*Bioinformatics*, 2024, **40**, i558–i566
https://doi.org/10.1093/bioinformatics/btae210
**ISMB 2024**

OXFORD

# Efficient parameter estimation for ODE models of cellular processes using semi-quantitative data

**Domagoj Dorešić [1,2], Stephan Grein [1], Jan Hasenauer [1,2,3,*]**

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

*Corresponding author. Life and Medical Sciences (LIMES) Institute, University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany. E-mail: jan.hasenauer@uni-bonn.de (J.H.)

## Abstract

**Motivation:** Quantitative dynamical models facilitate the understanding of biological processes and the prediction of their dynamics. The parameters of these models are commonly estimated from experimental data. Yet, experimental data generated from different techniques do not provide direct information about the state of the system but a nonlinear (monotonic) transformation of it. For such semi-quantitative data, when this transformation is unknown, it is not apparent how the model simulations and the experimental data can be compared.

**Results:** We propose a versatile spline-based approach for the integration of a broad spectrum of semi-quantitative data into parameter estimation. We derive analytical formulas for the gradients of the hierarchical objective function and show that this substantially increases the estimation efficiency. Subsequently, we demonstrate that the method allows for the reliable discovery of unknown measurement transformations. Furthermore, we show that this approach can significantly improve the parameter inference based on semi-quantitative data in comparison to available methods.

**Availability and implementation:** Modelers can easily apply our method by using our implementation in the open-source Python Parameter EStimation TOolbox (pyPESTO) available at https://github.com/ICB-DCM/pyPESTO.

## 1 Introduction

The use of mechanistic mathematical models has greatly contributed to the understanding of biological processes at the cellular (Kitano 2002, Schöberl *et al.* 2009), patient (Fey *et al.* 2015, Hass *et al.* 2017) and population level (Giordano *et al.* 2020, Zhao and Chen 2020). In particular, mechanistic ordinary differential equation (ODE) models are used for a broad spectrum of applications, ranging from cellular signaling, metabolism, and gene regulation over pharmacokinetics and -dynamics to the spread of diseases. However, ODE models often contain parameters that cannot be measured directly. Instead, the parameters have to be estimated from experimental data (Mitra and Hlavacek 2019). This is commonly achieved by numerical optimization of an objective function, which quantifies how well the model simulations fit the given experimental data, such as the likelihood function.

The experimental data used for parameter estimation are collected using a broad spectrum of experimental techniques. For example, early studies in the field of systems biology employed well-calibrated Western blot experiments and performed an in-depth assessment of the mapping of concentration to measured intensities (Kreutz *et al.* 2007). In this case, the data were ensured to fall within the linear regime of the experimental technique, and often even absolute quantification was performed. However, many, even state–of–the–art, measurement techniques do not ensure a linear relationship between the abundance of the biochemical quantities of interest and the measured output (Fig. 1). Well-known examples include fluorescence microscopy data such as Förster resonance energy transfer (FRET) data (Birtwistle *et al.* 2011), optical density (OD) measurement (Stevenson *et al.* 2016) and imaging data for certain stainings (Pargett *et al.* 2014). In addition, many experimental techniques suffer from lower limits of detection and/or saturation effects.

Quantitative data are easy to use for the parameterization of ODE models and the same holds for data that are collected in the linear regime of measurement devices. This is showcased in a large number of published articles [see, e.g. (Hass *et al.* 2019) for a collection of models and datasets]. In fact, there are custom methods for experimental data for which a linear mapping with unknown scaling and offset parameters can be assumed (Loos *et al.* 2018, Schmiester *et al.* 2019). If the linearity assumption is not fulfilled, it is usually assumed that the mapping from biochemical quantities of interest to measured output is monotone. This monotonicity ensures that the ordering is preserved and allows the use of approaches for ordinal data, such as the optimal scaling method (Shepard 1962). For ODE models, this approach has recently been accelerated by using a reformulation of the optimization problem (Schmiester *et al.* 2020) and gradient information (Schmiester *et al.* 2021b). However, in this approach, all quantitative information is discarded and the defined objective function is not based on probabilistic grounds, disallowing any uncertainty analysis.

In this manuscript, we introduce a spline-based approach to use semi-quantitative data—which are obtained using an experimental technique with a nonlinear but monotone

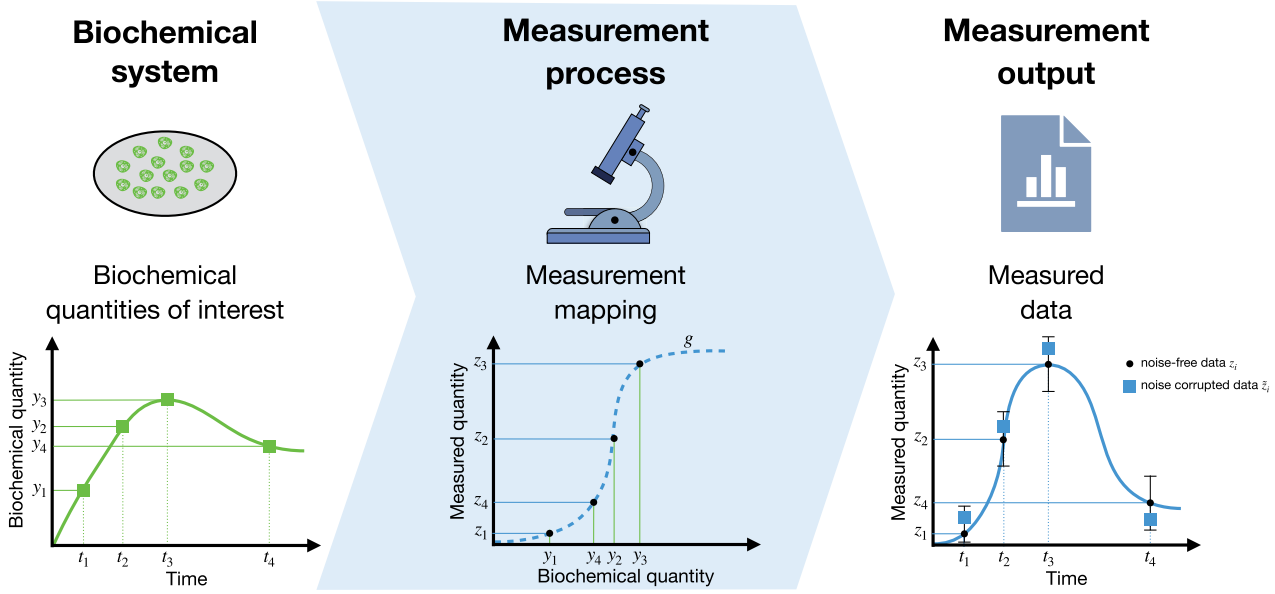


**Figure 1.** A nonlinear measurement mapping. (Biochemical system) True values of a biochemical quantity of interest $\{(t_i, y_i)\}_{i=1}^4$. (Measurement process) A measurement process can introduce unknown nonlinear data mappings (dashed blue line). In that case, a mapping function transforms the biochemical quantities $\{y_i\}_{i=1}^4$ and yields nonlinearly mapped measured quantities $\{z_i = g(y_i)\}_{i=1}^4$. (Measurement output) The measurement quantities $\{z_i\}_{i=1}^4$ are corrupted by noise, resulting in a noise-corrupted dataset $\mathcal{D} = \{(t_i, \tilde{z}_i)\}_{i=1}^4$

mapping—for parameter estimation. We assume that the measurement mappings are increasing monotonically. This is frequently observed in experiments: values with a larger measurement value are assumed to correspond to larger biochemical quantities of interest. The method reconstructs the unknown mapping function using a statistically coherent formulation that facilitates uncertainty analysis. We demonstrate the credibility of the proposed approach as a tool for uncovering measurement mapping shapes. To illustrate the parameter inference capabilities of the method, we benchmark its performance with a collection of published models. Furthermore, we derive formulas for the analytical calculation of the gradients of the objective function in hierarchical optimization. To evaluate this optimization framework, we compare its efficiency with alternative approaches.

