012; Pargett *et al.*, 2014), Förster resonance energy transfer (FRET) data (Birtwistle *et al.*, 2011) or phenotypic observations (Chen *et al.*, 2004). Although qualitative measurements do not provide numerical values, they contain valuable information to infer parameters (Pargett and Umulis, 2013). Therefore, several tailored parameter estimation approaches have been developed: (i) Toni *et al.* (2011) used simple distance functions and employed approximate Bayesian computing to explore the parameter space. (ii) Oguz *et al.* (2013) optimized the number of qualitative observations that were correctly captured by the model.

(iii) Mitra et al. used qualitative observations as static penalty functions (Mitra *et al.*, 2018, 2019) and proposed a statistically motivated objective function (Mitra and Hlavacek, 2020). (iv) Pargett *et al.* (2014) employed the concept of the optimal scaling approach (introduced by Shepard, 1962), which is based on finding the best possible quantitative representation (so-called surrogate data) of the qualitative observations. This is based on a hierarchical optimization problem, where in an outer optimization loop the model parameters are estimated, and in an inner optimization loop the optimal surrogate data is calculated. The approaches (i)–(iv) facilitate the extraction of information about the model parameters from qualitative data. Yet, the objective functions are either intrinsically discontinuous or an analytical formulation for the objective function gradient was unknown. Accordingly, only gradient-free optimization methods could be employed.

Here, we derive formulas for semi-analytical calculation of the gradients of the objective function arising in the optimal scaling approach. This allows for the use of gradient-based optimization in the outer loop (which optimizes the model parameters), and complements our previous work (Schmiester *et al.*, 2020) on the reformulation of the inner loop (which optimizes the surrogate data). We evaluate our gradient-based framework on several application examples and compare it to gradient-free optimization. We show that the proposed method yields accurate gradients, substantially accelerates parameter estimation and profile calculation and often yields improved final objective function values.

2 Materials and methods

2.1 Mathematical modeling of biological processes

We consider ODE models

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

for the dynamics of the concentrations of biochemical species, $x(t,\theta) \in \mathbb{R}^{n_x}$. The vector field $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$ describes the temporal evolution of the modeled species. The unknown model parameters are denoted by $\theta \in \mathbb{R}^{n_\theta}$ and the initial states at time point t_0 are given by $x_0(\theta)$. The total numbers of state variables and unknown parameters are denoted by n_x and n_θ, respectively.

The state variables $x(t,\theta)$ can be linked to experimental data by introducing an observation function $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}$, which maps the n_x state variables to an observable, $y(t,\theta) \in \mathbb{R}$, via

$$y(t,\theta) = h(x(t,\theta),\theta).$$

Note that we consider here the case of a single observable to simplify the notation. The general case is captured in the Supplementary Section S2.

Quantitative data provide information about the observable $y(t,\theta)$. Yet, the measurements are usually subject to measurement errors, which are often assumed to be i.i.d. additive Gaussian noise. In this case, the data $\bar{y}$ can be linked to the observables by

$$\bar{y}_i = y(t_i,\theta) + \varepsilon_i, \quad i = 1,\ldots,n_t$$

with measurement noise $\varepsilon_i \sim \mathcal{N}(0,\sigma_i^2)$, in which σ_i is the standard deviation for the ith observation and n_t denotes the number of time points. Alternative noise models can be used depending on the measurement characteristics (see discussion in Maier *et al.*, 2017).

Qualitative data do not provide information on the values of the observable, but rather on the ordering of different datapoints. Following the notation from Schmiester *et al.* (2020), we denote a qualitative readout as $z(t,\theta) \in \mathbb{R}$. For these readouts no exact quantitative relation to $y(t,\theta)$ is known and only monotonicity in the mapping between $y(t,\theta)$ and $z(t,\theta)$ can be assumed. A qualitative observation provides the ordering of two readouts. It is possible that these observations are indistinguishable ($z_i \approx z_j$), or that one observable is clearly larger or smaller ($z_i\{>,<\}z_j$). We group the different indistinguishable qualitative measurements in n_k different categories which are denoted by $\mathcal{C}_k$, $k = 1,\ldots,n_k$, i.e. $z_i, z_j \in \mathcal{C}_k \Rightarrow z_i \approx z_j$. We assume in the following that the categories are ordered as $\mathcal{C}_1 \prec \ldots \prec \mathcal{C}_{n_k}$.

2.2 Parameter estimation using qualitative data

In this study, we build upon the *optimal scaling* approach for parameter estimation using qualitative data (Pargett *et al.*, 2014; Schmiester *et al.*, 2020). The optimal scaling approach introduces *surrogate data* $\tilde{y}_i$, which are the best quantitative representations of the qualitative measurements $\bar{z}_i$. Therefore, for some parameter vector θ, the surrogate data $\tilde{y}_i$ aim to describe the model simulation $y(t_i,\theta)$ optimally, while fulfilling the ordering of the qualitative categories (Fig. 1A). To this end, we introduce intervals $[l_k,u_k]$ for each category $\mathcal{C}_k$, with lower and upper bounds l_k and u_k. The intervals are ordered, $u_k \leq l_{k+1}$, and the surrogate datapoints can be freely placed within the corresponding interval. As the bounds of the intervals and the surrogate data are a priori unknown, they are subject to optimization. Using a weighted sum of squared distances function, the corresponding optimization problem is

$$\begin{aligned} &\min_{\tilde{y},l,u}\left\{J := \sum_{i=1}^{n_t} w_i(\tilde{y}_i - y(t_i,\theta))^2\right\} \\ &\text{s.t. } \mathrm{l_{k(i)}} \leq \mathrm{\tilde{y}_i} \leq \mathrm{u_{k(i)}}, \mathrm{i} = 1,\ldots,\mathrm{n_t}, \\ &\quad u_k \leq l_{k+1}, k = 1,\ldots,n_k - 1, \end{aligned} \tag{2}$$

in which $k(i)$ is the index of the category of the surrogate datapoint $\tilde{y}_i$ and w_i are datapoint-specific weights. The weights are usually chosen such that the objective function value is independent of the scale of the simulation (Pargett *et al.*, 2014; Schmiester *et al.*, 2020). The first inequality constraint of (2) guarantees that the surrogate datapoints are placed inside the respective interval, and the second inequality constraint assures that the ordering of the categories is fulfilled.

