BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems

Attila Gábor[1,2†], Alejandro F. Villaverde[1†] and Julio R. Banga[1*]

Abstract

Background: Kinetic models of biochemical systems usually consist of ordinary differential equations that have many unknown parameters. Some of these parameters are often practically unidentifiable, that is, their values cannot be uniquely determined from the available data. Possible causes are lack of influence on the measured outputs, interdependence among parameters, and poor data quality. Uncorrelated parameters can be seen as the key tuning knobs of a predictive model. Therefore, before attempting to perform parameter estimation (model calibration) it is important to characterize the subset(s) of identifiable parameters and their interplay. Once this is achieved, it is still necessary to perform parameter estimation, which poses additional challenges.

Methods: We present a methodology that (i) detects high-order relationships among parameters, and (ii) visualizes the results to facilitate further analysis. We use a collinearity index to quantify the correlation between parameters in a group in a computationally efficient way. Then we apply integer optimization to find the largest groups of uncorrelated parameters. We also use the collinearity index to identify small groups of highly correlated parameters. The results files can be visualized using Cytoscape, showing the identifiable and non-identifiable groups of parameters together with the model structure in the same graph.

Results: Our contributions alleviate the difficulties that appear at different stages of the identifiability analysis and parameter estimation process. We show how to combine global optimization and regularization techniques for calibrating medium and large scale biological models with moderate computation times. Then we evaluate the practical identifiability of the estimated parameters using the proposed methodology. The identifiability analysis techniques are implemented as a MATLAB toolbox called VisId, which is freely available as open source from GitHub (https://github.com/gabora/visid).

Conclusions: Our approach is geared towards scalability. It enables the practical identifiability analysis of dynamic models of large size, and accelerates their calibration. The visualization tool allows modellers to detect parts that are problematic and need refinement or reformulation, and provides experimentalists with information that can be helpful in the design of new experiments.

Keywords: Parameter estimation, Dynamic models, Identifiability, Global optimization, Regularization, Overfitting

*Correspondence: julio@iim.csic.es
†Equal contributors
[1]BioProcess Engineering Group, IIM-CSIC, Eduardo Cabello 6, 36208 Vigo, Spain
Full list of author information is available at the end of the article

Background

The development of mechanistic (kinetic) models in order to quantitatively describe the dynamics of biological phenomena is one of the core research themes in systems biology. During the last decade, fostered by the greater availability of the necessary experimental data, the development of large (up to genome-scale) kinetic models has become one of the main objectives in the field, as well as in related areas such as synthetic biology, metabolic engineering, or industrial biotechnology [1–10]. More recently, the first steps towards comprehensive whole-cell models have been taken [11], which has great potential for applications e.g. in personalized medicine [12]. However, the development of these large-scale integrated dynamic models poses severe challenges [13, 14]. Those associated with model building are common to the more general problem of reverse engineering of biological systems [15]. In this context, parameter estimation (i.e. model calibration) is arguably one of the most studied [16–19], yet more challenging step in model building.

Parameter estimation in nonlinear dynamic models can be an extremely hard problem mostly due to the following issues [15]: lack of identifiability, ill-conditioning, multimodality and over-fitting. The latter three can be handled via global optimization and regularization methods, as reviewed and illustrated recently [20]. The present paper begins by continuing the line of work in [20], addressing these three issues. To this end we introduce a combination of a global optimization metaheuristic, eSS [21], and an efficient local search method, the adaptive algorithm NL2SOL [22]. By using this optimization technique jointly with regularization it is possible to *reduce the calibration times* of large dynamic models and simultaneously avoid over-fitting. We show this for models from the recently presented BioPreDyn benchmark collection [23]. Then we focus on the remaining issue, that is, *identifiability analysis* of large dynamic models. Our aim is to develop a methodology which (i) is able to characterize high-order relationships among parameters, and (ii) scales up well with model size. Thus, our objective goes beyond finding the subset of identifiable parameters: we also aim to systematically characterize the space of non-identifiable parameters, and to facilitate the advanced analysis of the results with scalable visualization tools.