## 2 Materials and methods

### 2.1 Mechanistic modeling of biological systems

We consider models of biological processes based on systems of ODEs:

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

in which the temporal evolution of the state variables $x(t,\theta) \in \mathbb{R}^{n_x}$ is determined by the vector field $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \times \mathbb{R}_+ \to \mathbb{R}^{n_x}$ with the unknown mechanistic model parameters $\theta \in \mathbb{R}^{n_x}$. State variables can, e.g. describe protein concentrations at the level of cellular processes or groups of individuals at the level of population modeling. The parameters $\theta$ usually consist of kinetic rate constants and initial species conditions. In cellular models, often not all metabolites are measurable, or in some cases only the sum of concentrations of multiple metabolites can be observed. These measured properties of a model are its observables, denoted as $y \in \mathbb{R}^{n_y}$,

$$y = h(x,\theta), \quad \tilde{y} = y + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0,\sigma), \tag{2}$$

in which $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_y}$ denotes the observation map which models the dependence of the observables on the model state variables and unknown mechanistic parameters, $\sigma \in \mathbb{R}_+^{n_t}$ is a noise parameter, and $\tilde{y} \in \mathbb{R}^{n_y}$ are noise-corrupted measurements. The dimensionalities of the state, parameter, and observable vector are denoted by $n_x$, $n_\theta$, and $n_y$, respectively. The number of time points is denoted by $n_t$.

#### 2.1.1 Linear semi-quantitative (relative) observables

Most measurement techniques provide only relative information on the biochemical quantity of interest. In this case, to obtain values comparable to the measured quantities, those observables need to be rescaled by scaling factors $a$ and offsets $b$. This is the case, for instance, for well-calibrated Western blot measurements, where the modeled protein concentrations have to be rescaled to be comparable to the optical density measurements. Most often, an additive Gaussian distributed noise model is assumed. Then the full relationship between measured and biochemical quantities of a relative observable is given by:

$$\tilde{z}_{ik} = \underbrace{a_i h_i(x(t_k,\theta),\theta) + b_i}_{:= g_i(h_i(x(t_k,\theta)))} + \varepsilon_{ik}, \tag{3}$$

$$\text{with } \varepsilon_{ik} \sim \mathcal{N}(0,\sigma_{ik}^2), \; g_i(x) = a_i \cdot x + b_i$$

in which $i$ is the observable index, $k$ is the time index, and $g_i : \mathbb{R} \to \mathbb{R}$ is a scaled and offset (affine) measurement mapping from the $i$th observable, $h_i(x(t_k,\theta))$, to its measurement. The scaling factors, offsets, and noise parameters of the $i$th observable are denoted as $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}$, and $\sigma_i \in \mathbb{R}_+^{n_t}$, respectively. These parameters are often unknown and need to be

estimated along with all other unknown parameters of the model.

### 2.1.2 Nonlinear semi-quantitative observables

In some cases, the measurement process induces a nonlinear mapping between the biochemical quantities of interest and the measured quantities. A common example is FRET measurements, which will be further investigated in the Section 3. Assuming an additive Gaussian distributed noise model the relationship is given by:

$$\tilde{z}_{ik} = g_i(h_i(x(t_k,\theta),\theta)) + \varepsilon_{ik}, \tag{4}$$

$$\text{with } \varepsilon_{ik} \sim \mathcal{N}(0,\sigma_{ik}^2)$$

in which $g_i : \mathbb{R} \to \mathbb{R}$ is a nonlinear measurement mapping from the $i$th observable, $h_i(x(t_k,\theta))$, to its measurement. The form and parameterization of nonlinear measurement mappings $g_i(h_i(x,\theta))$ are application-dependent. We provide examples in the Section 3. Measurement mappings are often unknown and need to be modeled in some way.

As this study considers the data-driven uncovering of measurement mappings, we select a class of approximations. Specifically, we consider the approximation of measurement mappings $g_i(h_i(x,\theta))$ with monotone piecewise linear splines $s_i : \mathbb{R} \times \mathbb{R}^{n_\xi^i} \to \mathbb{R}$, in which $i$ is the observable index. For the simplicity of further calculation, we parameterize the splines using the differences between the heights of neighboring spline knots $\{\xi_{ij}\}_{j=1}^{n_\xi^i}$:

$$s_i(x,\xi_i) := \begin{cases} \dfrac{x}{c_{i1}} \cdot \xi_1, & x \le c_{i1} \\ \dfrac{x - c_{i(j-1)}}{\Delta_c^i} \xi_{ij} + \sum_{l=1}^{j-1} \xi_{il}, & c_{i(j-1)} \le x \le c_{ij} \\ \xi_{in_\xi^i}, & x > c_{in_\xi^i} \end{cases} \tag{5}$$

in which $\{c_{ij}\}_{j=1}^{n_\xi^i}$ are the knot bases, and $n_\xi^i$ is the number of spline knots for the $i$th observable. Since $g_i$ is monotone, we constrain the spline parameters to be positive $\xi_{i,j} \ge 0$ for all $j = 1,\dots,n_\xi^i$. We regularized the spline by adding a penalty term to the objective function to promote linearity (Fig. 2B), which greatly improved the convergence of the estimation. For details on the definition of the spline, the distribution of the knot bases, and spline regularization, we refer to the first section of the supplementary materials.

The spline allows us to link the measured and biochemical quantities of the nonlinear monotone observable:

$$\tilde{z}_{ik} = s_i(h_i(x(t_k,\theta),\theta),\xi_i) + \varepsilon_{ik}, \tag{6}$$

$$\text{with } \varepsilon_{ik} \sim \mathcal{N}(0,\sigma_{ik}^2).$$

The model dataset $\mathcal{D} = \{\{(t_k,\tilde{z}_{ik})\}_{k=1}^{n_t}\}_{i=1}^{n_y}$ consists of observations of all model observables at time-points $\{t_k\}_{k=1}^{n_t}$. We denote the dataset of the $i$th observable as $\mathcal{D}_i = \{(t_k,\tilde{z}_{ik})\}_{k=1}^{n_t}$.