To estimate the unknown model parameters θ, the surrogate data optimization (2) can be nested into the model parameter optimization (Fig. 1B), yielding the hierarchical optimization problem

$$\begin{aligned} &\min_\theta \sum_{i=1}^{n_t} w_i(\tilde{y}_i - y(t_i,\theta))^2 \\ &\text{s.t. } (\tilde{\mathrm{y}},\mathrm{l},\mathrm{u}) \text{ are a solution to } (2). \end{aligned} \tag{3}$$

Therefore, in each iteration of the outer optimization (of the model parameters θ), the inner constrained optimization problem (2) (for the surrogate data and category bounds) has to be solved. We have previously shown that the inner optimization problem can be simplified to improve efficiency and robustness by only estimating upper bounds u and determining l and $\tilde{y}$ analytically (Schmiester *et al.*, 2020). However, for the derivation of the gradient computation algorithm introduced in the next section we will consider the full optimization problem (2).

For ease of notation, we rewrite the optimization problem in matrix-vector notation. For this, we denote the collection of all inner optimization variables by $\tilde{\xi} = (\tilde{y},l,u)^T \in \mathbb{R}^{n_\xi}$ and the vector of simulations by $\xi(\theta) = (y(\theta),0,0)^T \in \mathbb{R}^{n_\xi}$ which is filled with zeros such that $\tilde{\xi}$ and ξ have the same dimension. Here, $n_\xi = n_t + 2n_k$ is the number of inner optimization variables. With this, we can rewrite the optimization problem (3) as a hierarchical problem of the form

$$\min_\theta J(\xi(\theta),\tilde{\xi}^*(\theta)) \tag{4}$$

$$\begin{aligned} &\text{s.t. } \tilde{\xi}^*(\theta) = \mathrm{argmin}_{\tilde{\xi}}\, \mathrm{J}(\xi(\theta),\tilde{\xi}) \\ &\text{s.t. } \mathrm{C}\tilde{\xi} \leq 0, \end{aligned} \tag{5}$$

in which $\tilde{\xi}^*(\theta)$ is the optimal value of the inner optimization for a given θ. Note that, because the optimal solution to the inner optimization problem (5) depends on the model parameters θ, the optimal surrogate data $\tilde{\xi}^*$ is directly dependent on θ. The objective function is given by