Identifiability analysis aims at establishing whether it is possible to determine the values of the unknown model parameters [24]. It is common to distinguish between structural and practical identifiability. *Structural* or a priori identifiability analysis decides whether the model parameters are uniquely determinable based on the model formulation, which includes the dynamic equations, observation functions and stimuli [25]. A parameter θ of the model is structurally identifiable if $y(\theta) = y(\theta') \Leftrightarrow \theta = \theta'$, where y denotes the model predictions, which are observable in the experiments. A parameter θ is structurally *locally* identifiable if for almost any value θ^* there is a neighbourhood $V(\theta^*)$ in which the above relationship holds. It is *globally* identifiable if the relationship holds in all the range of values of the parameter. If there is some region with non-zero measure where the relationship does not hold, θ is structurally unidentifiable. Structural identifiability analyses usually involve a high computational burden, which makes them difficult to apply to large models [26–28]. Furthermore, structural identifiability is only a necessary but not sufficient condition for identifiability. Very often a structurally identifiable parameter is practically unidentifiable, that is, its value cannot be determined with precision due to limitations in the available data. This can be quantified using *practical* or *a posteriori* identifiability analysis, which provides confidence intervals of the parameter values. The two main sources of practical non-identifiability are (1) lack of influence of a parameter on the observables, and (2) interdependence among the parameters. Obviously, if a parameter does not influence the observables (case 1) it is not possible to determine its value. The second situation, in which the effect on the observables of a change in one parameter can be compensated by changes in other parameters, can also prevent parameter identification. Both problems are related to the sensitivities of the observables to changes in model parameters. While (1) is related to the average sensitivity of the model outputs to a specific parameter, (2) can be investigated based on the collinearity of the parametric sensitivities [29].

In this paper we combine global optimization and regularization techniques to calibrate medium and large scale biological models (in this context, we will use the term “medium-scale” for models with 10 to 50 parameters and “large-scale” for models with more than 50 parameters). Then we evaluate the practical identifiability of estimated model parameters using sensitivity analysis and collinearity measures. We determine the largest identifiable subsets of parameters, characterize the interplay among non-identifiable groups of parameters, and visualize the results using Cytoscape. The visualization tool shows the identifiable and non-identifiable groups of parameters together with the model structure in the same graph. In this way, modellers can detect parts that are problematic and need refinement or reformulation, and experimentalists obtain information that can be helpful in the design of new experiments. The methods for identifiability analysis and visualization presented here have been implemented as a MATLAB toolbox called VisId, which is available from GitHub (https://github.com/gabora/visid) and as Additional file 1.

Methods

Parameter estimation with regularization and global optimization

Mathematical model

We consider deterministic models of biological systems that can be described by nonlinear ordinary differential equations (ODEs) in the following form:

$$\frac{dx(t,\theta)}{dt} = f(x(t,\theta), u(t), \theta), \tag{1}$$

$$y(x,\theta) = g(x(t,\theta),\theta), \tag{2}$$

$$x(t_0) = x_0(\theta), \quad t \in \left[t_0,\, t_f\right]. \tag{3}$$

Here $x \in \mathbb{R}^{N_x}$ denotes the state vector (often concentrations), f describes the interactions among the state variables (often constructed from the reaction rate functions), and $u(t)$ denotes the input variables (stimuli). The parameter vector $\theta \in \mathbb{R}^{N_\theta}$ contains the (positive) parameters, e.g. reaction rate coefficients or Hill exponents. Their values are often unknown and must be estimated from data.

The model variables x are mapped to the measurable output variables $y \in \mathbb{R}^{N_y}$, also known as observables or model predictions, by the observation function g. These y signals are the quantities that can be experimentally measured. We will denote by y_{ijk} the model prediction for the j-th observed quantity in the k-th experiment at time $t_i \in [\, t_0,\, t_f]$. The corresponding measured data is denoted by $\tilde{y}_{ijk}$.

Parameter estimation

The goal of parameter estimation is to determine the values of the unknown parameter vector θ. This is usually done by minimizing a distance between model prediction y_{ijk} and measured data $\tilde{y}_{ijk}$. One of the simplest, but yet general, choices of this distance is the weighted sum-of-squares

$$Q_{\mathrm{LS}}(\theta) = \sum_{k=1}^{N_\mathrm{e}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} w_{ijk} \left(y_{ijk}\left(x(t_i,\theta),\theta\right) - \tilde{y}_{ijk}\right)^2, \tag{4}$$

where N_e is the number of experiments, $N_{y,k}$ is the number of observed compounds in the k-th experiment, and $N_{t,k,j}$ is the number of measurement time points of the j-th observed quantity in the k-th experiment, and the weights are denoted by w_{ijk}. The total number of data in all experiments is denoted by $N_D = \sum_{k=1}^{N_\mathrm{e}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} 1$. In order to simplify the index triplet, from now on we will use only one index, i.e. the weights and observables are denoted by w_i and y_i for $i = 1,\ 2 \ldots N_D$.