## 2.2 Parameter estimation

For a dataset $\mathcal{D}$ consisting of independent observations of quantitative, linear semi-quantitative, and/or nonlinear semi-quantitative observables, the negative log-likelihood objective function is commonly defined as:

$$J(\theta,\psi) = \sum_{i=1}^{n_y} J_i(\theta,\psi_i) = \sum_{i=1}^{n_y} -\log \mathcal{L}_{\mathcal{D}_i}(\theta,\psi_i) \tag{7}$$

in which $\psi_i$ are the observable parameters of the $i$th observable: for a relative observable these are scaling $a_i$ and offset $b_i$, while for a nonlinear semi-quantitative observable they are the spline parameters $\xi_i$. Minimizing the objective function, we obtain the maximum likelihood estimate (MLE): $\theta^*,\psi^* = \underset{\theta,\psi}{\operatorname{argmin}} J(\theta,\psi)$.

### 2.2.1 Hierarchical optimization problem and analytical gradients

The objective function minimization can be executed jointly in all mechanistic parameters $\theta$ and observable parameters $\psi$. However, this leads to a high-dimensional optimization problem and long computation times. Alternatively, the optimization problem can be separated hierarchically (Fig. 2A):

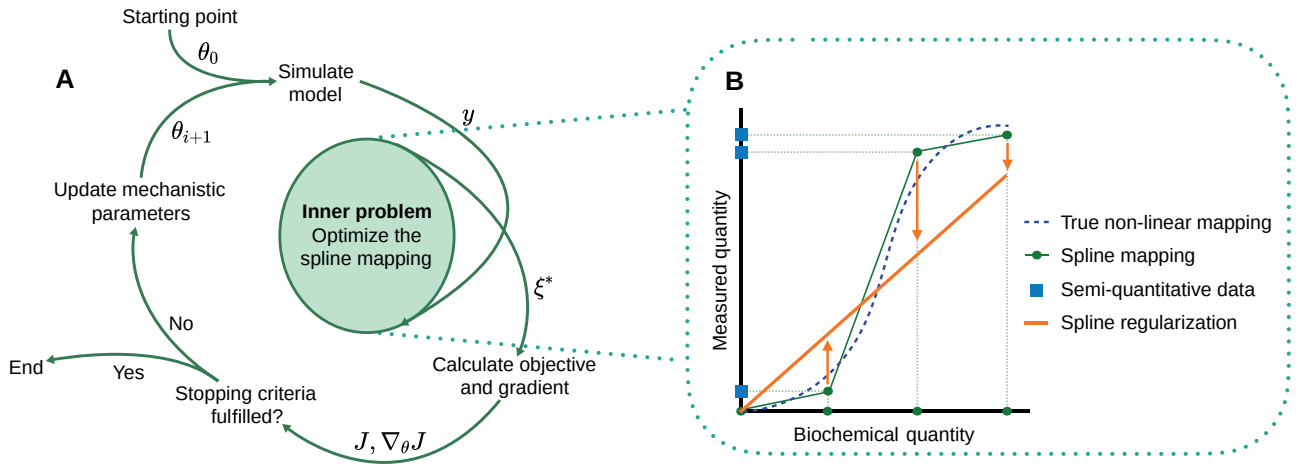


**Figure 2.** Illustration of the spline estimation approach. (A) Model mechanistic parameters $\theta$ are iteratively updated during parameter estimation. For each vector of trial parameters, the model is simulated to obtain the simulation $y$. Then, the spline parameters $\xi^*$ are optimized and used to calculate the objective function $J$ and its gradient $\nabla_\theta J$. These are then passed on to obtain the next trial parameter vector, or the optimization is halted. (B) The spline (green) enables mapping of the simulation of the model $y$ (biochemical quantities) to the measurement axis. This allows for the definition of a likelihood objective function. In the inner problem, this objective function is minimized with respect to the spline parameters to obtain optimal spline parameters $\xi^*$. The spline is additionally regularized by the distance to the linear mapping (orange)

$$\min_{\theta} \; J(\theta, \psi^*(\theta)) \tag{8}$$

$$s.t. \; \{\psi_i^*(\theta) = \underset{\psi_i}{\operatorname{argmin}} J_i(\theta, \psi_i) \text{ for } i = 1, \ldots, n_y. \tag{9}$$

In the outer optimization problem (8), we estimate the mechanistic parameters $\theta$, and in the inner optimization problems (9) we estimate the observable parameters $\psi_i$ of each observable. For nonlinear semi-quantitative observables, the inner problem is additionally constrained by the positivity of spline parameters. Since the objective function can be additively separated into components $\{J_i\}_{i=1}^{n_y}$ depending on the observable parameters of a single observable $\psi_i$, the inner optimization problem is a set of $n_y$ small inner optimization problems (9). For relative observables, the inner problems (9) can be solved analytically (Schmiester *et al.* 2019). For nonlinear semi-quantitative observables, they still need to be numerically minimized, but the inner problems are convex and thus easy to minimize. Second, the gradient of the objective function can be calculated analytically. We formulate and prove the two statements above in Theorems 1 and 2 of the supplementary materials.

### 2.3 Confidence region of a parameter vector $\theta$

Here, we define what it means for a parameter vector to lie in a confidence region of a certain significance. We do this using the likelihood-ratio test in which we define the corresponding test statistic as

$$\Lambda(\theta) = -2 \log \left( \frac{\mathcal{L}_{\mathcal{D}}(\theta)}{\sup_{\theta}\{\mathcal{L}_{\mathcal{D}}(\theta)\}} \right) = 2(J(\theta) - J(\theta^*)). \tag{10}$$

In the asymptotic case of a large number of data points, the test statistic converges to a chi-square distribution $\chi^2_{\text{df}}$ with $\text{df} = n_\theta$ degrees of freedom [see Tönsing *et al.* (2023) for more details]. Then we define the confidence region of significance $\alpha$ as

$$\text{CR}^\alpha = \{\theta | \Lambda(\theta) \leq \Delta_\alpha\} \tag{11}$$

in which $\Delta_\alpha$ is the $\alpha$th percentile of the $\chi^2_{n_\theta}$ distribution.

### 2.4 Scalability and complexity of the proposed method

The inner optimization objective functions $J_i(\theta, \psi_i)$ for semi-quantitative observables are convex and self-concordant. Thus, their numerical optimization using barrier methods scales mostly with the number of inequality constraints in the inner problem (Boyd and Vandenberghe 2004), i.e. with the number of spline parameters $\xi_i$ of the spline mapping (5). In most applications, this number should be set to a small value (5–10), as this already provides sufficient modeling capacity for most nonlinear measurement mappings and also reduces overfitting. Therefore, in larger ODE systems, the optimization of inner problems for semi-quantitative observables constitutes a small part of computational complexity and the proposed method scales linearly in the number of semi-quantitative observables.