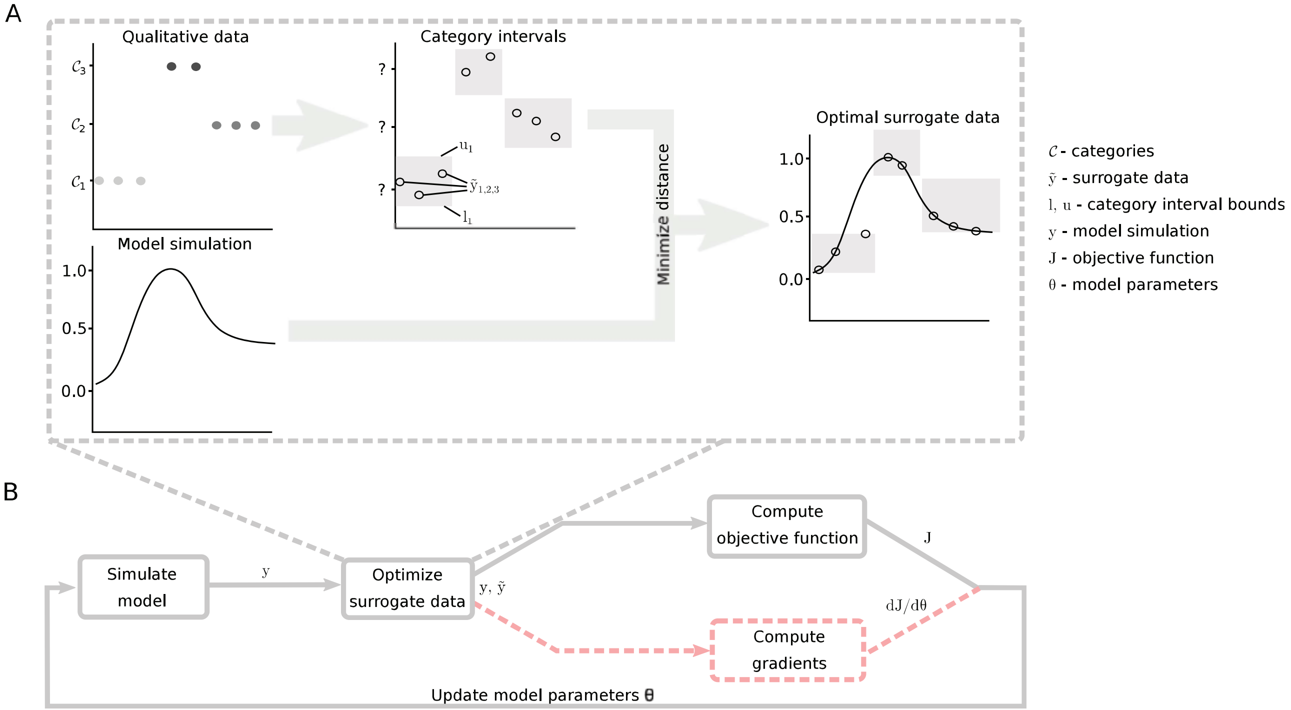

Fig. 1. Illustration of the optimal scaling approach. (**A**) Qualitative data and model simulation are integrated by introducing intervals for the qualitative categories $\mathcal{C}$, which have to be optimized to minimize the difference to the model simulation. Surrogate data ($\tilde{y}$) can then be placed optimally inside the intervals $[l, u]$. (**B**) Model parameters θ are updated iteratively during parameter estimation. For each trial parameter vector, the model output y is simulated. Then, the surrogate data is optimized and used to compute the objective function J and, if required by the employed optimizer, gradients $dJ/d\theta$

$$J(\xi(\theta), \tilde{\xi}) = \left(\tilde{\xi} - \xi(\theta)\right)^T W \left(\tilde{\xi} - \xi(\theta)\right).$$

The matrix $C \in \mathbb{R}^{n_c \times n_{\tilde{\xi}}}$ encodes the inequality constraints of the inner problem, with the total number of constraints n_c, and the weight matrix W is given by

$$W = \begin{pmatrix} \mathrm{diag}(w) & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n_{\tilde{\xi}} \times n_{\tilde{\xi}}}.$$

W is augmented with zeros such that the bounds l and u do not contribute to the objective function J and the dimensions of W and C are consistent, which is necessary for the following calculations. An illustrating example of the reformulation is given in the Supplementary Section S3.

2.3 Gradient computation for the optimal scaling objective function

In this section, we derive an algorithm to calculate the gradients of the optimal scaling objective function, based on ideas from the field of bi-level optimization (Fiacco, 1976; Kolstad and Lasdon, 1990). We are interested in calculating the derivative of the objective function J with respect to the model parameters θ, evaluated at the optimal surrogate data $\tilde{\xi}^*(\theta)$, which is given by

$$\begin{aligned} \frac{dJ}{d\theta}\Big|_{\xi(\theta),\tilde{\xi}^*(\theta)} = & \frac{\partial J}{\partial \xi}\Big|_{\xi(\theta),\tilde{\xi}^*(\theta)} \frac{\partial \xi(\theta)}{\partial \theta} \\ & + \frac{\partial J}{\partial \tilde{\xi}}\Big|_{\xi(\theta),\tilde{\xi}^*(\theta)} \frac{\partial \tilde{\xi}^*(\theta)}{\partial \theta}. \end{aligned} \tag{6}$$

All parts of (6) except for $\partial\tilde{\xi}^*(\theta)/\partial\theta$ can be easily calculated (see Supplementary Section S1 for details). To obtain $\frac{\partial \tilde{\xi}^*(\theta)}{\partial \theta}$, we introduce the Lagrangian function $\ell(\tilde{\xi}, \mu) = J(\xi(\theta), \tilde{\xi}) + \mu^T C\tilde{\xi}$, with Lagrange multipliers $\mu \in \mathbb{R}^{n_c}$. The necessary first order optimality conditions of problem (5) for a given θ are then given by the Karush-Kuhn-Tucker conditions (Boyd and Vandenberghe, 2004)

$$\nabla_{\tilde{\xi}} \ell(\tilde{\xi}^*(\theta), \mu(\theta)) = 2\left(\tilde{\xi}^*(\theta) - \xi(\theta)\right)^T W + \mu(\theta)^T C = 0 \tag{7}$$

$$\mu_i(\theta) C_i \tilde{\xi}^*(\theta) = 0 \tag{8}$$

$$C_i \tilde{\xi}^*(\theta) \leq 0 \tag{9}$$

$$\mu_i(\theta) \geq 0 \tag{10}$$

for $i = 1, \ldots, n_c$, where C_i is the ith row of C. Given the optimal values of the inner optimization variables $\tilde{\xi}^*(\theta)$, the Lagrange multiplier μ can be calculated by solving (7)–(10). To obtain $\partial\tilde{\xi}^*(\theta)/\partial\theta$, we calculate the derivative of the optimality conditions (7) and (8) w.r.t. θ_j, which results in

$$\begin{aligned} 2\left(\tfrac{\partial \tilde{\xi}^*(\theta)}{\partial \theta_j} - \tfrac{\partial \xi(\theta)}{\partial \theta_j}\right)^T W + \tfrac{\partial \mu(\theta)}{\partial \theta_j}^T C = 0 \\ \frac{\partial \mu_i(\theta)}{\partial \theta_j} C_i \tilde{\xi}^*(\theta) + \mu_i(\theta) C_i \frac{\partial \tilde{\xi}^*(\theta)}{\partial \theta_j} = 0. \end{aligned}$$

This yields a linear system of equations that needs to be solved for every parameter θ_j:

$$\begin{pmatrix} 2W & C^T \\ \mathrm{diag}(\mu)C & \mathrm{diag}(C\tilde{\xi}^*) \end{pmatrix} \begin{pmatrix} \dfrac{\partial \tilde{\xi}^*}{\partial \theta_j} \\ \dfrac{\partial \mu}{\partial \theta_j} \end{pmatrix} = \begin{pmatrix} 2W\dfrac{\partial \xi}{\partial \theta_j} \\ 0 \end{pmatrix}, \tag{11}$$

where we omitted the dependency of μ, ξ and $\tilde{\xi}^*$ on θ for simplicity of presentation. After solving (11), we can calculate the gradients of the objective function w.r.t. the model parameters θ using equation (6).