Then the parameter estimation problem is formulated as an optimization problem in the following form:

$$\underset{\theta}{\text{minimize}}\ Q_{\mathrm{LS}}(\theta) + \alpha \Gamma(\theta) \tag{5}$$

$$\text{subject to } \theta_{\min} \le \theta \le \theta_{\max}, \tag{6}$$

$$\frac{dx(t,\theta)}{dt} = f(u(t), x(t,\theta), \theta), \tag{7}$$

$$y(x,\theta) = g(x(t,\theta),\theta), \tag{8}$$

$$x(t_0) = x_0(\theta), \quad t \in \left[t_0,\, t_f\right]. \tag{9}$$

Here $\Gamma(\theta)$ is a a regularization term, which is described in the following subsection, and $\theta_{\min}$ and $\theta_{\max}$ are lower and upper bounds of the parameter values. The parameter vector $\hat{\theta}$ that solves this minimization problem is called the *optimal parameter vector* or the parameter estimates.

Regularization

Large scale dynamic models are often over-parametrized, turning the estimation of their parameters into an ill-posed problem [30]. This means that the minimum of the least-squares cost function (4) is non-unique, or that even a very small perturbation of the data results in very different estimated parameters. Furthermore, due to the large number of degrees of freedom, these models tend to capture the artificial dynamics of measurement noise. This is known as overfitting [31, 32] and it usually results in poor predictive capability of the calibrated model.

Regularization techniques incorporate a priori knowledge about the parameter values to make the problem well-posed. The regularization parameter α in (5) balances the strength of this knowledge; its value can be found by regularization tuning methods [33]. Here we followed the guidelines presented in [20] and chose a small regularization parameter ($\alpha = 0.1$), since we assume that we do not have good a priori estimates of the parameters.

Regarding the regularization function, $\Gamma(\theta)$, we chose the Tikhonov regularization framework to match the form of the penalty to the least squares formalism of the objective function. In this case the penalty is a quadratic penalty function,

$$\Gamma(\theta) = \left(\theta - \theta^{\mathrm{ref}}\right)^T W^T W \left(\theta - \theta^{\mathrm{ref}}\right), \tag{10}$$

where $W \in \mathbb{R}^{N_\theta \times N_\theta}$ is a diagonal scaling matrix and $\theta^{\mathrm{ref}} \in \mathbb{R}^{N_\theta}$ is a reference parameter vector, which is problem dependent and determined by the available information about the model parameters.

Global optimization

We solve the minimization problem defined by (5)–(10) using optimization. Since the cost function (5) is usually multi-modal (i.e. it usually has several local minima) [34–37], it is necessary to use an efficient global optimization method. Deterministic global optimization methods [38–42] can guarantee global optimality of the solution. However, their computational cost increases exponentially with the number of parameters, which makes them unsatisfactory for large scale models. Stochastic and metaheuristic methods [17, 18, 35, 36, 43, 44], on the other hand, do not provide such guarantees, but are often capable of finding adequate solutions in reasonable computation times.

For this reason we use a method called enhanced scatter search (eSS) [21], which is an advanced implementation of a population-based algorithm called scatter search. The scatter search metaheuristic works by evolving a number of solutions (population members), which constitute the reference set (RefSet). Members of this set are selected due to their quality and diversity. They are updated at every iteration by combining them with other RefSet members and, occasionally, by applying an improvement method. This improvement consists of a local search to speed-up the convergence to optimal solutions. In the present work we have chosen NL2SOL [22] as a local method. NL2SOL is a quasi-Newton algorithm with trust region strategy that exploits the structure of the nonlinear least squares problem. Note that the combination of a global method (scatter search) with a local one makes eSS a hybrid algorithm.

Practical identifiability analysis

The shape of the cost function (5) in the surroundings of its optima determines the local identifiability of the parameters. We assess parametric identifiability in two consecutive steps:

1. First we calculate the sensitivity of the model outputs (observables) with respect to changes in the parameters. Those parameters which have no effect (or very little) on the observed signals are classified as non-identifiable. Note that this label is assigned on an individual basis, that is, taking only into account the effect of each parameter individually.
2. Even if a parameter influences the model output, it may still be unidentifiable if its effect can be compensated by changes in other(s) parameter(s). Hence in the second step we consider the interplay among parameters, aiming at finding groups of parameters which are non-identifiable due to their collinearity.

Note that, while it would be possible at least in principle to perform both steps simultaneously, in practice the curse of dimensionality hampers the application of such a global sensitivity approach to large models [45, 46].