### 2.5 Benchmark models

For the evaluation of the proposed method, we consider one exemplary model and four published models that were previously developed and calibrated for different biological systems (Table 1). As the published models originally did not contain nonlinear semi-quantitative observables, we generated synthetic data at the same time points, chose nonlinear monotone measurement mappings, applied them to the observables, and corrupted them with the same type of noise as in the original model. For details on the synthetic data generation, chosen nonlinear measurement mappings, and model structure, we refer to the second section of the supplementary material.

**Table 1.** Benchmark models.[a]

| Model | $n_x$ | $n_\theta$ | $n_y$ | $\lvert\mathcal{D}\rvert$ | Reference |
|---|---|---|---|---|---|
| T1 | 2 | 4 | 1 | 12 | Birtwistle *et al.* (2011) |
| M1 | 8 | 6 | 3 | 48 | Boehm *et al.* (2014) |
| M2 | 7 | 9 | 1 | 23 | Rahman et al. (2016) |
| M3 | 8 | 18 | 1 | 58 | Elowitz and Leibler (2000) |
| M4 | 14 | 18 | 8 | 205 | Raia *et al.* (2011) |

[a] By $|\mathcal{D}|$ we denote the cardinality of the dataset.

## 3 Results

### 3.1 The proposed method uncovers measurement mapping for FRET probe activation

To illustrate an application of the proposed method, we have applied it to a FRET probe activation model introduced by Birtwistle *et al.* (2011). In general, the transition of inactive FRET probes $P$ to an active state $P^*$ can be represented by the scheme in Fig. 3A. The quantity of interest in this model is the ratio of activated probes to total probes $P^*/P_{TOT}$. The most common way to measure this value is through a measurement technique called ratiometric imaging. Cells are exposed to excitation light from the donor channel, and then fluorescence emission is divided into donor and acceptor channels. The output of ratiometric imaging, R, is the intensity in the acceptor channel, $I_A$, divided by the intensity in the donor channel, $I_D$. Previous studies have shown that this measured R value can have a highly nonlinear relationship to the fraction of active FRET probes (Birtwistle *et al.* 2011) (Fig. 3A).

One approach of modeling this nonlinear mapping is to parameterize a function of a similar shape and to estimate its parameters. For FRET probe activation it has been shown that the relation between state variables and measurement is of the form:

$$g(P^*) = \alpha \cdot \frac{P^*}{P_{TOT} - P^*} + \beta \tag{12}$$

with experiment- and probe-specific parameters $\alpha$ and $\beta$. However, this requires prior knowledge of the shape of the measurement mapping. Without such prior knowledge, the measurement mapping has to be inferred. A simple and easy-to-implement approach is to assume that the mapping is linear. This linear approximation can be sufficiently correct if the measurement region is locally linear. For highly nonlinear measurement mappings, this is not true, so one has to resort to more flexible approaches such as spline estimation.

To evaluate how well the three modeling approaches can recover the true measurement mapping, we performed 1000 local optimizations for each and chose the best measurement mapping estimates in the 95% confidence region (Fig. 3B).

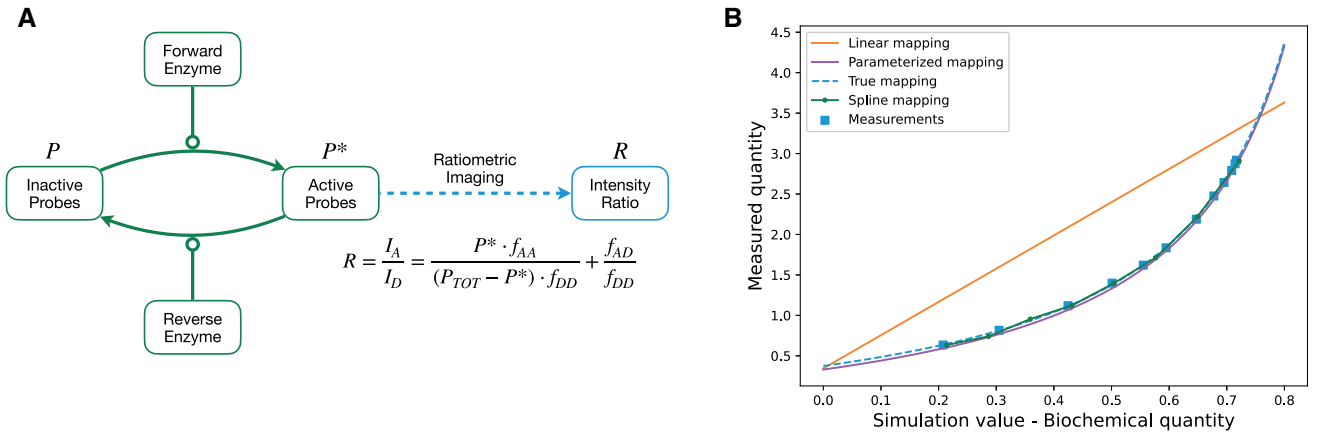


**Figure 3.** A model of FRET probe activation. (A) A forward enzyme catalyzes the activation of inactive FRET probes $P$, and a reverse enzyme catalyzes the conversion of an active probe $P^*$ into the inactive state. Active probe concentration can be observed via ratiometric imaging. The measurement mapping of this process is highly nonlinear. $f_{AA}$ and $f_{AD}$ are fractions of the acceptor and donor emissions that the acceptor channel captures, respectively. $f_{DD}$ is the fraction of donor emissions that the donor channel captures. (B) Comparison of the estimation of the measurement mapping using a linear function, proposed spline mapping, and parameterization of the true mapping. For all three models, we performed 1000 local optimizations. Depicted are the estimated mappings closest to the true mapping of starts with mechanistic parameters in the 95% confidence region. We show the synthetic noise-corrupted data used in all model optimizations in blue squares