To summarize, the gradient computation scheme consists of the following steps:

1. Simulate the ODE (1) to obtain $\xi(\theta)$ and $\frac{\partial \xi}{\partial \theta}$.
2. Calculate optimal surrogate data $\tilde{y}$ and category bounds l, u by solving (2) or a reformulation of this problem (Schmiester *et al.*, 2020).
3. Solve the optimality conditions (7)–(10) for the Lagrange multiplier μ.
4. Solve the linear system of Equations (11) to obtain $\frac{\partial \tilde{\xi}^*}{\partial \theta}$.
5. Evaluate the gradient $\frac{dJ}{d\theta}$ of the objective function using Equation (6).

In practice, it is sometimes preferable to choose weights, that are dependent on the parameters θ or the simulation ξ. In addition, minimal sizes on the intervals $s(\theta, \xi(\theta)) \leq u_k - l_k$ and on the gaps between the intervals $g(\theta, \xi(\theta)) \leq l_{k+1} - u_k$, which can also depend on θ and ξ, can be imposed. Assuming that $g(\theta, \xi(\theta))$, $s(\theta, \xi(\theta))$ and $W(\theta, \xi(\theta))$ are differentiable functions, similar formulas for gradient computation can be derived (see Supplementary Section S1). Collecting minimal interval gaps g and interval sizes s in the vector d, yielding the inequality constraints $C\tilde{\xi} + d(\theta, \xi(\theta)) \leq 0$, we obtain the linear system of equations

$$\begin{pmatrix} 2W & C^T \\ \mathrm{diag}(\mu)C & \mathrm{diag}(C\tilde{\xi}^* + d) \end{pmatrix} \begin{pmatrix} \dfrac{\partial \tilde{\xi}^*}{\partial \theta_j} \\ \dfrac{\partial \mu}{\partial \theta_j} \end{pmatrix} = \begin{pmatrix} 2W\dfrac{\partial \xi}{\partial \theta_j} - 2\left(\dfrac{\partial W}{\partial \xi}\dfrac{\partial \xi}{\partial \theta_j} + \dfrac{\partial W}{\partial \theta_j}\right)(\tilde{\xi}^* - \xi) \\ -\mathrm{diag}(\mu)\left(\dfrac{\partial d}{\partial \xi}\dfrac{\partial \xi}{\partial \theta_j} + \dfrac{\partial d}{\partial \theta_j}\right) \end{pmatrix}. \qquad (12)$$

The linear systems (11) and (12) are sparse and can be solved efficiently. These systems can be solved and the gradients can be calculated for arbitrary parameter vectors as long as the solution of the ODE model is available.

2.4 Implementation

We implemented the gradient calculation method in the Python Parameter EStimation TOolbox (pyPESTO) (Schälte *et al.*, 2020). The qualitative data can be defined using an extension to the PEtab format, which is a standardized format for the definition of parameter estimation problems (Schmiester *et al.*, 2021). Model simulation was carried out using the AMICI toolbox (Fröhlich *et al.*, 2020), which internally interfaces the Sundials solver CVODES (Hindmarsh *et al.*, 2005). Parameter estimation was performed using multi-start local optimization with 500 starts per model and method. To guarantee comparability, optimizations were started from the same initial parameters for each method. For gradient-free optimization we used the Powell algorithm implemented in the SciPy package (Jones *et al.*, 2001), which performed well among SciPy's gradient-free optimizers in a previous study using the optimal scaling approach (Schmiester *et al.*, 2020). For gradient-based optimization we used SciPy's L-BFGS-B algorithm, which is the default optimizer in pyPESTO. For more details on the implementation, we refer to the Supplementary Section S4.

3 Results

3.1 Model overview

To analyze the gradient algorithm and compare it to gradient-free optimization, we considered six models. We included one toy model and five larger application examples. An overview of the models and datasets used for parameter estimation is given in Table 1.

T1 is a small model used for illustration. Models M1–M5 are published models with experimental data and describe different biological processes. They are taken from a collection of parameter estimation problems in the PEtab format, which is based on the benchmark collection by Hass *et al.* (2019). The models comprise different numbers of datapoints and unknown parameters and were originally calibrated on quantitative measurements. These quantitative measurements were converted to qualitative observations based on their ordering, where we assumed that measurements are indistinguishable, if their numeric values are equal. In addition, we assumed that data was comparable within an observable but not across observables, leading to one optimal scaling problem per observable that needed to be solved.

3.2 Semi-analytical approach yields accurate gradients

As the gradient computation algorithm involves numerically solving a linear system of equations, we first evaluated the accuracy of the obtained solution. We considered the model T1 and compared the semi-analytical gradients with gradients obtained using a finite difference approach at different parameter vectors (Fig. 2A). The analysis revealed that for all tested parameters the approaches yielded almost identical gradients (with absolute differences $< 10^{-9}$). We additionally compared computation times of finite differences and our semi-analytical gradient algorithm for models M1–M5, which showed that finite differences require substantially more computation time (Supplementary Fig. S1). Therefore, we restrict ourselves to gradient computations using the semi-analytical approach in the following analysis.

3.3 Gradient information increases optimizer efficiency

To illustrate the differences of gradient-free and gradient-based optimization, we estimated parameters for both optimizers starting from the same initial parameters. As the model T1 only contains two parameters, we inspected the whole objective function landscape and the respective optimizer trajectories (Fig. 2B and C). While the gradient-free optimizer used a rather naive updating scheme (Fig. 2B) that often moved along sub-optimal directions, the gradient-based optimizer moved towards the optimal point within a few iterations (Fig. 2C), indicating a potential advantage of using gradient-based optimizers. Figure 2B and C revealed flat regions in the objective function. We additionally computed the landscapes of the gradients to show that they are correctly calculated close to zero in these areas (Supplementary Fig. S5).

To assess the performance of gradient-based and gradient-free optimization in a realistic setting, we performed multi-start local optimization for the application examples M1–M5. The results show

Table 1. Key numbers of the different considered models and datasets used for parameter estimation

Model	No. of states	No. of parameters	No. of observables	No. of datapoints	No. of categories	Description	Reference
T1	6	2	2	18	2×3	RAF inhibition	Mitra *et al.* (2018)
M1	7	9	1	23	19	Infectious diseases dynamics	Rahman *et al.* (2016)
M2	8	6	3	48	3×16	STAT5 dimerization	Boehm *et al.* (2014)
M3	8	18	1	58	43	Transcriptional regulation	Elowitz and Leibler (2000)
M4	14	18	8	205	6–38	IL13-induced signaling	Raia *et al.* (2011)
M5	6	12	8	72	9–11	RAF-MEK-ERK signaling	Fiedler *et al.* (2016)

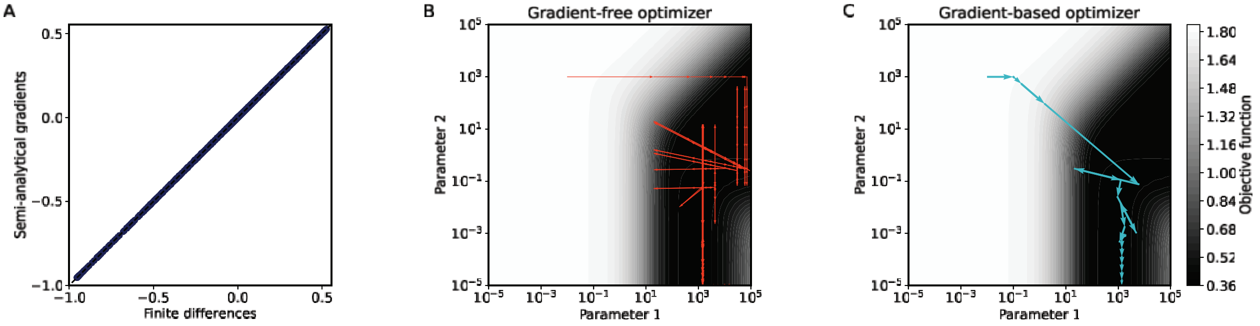

Fig. 2. (**A**) Absolute gradients of the two parameters of model T1 evaluated at 2500 uniformly sampled parameter vectors using the semi-analytical approach and central finite differences. (**B and C**) Objective function landscape and optimizer trajectories of a gradient-free (B) and a gradient-based (C) optimizer