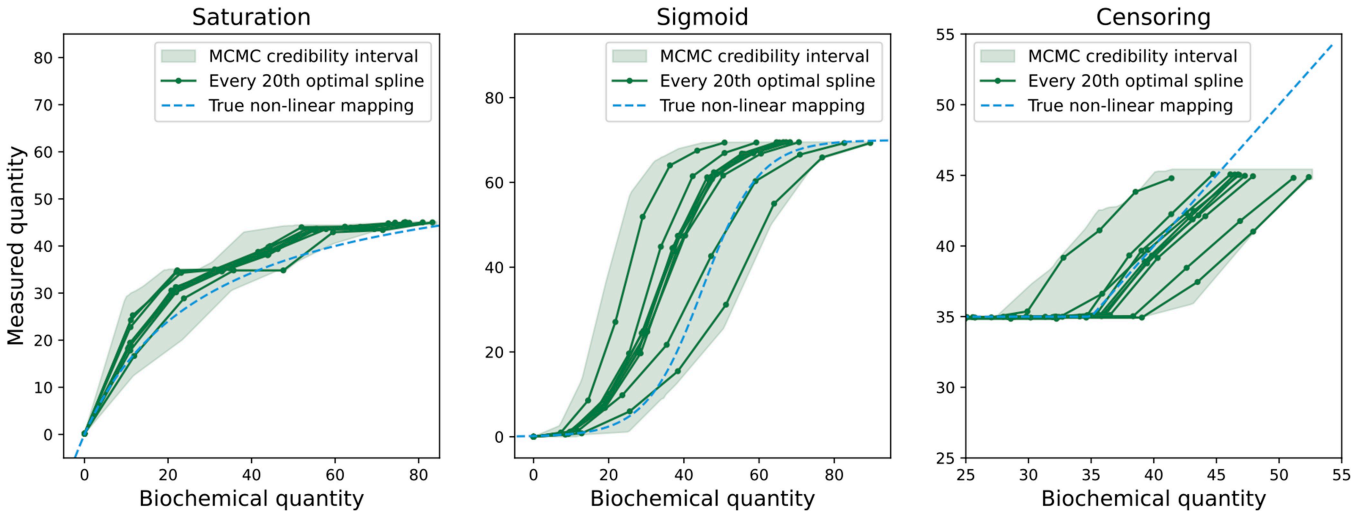


**Figure 4.** Credibility intervals of the estimated spline mappings. MCMC sampling of the model M1. The model contains different synthetic nonlinear measurement mappings for each of its observables (dashed blue). Splines were estimated for each 500th sample of the MCMC run with which we constructed the credibility intervals of the estimated spline mappings (light green). For visibility, we show only each 20th estimated spline (dark green)

We found that the reconstruction using spline estimation agrees well with the true measurement mapping. Indeed, it yields similar results to using parametric representations with unknown parameters. In contrast, a linear model for the measurement mapping proved to be insufficiently flexible and resulted in biased reconstruction of the measurement mapping.

Overall, our assessment revealed that, unlike a simple linear approximation for the measurement mapping, a spline-based approximation enables the reconstruction of nonlinear mappings as observed for FRET probe activation.

## 3.2 The spline estimation approach as a tool for uncovering measurement mapping shapes

In the previous subsection, we have shown that the estimation of an unknown measurement mapping using a spline can yield results similar to the estimation of a parametric representation. Yet, we only considered a point estimate and did not assess the reliability of the reconstruction. To determine whether the proposed approach provided statistically coherent estimates, we considered model M1 with measurement mappings of various shapes across observables (Fig. 4A–C). Using the resulting dataset, we performed a multi-start optimization ($10^3$ runs) to obtain optimal parameters and Markov chain Monte Carlo sampling using an adaptive Metropolis-Hastings algorithm ($10^5$ iterations). The resulting chain was thinned by a factor of 500. We computed the optimal spline for each of the remaining samples and, with them, constructed the credibility intervals of the optimal spline mappings.

The inspection of the results confirmed that the optimal splines are qualitatively similar to the measurement mappings used for data generation. Furthermore, more importantly, the measurement mappings used for data generation lie within the credibility intervals. This showcases the reliability of the method as a tool for discovering curve shapes of unknown measurement mappings.

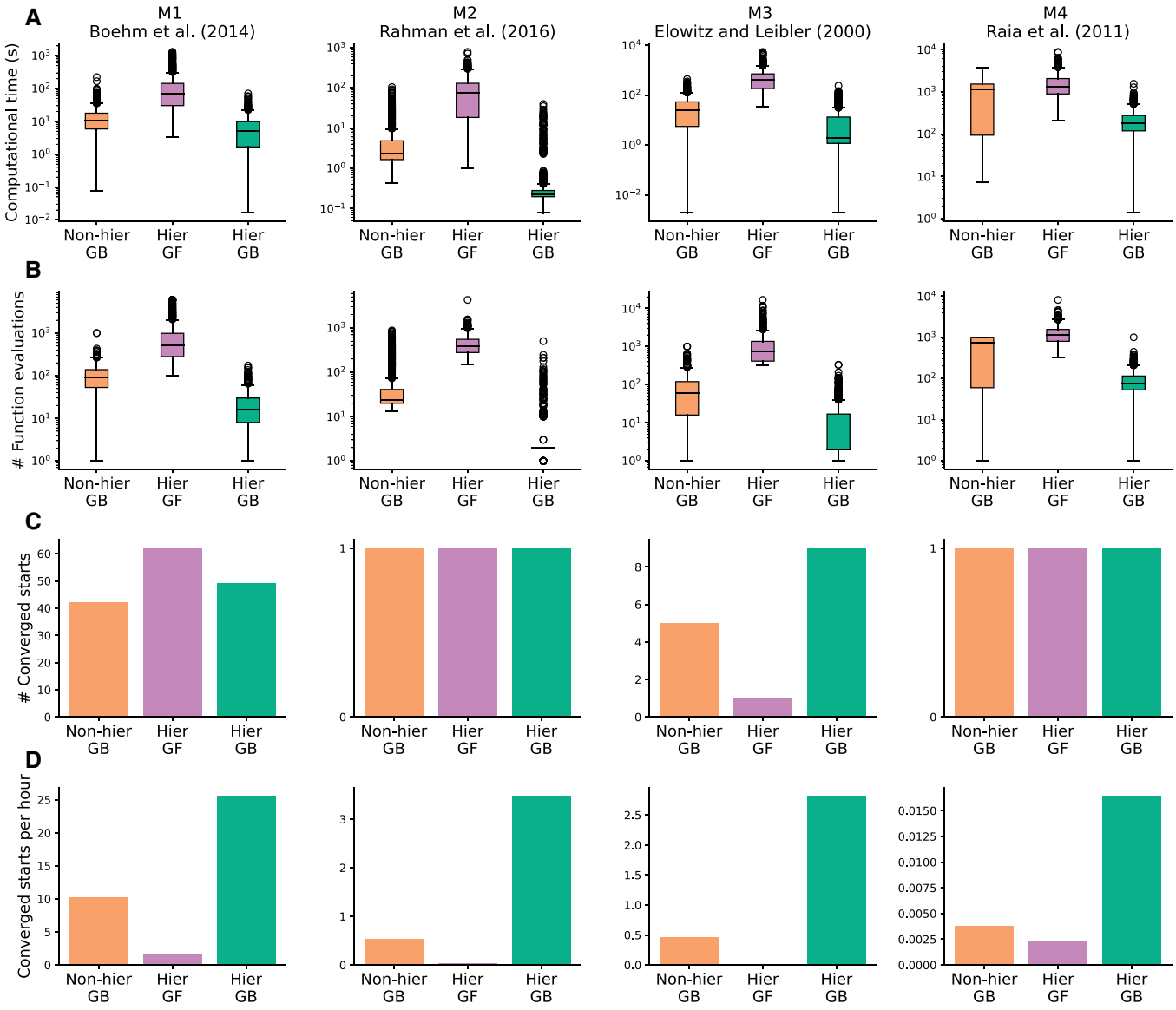


**Figure 5.** Evaluation of the gradient-based nonhierarchical, gradient-free hierarchical, and gradient-based hierarchical estimation approaches. Models M1–M4 are shown from left to right. (A) Comparisons of computation time. (B) Comparisons of the number of function evaluations. (C) Comparisons of the number of converged starts. Converged starts are defined as the starts with estimated mechanistic parameters within the 95% confidence region. (D) Comparison of the number of converged starts per CPU hour

### 3.3 Hierarchical optimization and analytical gradients increase the estimation efficiency

As the method provides reliable estimates for the mappings, we turn to the assessment of the computational cost, which is of high practical relevance. Here, we also wanted to evaluate the impact of (i) reformulation as a hierarchical problem and (ii) the availability of analytical gradients. For this assessment, we considered the published models M1 to M4 with synthetic data with a range of different measurement mappings, as detailed in the second section of the supplementary material. For all models, we performed 1000 local optimizations with equal start points across approaches. We then determined the overall computational cost, the number of function evaluations, and the number of converged starts per computation time.

We found that in general, the proposed hierarchical approach with analytical gradients achieves the best performance (Fig. 5). This appears to be mostly related to a reduction in the computation time, respectively, the number of function evaluations (Fig. 5A and B), while the number of converged starts remains rather similar (Fig. 5C). The number of converged starts per computation time is, for the hierarchical approach with analytical gradients, at least twice as high for the other approaches (Fig. 5D). Interestingly, a hierarchical approach without gradient information does not perform well, and is worse than the nonhierarchical approach with gradient information for all models.

For the proposed hierarchical approach with analytical gradients, the number of converged starts per CPU hour was on average $\approx 7.96$. As the spread between models was large, this finding clearly suggests that the approach is computationally tractable.

### 3.4 Spline approach improves the parameter inference of models with unknown measurement mappings

Our proposed method provides reliable estimates of measurement mappings. Here, we examine whether this leads to good

estimates of the mechanistic model parameters. Apart from spline estimation, a generally applicable approach to the integration of semi-quantitative data into parameter estimation is linear estimation of measurement mappings. Thus, we compare the parameter inference of these two approaches in a realistic setting. In addition, for reference, we include the approach of discarding data with unknown measurement mappings. We performed 1000 local optimizations for the application examples M1–M4. We evaluated the impact of an increasing number of unknown measurement mappings by turning quantitative observables into semi-quantitative observables. As a parameter inference metric, we use the mean L2 distance of the estimated to the true mechanistic parameters normalized by the number of mechanistic parameters. For details of the study setting, we refer to the fourth section of the supplementary materials.

The spline estimation outperforms other approaches (Fig. 6). In general, linear estimation has a stronger bias than variance. We observe this primarily for model M4, as the linear estimation has the smallest standard deviation between approaches (Fig. 6D). In some cases, this even causes the linear estimation to perform worse than the approach of discarding data with unknown mappings. In contrast, the higher flexibility of the spline estimation allows for the general attainment of better parameter estimates. This is the case even for model M4 with eight unknown measurement mappings, for which the spline estimation adds seven times more parameters than the linear estimation. For a small number of unknown measurement mappings, the spline estimation can perform almost equally well as the model with completely known measurement mappings. This showcases that the proposed method yields good estimates of the mechanistic model parameters, especially when the number of unknown measurement mappings is low.

## 4 Discussion

Semi-quantitative measurements represent a large portion of the available data that can be used to estimate unknown mechanistic parameters of ODE models. Among others, examples include spatial protein expression assays important in developmental biology, such as chemical staining, fluorescent expression, and immunohistochemistry (Brooks *et al.* 2012). When these are well-controlled, they are expected to linearly transform the true concentration into an image intensity. However, this is not always true: in the case of nonadequate procedural care, hard-to-control outer factors, or insufficient knowledge of the entire experimental system, the transformation function may not be