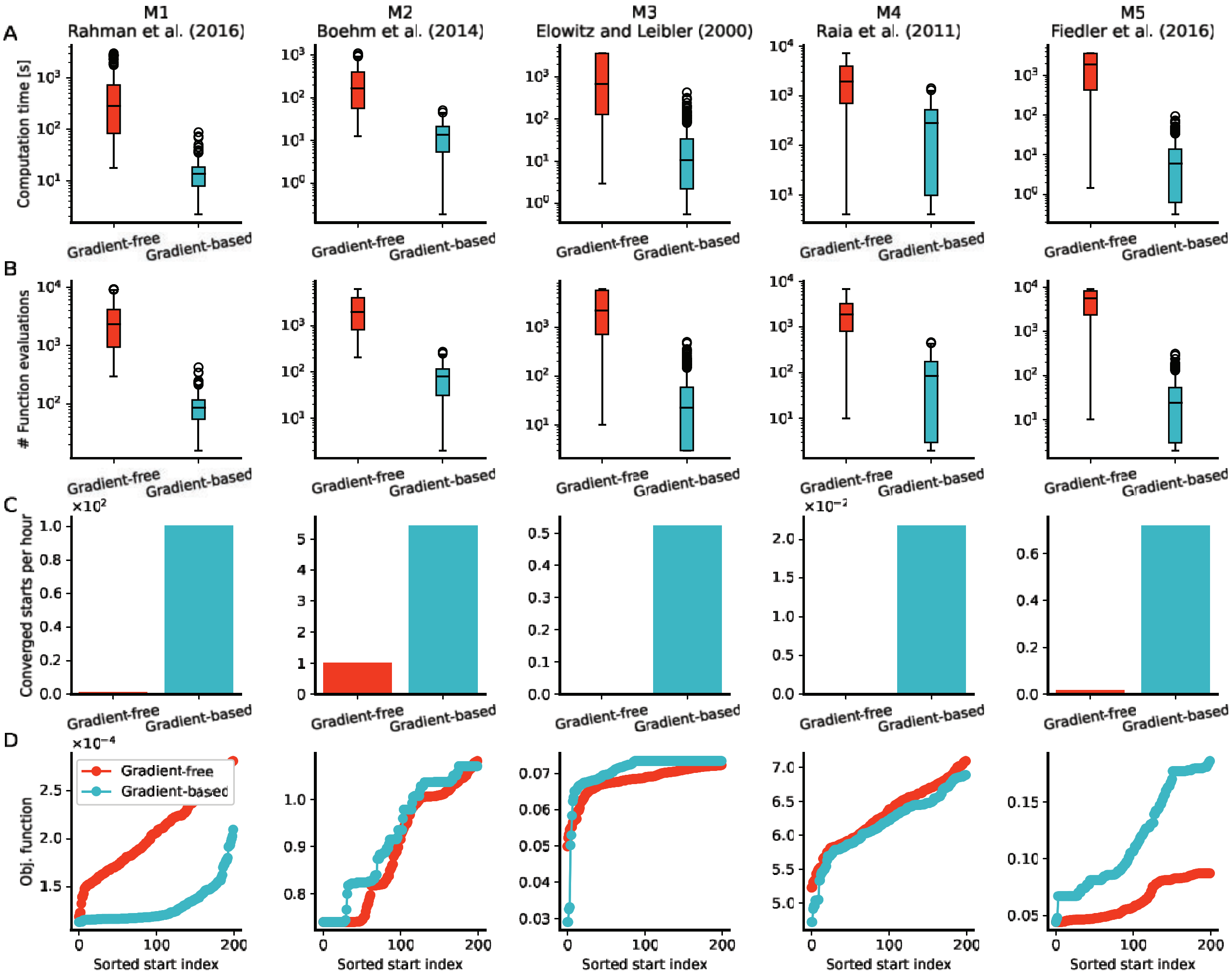

Fig. 3. Optimization results for all models using gradient-free and gradient-based optimization for 500 local optimizations. (**A**) Computation times until the optimizer terminates per local optimization. (**B**) Number of function evaluations per local optimization. (**C**) Converged starts per hour. A start is considered converged, if the absolute difference to the overall best value is less than 10^{-4} (see Supplementary Fig. S2 for results using different thresholds). (**D**) Waterfall plots for all five considered models using gradient-free and gradient-based optimization. Best 200 starts out of a total of 500 are shown. See Supplementary Figure S3 for results of all starts

considerably reduced computation times of the gradient-based optimization for all models (Fig. 3A). Depending on the considered model, the median CPU times are reduced by 1–2 orders of magnitude.

As illustrated in Figure 2, the gradient-based optimizer used a more intelligent updating scheme during optimization, resulting in reduced numbers of objective function evaluations. The reduction in function evaluations could also be observed for all employed models (Fig. 3B). Even though a single evaluation is more costly when gradients need to be calculated (Supplementary Fig. S1), the reduced number of necessary evaluations outweighs this, explaining the improved computation times.

In addition to the computation times, we also incorporated the final objective function values into our analysis by considering the number of local optimization runs which achieved values close to the overall best value (Fig. 3C). This revealed substantially improved efficiency of the gradient-based optimizer for all models. This result could be observed independent of the threshold for convergence to the optimal point (Supplementary Fig. S2). The improved efficiency of the gradient-based optimizer could be observed even for the models M2 and M5, for which the gradient-free optimizer found the optimal value more often (Fig. 3D). A possible explanation for the sometimes larger number of converged starts observed for gradient-free optimization methods could be flat regions in the objective function. This would result in a vanishing gradient, which in turn could lead to a termination of the gradient-based optimization.

3.4 Gradient-based optimization yields improved model fits

The waterfall plots show that the gradient-based optimization yielded equal (M1, M2 and M5) or even better (M3 and M4) final objective function values compared to gradient-free optimization (Fig. 3D and Supplementary Fig. S3). We simulated the models for which larger differences in the best obtained objective function values were observed to analyze if the different parameters resulted in substantial differences in the model fits. Especially for model M3 we observed a better representation of the data with the parameters from gradient-based optimization (Fig. 4). Indeed, only the parameters obtained using the gradient-based optimization correctly captured the oscillations of the measurements.

For the model M4, we also observed smaller improvements in the model fits for some observables, when using the parameters obtained from gradient-based optimization (Supplementary Fig. S4). The improved objective function values are likely owing to the additional information coming from the gradient that helps the optimizer move towards an optimal point, in particular for high-dimensional problems. Indeed, the models M3 and M4, were this improvement was observed, are the problems with the largest number of estimated parameters.

3.5 Gradient-based approach facilitates uncertainty quantification

Qualitative data is often considered to be less informative than quantitative measurements. As this can result in reduced parameter identifiability, it is even more important to assess the uncertainties associated with the estimated parameters when using qualitative measurements. For uncertainty analysis, we used objective function ratio profiles analogously to profile likelihoods in the case of a likelihood function (Raue *et al.*, 2009). While the objective function differences in the profiles cannot easily be interpreted statistically, they can still be valuable to indicate uncertainties of the estimated parameters. As a proof of concept, we calculated objective function profiles for the model M2 (Fig. 5A). The gradient-based approach yielded mostly smooth profiles indicating that several parameters could be well identified using the qualitative dataset. In contrast, the gradient-free approach resulted in several discontinuities in the profiles probably caused by impaired optimization. This shows that only the gradient-based approach was able to yield meaningful profiles for this model. In addition to the improved profiles, the gradient-based approach required on average an order of magnitude less computation time than the gradient-free optimizer (Fig. 5B).

4 Discussion

Qualitative data can contain valuable information for parameter estimation but current methods to integrate such data are computationally demanding and more efficient algorithms are required. Here, we developed a framework for semi-analytical computation of gradients for the optimal scaling approach. We validated the accuracy of the obtained gradients by comparing them to finite differences and assessed the advantage of using gradient information on five application examples by performing optimization with a gradient-free and a gradient-based algorithm. This revealed speedups of more than one order of magnitude using the gradient-based approach. In addition, the gradient-based algorithm resulted in equal or even improved final objective function values and model fits. The gradient-based approach was further used to reliably calculate objective function profiles to assess the uncertainty of parameter estimates when using qualitative data.