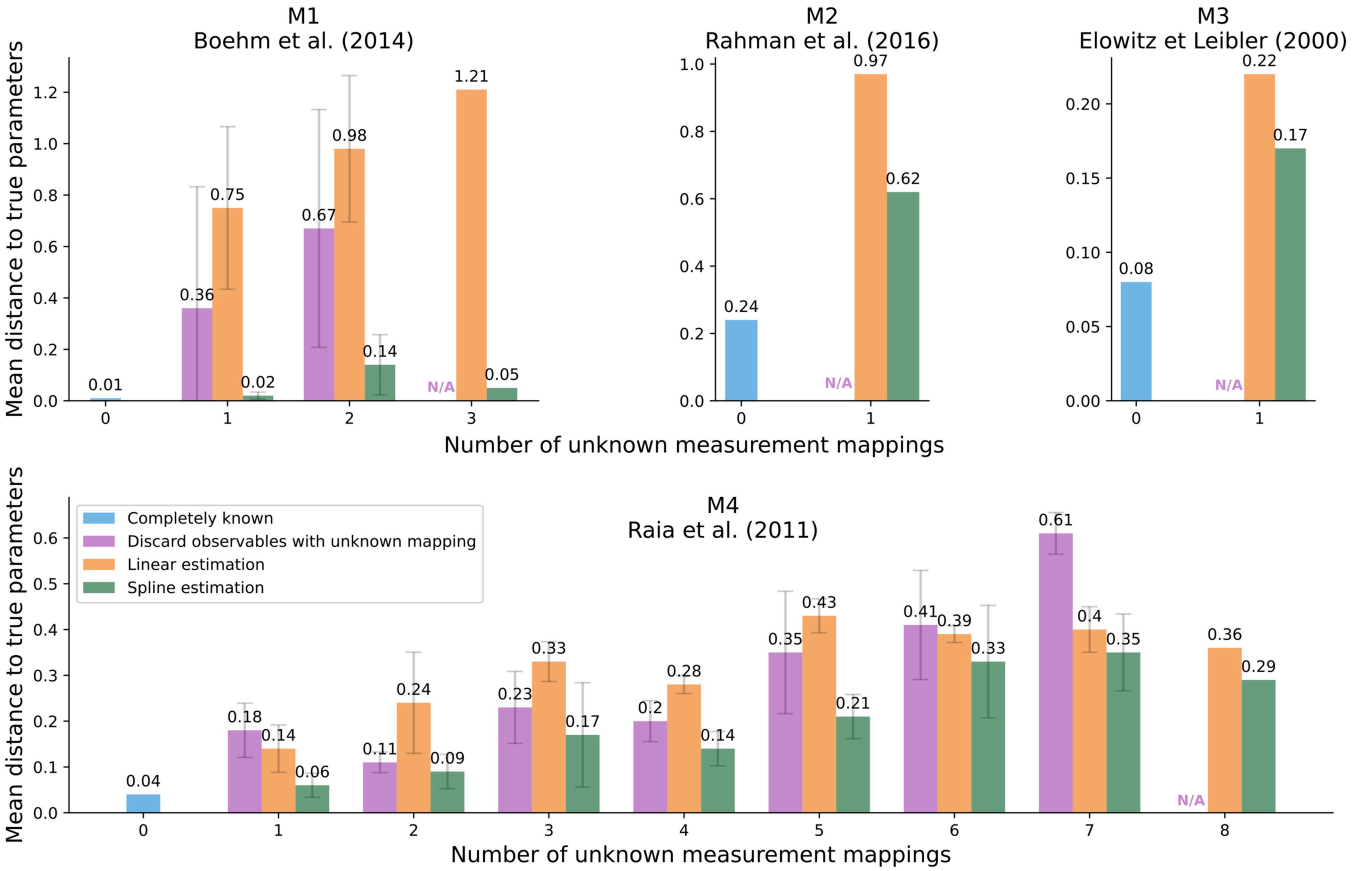


**Figure 6.** Evaluation of parameter inference across the number of unknown measurement mappings for linear and spline estimation. The parameter inference is measured by the L2 distance of the estimated to the true mechanistic model parameters. On the *x*-axis of each plot, we mark the model variant with a certain number of unknown measurement mappings, ranging from 0 to the number of model observables. The distance for each model variant is normalized by the number of mechanistic parameters and averaged across combinations of known-unknown observables. The best-case scenario for each model M1–M4 is the case of completely known measurement mappings (blue). We compare the distance to the true parameters for the linear (orange) and spline (green) approach for each number of unknown measurement mappings. The approach of discarding the data of observables with unknown measurement mappings (purple) is depicted for reference. This approach is not feasible for the model variants with a maximum number of unknown measurement mappings, as that would involve the removal of all data, so we denote this with N/A

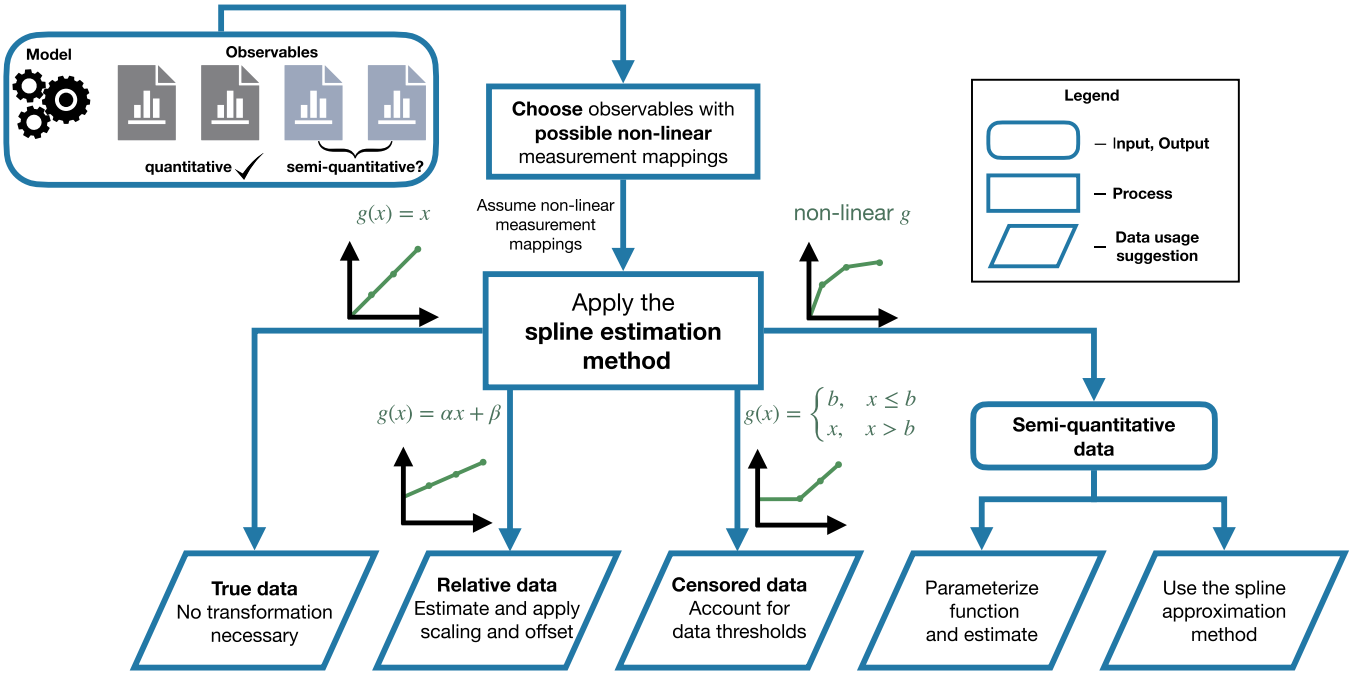


**Figure 7.** Diagram of application of the spline method to data with possible nonlinear measurement mappings. Estimated optimal splines are depicted in green. Each arrow from the central square is a possible outcome of the nonlinear mapping estimation

available and can take on a nonlinear form. Here, we address this challenge by introducing a spline-based method for the estimation of unknown nonlinear mappings. The approach can be applied to models with quantitative data for which it is unclear whether the data are truly linear (Fig. 7). Depending on the estimated optimal splines, the data can be deemed to be quantitative, relative, censored, or semi-quantitative, so that an appropriate method can be used. If the estimated optimal spline is nonlinear, one can choose to estimate a parameterized function of a similar shape, or continue using the optimal splines as the measurement mapping. In this way, the method allows for the integration of data previously usable only as qualitative. Furthermore, it can give a clearer understanding of the underlying experimental procedures.

However, one has to be conservative with the number of observables chosen for spline estimation, as it can lead to a large expansion of the parameter space dimension. We evaluated the reliability of this process using an example, showing consistent qualitative measurement mapping shapes. An obvious extension of this approach is the inclusion of symbolic function identification from the estimated optimal splines. This would constitute an automatic parameterization of the unknown measurement mappings.

To increase the method's efficiency, we employed a hierarchical gradient-based optimization approach. We evaluated its performance and compared it with alternative approaches for four published models with differences in their complexity. This revealed a higher computational efficiency across all models, allowing for faster estimation of parameters for a given model. Further optimization acceleration could be achieved by including adjoint sensitivity analysis (ASA) (Kokotovic and Heller 1967, Fröhlich *et al.* 2017). Although our inner problem is not solved exactly, in Theorem 2 of the supplementary, we show that its gradient contribution is still zero. Thus, existing ASA software implementations from Schmiester et al. (2019) can be used, since the gradient computation is the same as in hierarchical optimization with an exactly solved inner problem. Complementary to this, the derivation of second-order derivatives could further improve the method's convergence and, with it, its computational efficiency.

The proposed method employs piecewise linear splines to estimate general nonlinear mappings. This was the simplest first-pass option, but, as they are not smooth, they had unavoidable approximation errors. Therefore, it is valuable to explore alternative smooth and flexible parameterized functions. Furthermore, they should retain the convexity of the inner optimization problems and the possibility of analytical gradient calculation. Interesting candidates are the scaled cumulative distribution functions (CDFs) of the beta distribution. They are monotone by definition, parameterized by only three parameters, and with promising flexibility to be able to model most types of measurement nonlinear mappings.

The models for which the method was developed are based on ODE systems primarily because of their widespread prevalence. However, the method can be used more generally. It requires only the model simulations, sensitivities, and the definition of the objective function as a negative log-likelihood. Thus, any model that satisfies these constraints can be incorporated to integrate semi-quantitative data into its estimation of parameters.