A linear system needs to be solved for every parameter to obtain the gradients of the optimization problem. As this is the most time consuming part during gradient computation, more efficient approaches could further decrease computation times. The computation time could for instance be reduced by splitting it into active and inactive constraints (Kolstad and Lasdon, 1990) or by parallelizing gradient computation over the parameters. Complementary to this, it remains open whether—similar to other hierarchical optimization approaches (Schmiester *et al.*, 2019) – adjoint sensitivity analysis can be used to further accelerate optimization (Fröhlich *et al.*, 2017). Another possible extension would be the derivation of second-order derivatives that could be used for parameter estimation and profile calculation (Stapor *et al.*, 2018).

The application examples considered here were all based on synthetic qualitative data by transforming real quantitative measurements to qualitative observations based on their ordering. The application to real-world qualitative data, as given e.g. in Mitra *et al.* (2018) and Pargett *et al.* (2014), is left for future work. For this, it will often be advantageous to combine qualitative and quantitative measurements. In Mitra *et al.* (2018) it has been shown, that complementing quantitative data with qualitative data can improve parameter identifiability. This can in principle be done by formulating a similar objective function for quantitative data by replacing the surrogate data with the measured quantitative values. As the gradient for quantitative data can easily be calculated, an overall objective function value and gradient can be obtained by summing up the respective values for qualitative and quantitative data. The integration of both datatypes could further be improved by defining a

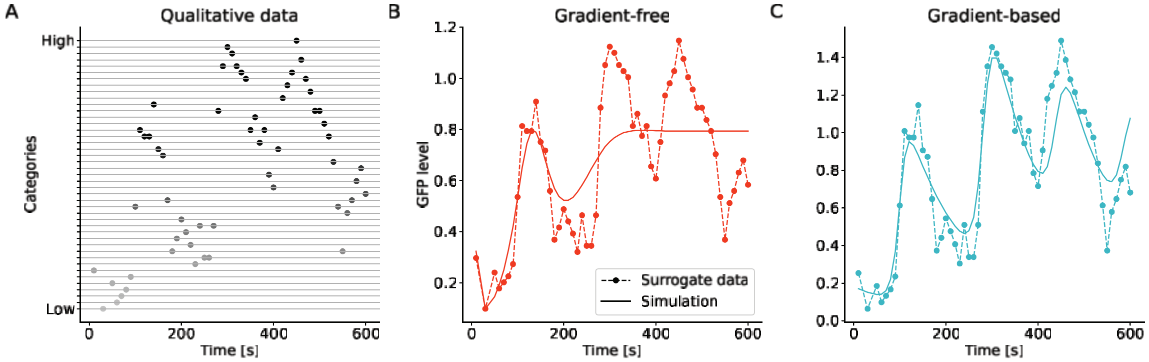

Fig. 4. Fits of model and qualitative data for model M3. (**A**) Qualitative observations. The gray-scale and horizontal lines indicate the categories to which the qualitative datapoints belong. (**B and C**) Simulated data and optimal surrogate data for the best parameters from gradient-free (B) and gradient-based (C) optimization. The surrogate data are ordered according to the qualitative data (A), but can have different numeric values in B and C as they are optimized to the respective model simulation

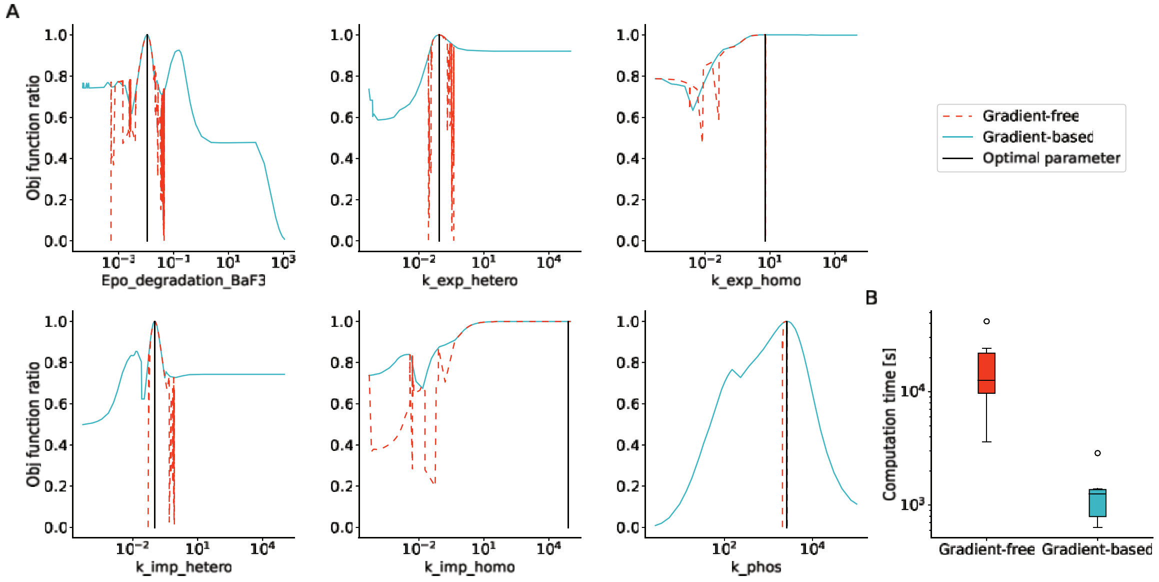

Fig. 5. (**A**) Objective function ratio profiles for the six parameters of model M2 using gradient-free and gradient-based optimization. Shown is the ratio of the overall best objective function value and the value along the profile. Profiles were initiated on the overall best found parameter. (**B**) Boxplots for computation times for the profiles per parameter

proper likelihood function, which seems to be problematic in the optimal scaling approach. A likelihood function for qualitative data would also facilitate the use of Bayesian problem formulations, which are not feasible with the here considered objective function.