In conclusion, we developed and implemented an easy-to-use, computationally efficient framework to uncover unknown nonlinear measurement mappings and to integrate semi-quantitative data into the parameter estimation of ODE models. The approach has a user-friendly implementation in the open-source Python Parameter Estimation TOolbox (pyPESTO). As it is agnostic to the structure of the underlying dynamical model, the method can be applied to models from different research fields, such as physics and engineering.

## Author contributions

J.H. conceived the project. D.D. implemented the proposed approach and conducted all studies. S.G. and J.H. provided

critical feedback on the implementation development. D.D. and J.H. wrote the manuscript. All authors discussed the results and commented on the manuscript.

## Supplementary data

Supplementary data are available at *Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy [EXC 2047—390685813, EXC 2151—390873048] and under the project IDs 432325352—SFB 1454 and 443187771—AMICI, by the German Federal Ministry of Education and Research (BMBF) within the e:Med funding scheme [junior research alliance PeriNAA, 01ZX1916A] and under the CompLS program [EMUNE, 031L0293C], and by the University of Bonn (via the Schlegel Professorship of J.H.).

## Data availability

The proposed method is implemented in the open-source Python Parameter Estimation TOolbox (pyPESTO) (schalte2023pypesto). Models M1-M4 were taken from the PEtab benchmark collection (Schmiester et al. 2021a) based on Hass *et al.* (2019). For ODE integration, we used the AMICI Python toolbox (Fröhlich *et al.* 2021). Gradient-based optimization was performed using the fides optimizer (Fröhlich and Sorger 2022) and gradient-free optimization was performed with the SciPy Powell algorithm (Jones *et al.* 2001). Both optimizers were used through the pyPESTO interface with the default optimizer settings. All the code and models used in this study are available from the Zenodo database at https://doi.org/10.5281/zenodo.10568951.

## References

Birtwistle MR, von Kriegsheim A, Kida K *et al.* Linear approaches to intramolecular Förster resonance energy transfer probe measurements for quantitative modeling. *PLoS One* 2011;**6**:e27823.

Boehm ME, Adlung L, Schilling M *et al.* Identification of isoform-specific dynamics in phosphorylation-dependent stat5 dimerization by quantitative mass spectrometry and mathematical modeling. *J Proteome Res* 2014;**13**:5685–94.

Boyd S, Vandenberghe L. *Convex Optimisation*. UK: Cambridge University Press, 2004.

Brooks A, Dou W, Yang X *et al.* BMP signaling in wing development: a critical perspective on quantitative image analysis. *FEBS Lett* 2012;**586**:1942–52.

Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature* 2000;**403**:335–8.

Fey D, Halasz M, Dreidax D *et al.* Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci Signal* 2015;**8**:ra130. https://doi.org/10.1126/scisignal.aab0990

Fröhlich F, Sorger PK. Fides: reliable trust-region optimization for parameter estimation of ordinary differential equation models. *PLoS Comput Biol* 2022;**18**:e1010322. https://doi.org/10.1371/journal.pcbi.1010322

Fröhlich F, Weindl D, Schälte Y *et al.* AMICI: high-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics* 2021;**37**:3676–7. https://doi.org/10.1093/bioinformatics/btab227

Fröhlich F, Kaltenbacher B, Theis FJ *et al.* Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput Biol* 2017;**13**:e1005331. https://doi.org/10.1371/journal.pcbi.1005331

Giordano G, Blanchini F, Bruno R *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med* 2020;**26**:855–60. https://doi.org/10.1038/s41591-020-0883-7

Hass H, Masson K, Wohlgemuth S *et al.* Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ Syst Biol Appl* 2017;**3**:27. https://doi.org/10.1038/s41540-017-0030-3

Hass H, Loos C, Raimúndez-Álvarez E *et al.* Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics* 2019;**35**:3073–82. https://doi.org/10.1093/bioinformatics/btz020

Jones E, Oliphant T, Peterson P *et al. SciPy: Open Source Scientific Tools for Python*, 2001. http://www.scipy.org/

Kitano H. Systems biology: a brief overview. *Science* 2002;**295**:1662–4.

Kokotovic P, Heller J. Direct and adjoint sensitivity equations for parameter optimization. *IEEE Trans Automat Contr* 1967;**12**:609–10. https://doi.org/10.1109/tac.1967.1098670

Kreutz C, Bartolome Rodriguez MM, Maiwald T *et al.* An error model for protein quantification. *Bioinformatics* 2007;**23**:2747–53. https://doi.org/10.1093/bioinformatics/btm397

Loos C, Krause S, Hasenauer J. Hierarchical optimization for the efficient parametrization of ODE models. *Bioinformatics* 2018;**34**:4266–73. https://doi.org/10.1093/bioinformatics/bty514

Mitra ED, Hlavacek WS. Parameter estimation and uncertainty quantification for systems biology models. *Curr Opin Syst Biol* 2019;**18**:9–18.

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

Rahman SMA, Vaidya NK, Zou X. Impact of early treatment programs on HIV epidemics: an immunity-based mathematical model. *Math Biosci* 2016;**280**:38–49. https://doi.org/10.1016/j.mbs.2016.07.009

Raia V, Schilling M, Böhm M *et al.* Dynamic mathematical modeling of il13-induced signaling in hodgkin and primary mediastinal b-cell lymphoma allows prediction of therapeutic targets. *Cancer Res* 2011;**71**:693–704.

Schmiester L, Schälte Y, Fröhlich F *et al.* Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics* 2019;**36**:594–602. https://doi.org/10.1093/bioinformatics/btz581

Schmiester L, Weindl D, Hasenauer J. Parameterization of mechanistic models from qualitative data using an efficient optimal scaling approach. *J Math Biol* 2020;**81**:603–23. https://doi.org/10.1007/s00285-020-01522-w

Schmiester L, Schälte Y, Bergmann FT *et al.* PEtab—interoperable specification of parameter estimation problems in systems biology. *PLoS Comput Biol* 2021a;**17**:e1008646. https://doi.org/10.1371/journal.pcbi.1008646

Schmiester L, Weindl D, Hasenauer J. Efficient gradient-based parameter estimation for dynamic models using qualitative data. *Bioinformatics* 2021b;**37**:4493–500. https://doi.org/10.1093/bioinformatics/btab512

Schöberl B, Pace EA, Fitzgerald JB *et al.* Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Sci Signal* 2009;**2**:ra31.

Shepard RN. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* 1962;**27**:125–40.

Stevenson K, McVey AF, Clark IBN *et al.* General calibration of microbial growth in microplate readers. *Sci Rep* 2016;**6**:38828. https://doi.org/10.1038/srep38828

Tönsing C, Steiert B, Timmer J *et al.* Likelihood-ratio test statistic for the finite-sample case in nonlinear ordinary differential equation models. *PLoS Comput Biol* 2023;**19**:e1011417. https://doi.org/10.1371/journal.pcbi.1011417

Zhao S, Chen H. Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant Biol* 2020;**8**:11–9. https://doi.org/10.1007/s40484-020-0199-0

Bioinformatics, 2024, 40, 558–566
https://doi.org/10.1093/bioinformatics/btae210
ISMB 2024