In conclusion, we developed a framework to compute gradient information for parameter estimation problems that include qualitative data and showed that it substantially improves computational efficiency. The open-source implementation of the approach we provide will facilitate reusability and might improve the usage of qualitative data for the parameterization of quantitative models.

Author Contributions

L.S. and J.H. derived the theoretical foundation. L.S. wrote the implementations and performed the case study. L.S., J.H. and D.W. analyzed the results. All authors wrote and approved the final manuscript.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program (CanPathPro; Grant No. 686282; J.H., D.W., L.S.) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2047 & EXC 2151 (J.H.).

Conflict of Interest: none declared.

References

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

Birtwistle,M.R. *et al.* (2011) Linear approaches to intramolecular förster resonance energy transfer probe measurements for quantitative modeling. *PLoS One*, **6**, e27823.

Boehm,M.E. *et al.* (2014) Identification of isoform-specific dynamics in phosphorylation-dependent stat5 dimerization by quantitative mass spectrometry and mathematical modeling. *J. Proteome Res.*, **13**, 5685–5694.

Boyd,S. and Vandenberghe,L. (2004) *Convex Optimisation*. Cambridge University Press, UK.

Brooks,A. *et al.* (2012) BMP signaling in wing development: a critical perspective on quantitative image analysis. *FEBS Lett.*, **586**, 1942–1952.

Chen,K.C. *et al.* (2004) Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, **15**, 3841–3862.

Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.

Fiacco,A.V. (1976) Sensitivity analysis for nonlinear programming using penalty methods. *Math. Program.*, **10**, 287–311.

Fiedler,A. *et al.* (2016) Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints. *BMC Syst. Biol*, **10**, 80.

Fröhlich,F. *et al.* (2017) Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, **13**, e1005331.

Fröhlich,F. *et al.* (2018) Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Syst.*, **7**, 567–579.e6.

Fröhlich,F. *et al.* (2021) AMICI: high-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics, btab227*.

Hass,H. *et al.* (2019) Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, **35**, 3073–3082.

Hindmarsh,A.C. *et al.* (2005) SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM Ttans. Math. Softw.*, **31**, 363–396.

Intosalmi,J. *et al.* (2016) Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks. *Bioinformatics*, **32**, i288–i296.

Jones,E. *et al.* (2001) SciPy: Open source scientific tools for Python.

Kolstad,C.D. and Lasdon,L.S. (1990) Derivative evaluation and computational experience with large bilevel mathematical programs. *J. Optim. Theory Appl.*, **65**, 485–499.

Maier,C. *et al.* (2017) Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, **33**, 718–725.

Mitra,E.D. and Hlavacek,W.S. (2019) Parameter estimation and uncertainty quantification for systems biology models. *Curr. Opin. Syst. Biol.*, **18**, 9–18.

Mitra,E.D. and Hlavacek,W.S. (2020) Bayesian inference using qualitative observations of underlying continuous variables. *Bioinformatics*, **36**, 3177–3184.

Mitra,E.D. *et al.* (2018) Using both qualitative and quantitative data in parameter identification for systems biology models. *Nat. Commun.*, **9**, 3901.

Mitra,E.D. *et al.* (2019) PyBioNetFit and the biological property specification language. *iScience*, **19**, 1012–1036.

Oguz,C. *et al.* (2013) Optimization and model reduction in the high dimensional parameter space of a budding yeast cell cycle model. *BMC Syst. Biol.*, **7**, (53.

Ozbudak,E.M. *et al.* (2004) Multistability in the lactose utilization network of *Escherichia coli*. *Nature*, **427**, 737–740.

Pargett,M. and Umulis,D.M. (2013) Quantitative model analysis with diverse biological data: applications in developmental pattern formation. *Methods*, **62**, 56–67.

Pargett,M. *et al.* (2014) Model-based analysis for qualitative data: an application in drosophila germline stem cell regulation. *PLoS Comput. Biol.*, **10**, e1003498.

Rahman,S.M.A. *et al.* (2016) Impact of early treatment programs on HIV epidemics: an immunity-based mathematical model. *Math. Biosci.*, **280**, 38–49.

Raia,V. *et al.* (2011) Dynamic mathematical modeling of il13-induced signaling in Hodgkin and primary mediastinal b-cell lymphoma allows prediction of therapeutic targets. *Cancer Res.*, **71**, 693–704.

Raue,A. *et al.* (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929.

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

Schälte,Y. *et al.* (2018) Evaluation of derivative-free optimizers for parameter estimation in systems biology. *IFAC-PapersOnLine*, **51**, 98–101.

SchälteY., *et al.* (2020) ICB-DCM/pyPESTO: pyPESTO 0.0.11. Zenodo. doi: 10.5281/zenodo.3715448 (18 March 2020, date of deposit).

Schmiester,L. *et al.* (2019) Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics*, **36**, 594–602.

Schmiester,L. *et al.* (2020) Parameterization of mechanistic models from qualitative data using an efficient optimal scaling approach. *J. Math. Biol.*, **81**, 603–623.

Schmiester,L. *et al.* (2021) PEtab—interoperable specification of parameter estimation problems in systems biology. *PLoS Comput. Biol.*, **17**, e1008646.

Sengupta,B. *et al.* (2014) Efficient gradient computation for dynamical models. *NeuroImage*, **98**, 521–527.

Shepard,R.N. (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, 125–140.

Stapor,P. *et al.* (2018) Optimization and profile calculation of ODE models using second order adjoint sensitivity analysis. *Bioinformatics*, **34**, i151–i159.

Toni,T. *et al.* (2011) From qualitative data to quantitative models: analysis of the phage shock protein stress response in *Escherichia coli*. *BMC Syst. Biol.*, **5**, (69.

Villaverde,A.F. *et al.* (2018) Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics*, **35**, bty736.