Sensitivity analysis

The analysis of parametric sensitivity of kinetic models has a long tradition in model analysis [47, 48]. For the dynamical system (1)–(2), the parametric sensitivities of the observables can be accurately calculated by solving the forward sensitivity equations:

$$\frac{dX_i(t)}{dt} = \frac{\partial f(x,u,\theta)}{\partial x} X_i(t) + \frac{\partial f(x,u,\theta)}{\partial \theta} \quad \text{for } i = 1,\dots,N_\theta \tag{11}$$

$$s_i(t) = \frac{\partial g(x,\theta)}{\partial x} X_i(t) + \frac{\partial g(x,\theta)}{\partial \theta} \quad \text{for } i = 1,\dots,N_\theta \tag{12}$$

$$s_i(t_0) = \begin{cases} 0 & \text{if } \theta_i \text{ is a model parameter} \\ 1 & \text{if } \theta_i \text{ is an initial condition} \end{cases} \quad \text{for } i = 1,\dots,N_\theta. \tag{13}$$

Here $X_i = \frac{\partial x}{\partial \theta_i}$ denotes the *sensitivity of the state vector* with respect to the i-th parameter and the vector $s_i = \frac{\partial y}{\partial \theta_i}$ is the *sensitivity of the observables* with respect to this parameter. This calculation requires the solution of the $N_x \times N_\theta$ ordinary differential Eq. (11) with initial conditions (13) for each experiment. The numerical solution is determined for the time points for which there are experimental data available, and then the algebraic Eq. (12) are evaluated. If the partial derivatives of the dynamic equations are not available, an alternative is to calculate the sensitivities using finite differences or automatic differentiation.

The sensitivities of the observables are scaled using the same weights as in Eq. (4), resulting in scaled sensitivities for an output *j* and a parameter *i*:

$$\left[\tilde{s}_i\right]_j = \sqrt{w_j}\frac{\partial y_j}{\partial \theta_i}. \tag{14}$$

For each parameter we calculate an overall scoring called root mean squared sensitivity, $\tilde{s}_i^{\text{msqr}}$, to take into account changes in time or across experiments [29, 49]:

$$\tilde{s}_i^{\text{msqr}} = \sqrt{\frac{1}{N_D}\sum_{j=1}^{N_D} \tilde{s}_{ij}^2} \ \text{for } \ i = 1,\dots,N_\theta \ . \tag{15}$$

Below a certain threshold the parameters are considered non-influential to the outputs. We set the threshold to four orders of magnitude smaller than the maximum root mean square value (15). Parameters whose sensitivity falls below this cut-off value are considered practically non-identifiable and they are kept out of further analysis. The procedure is summarized in Algorithm 1.

We remark that the outcome of the sensitivity calculations depends not only on the parameters, but also on the choice of initial conditions and external stimuli, which can have a strong influence in the practical identifiability of a

Algorithm 1 Finding sensitive model parameters

```
Require: Obtain vector of calibrated parameters → $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_{N_\theta}]$ (solve Eqs. (5)–(10))
    Parameter index set $\mathcal{I} \leftarrow \{1, 2, \ldots, N_\theta\}$
    Compute the sensitivity matrix at the optimal parameter vector → $s(\hat{\theta})$ (solve Eqs. (11)–(13))
    Compute the weighted sensitivities → $\tilde{s}$ (solve Eq. (14))
    Find the sensitive parameters by ranking the mean-square values of the sensitivity columns as in Eq. (15) and setting a cut-off value. The corresponding index set → $\mathcal{I}_{\text{sensitive}} \subset \mathcal{I}$
```

model. If insufficiently excitatory stimuli or initial conditions result in poor practical identifiability, a solution – if it is possible to carry out additional measurements – is to design and perform a new experiment to generate maximally informative data [17].

Collinearity of parameters

Interplay among influential parameters can result in an unidentifiable model, because a variation in the cost function value due to a change in a parameter can be compensated by changes in other parameters. Pairwise interplay can be detected by plotting contours of the cost function versus pairs of parameters. Largely eccentric contours or "valleys" show that the cost function is almost unchanged in one direction, and the two parameters are highly correlated. This approach has two drawbacks: it involves a large computational effort and is limited to interplay between pairs of parameters. To compute higher dimensional interactions we use a different measure: the *collinearity* of parametric sensitivities.

To calculate collinearity we first normalize the scaled sensitivities (14) as follows:

$$\bar{s}_i = \frac{\hat{s}_i}{\|\hat{s}_i\|} \quad \text{for } i = 1, \ldots, N_\theta. \tag{16}$$

This normalization avoids biases caused by differences in the absolute values of the individual sensitivity vectors.

Let us consider a set K of k parameters and their corresponding sensitivity vectors. The parameters are linearly dependent if there exist k constants $\alpha_i \neq 0$ such that

$$\alpha_1 \bar{s}_{K_1} + \alpha_2 \bar{s}_{K_2} + \ldots \alpha_k \bar{s}_{K_k} = 0 \tag{17}$$

If the above relation does not hold, the set is independent. When the equality (17) holds only approximately, the parameters are nearly dependent or nearly collinear. The degree of collinearity among a set of parameters can be measured by the collinearity index, CI_K, which is defined as [29]:

$$\mathrm{CI}_K = \frac{1}{\min_{\|\alpha\|=1} \|\bar{S}_K \alpha\|} = \frac{1}{\sqrt{\lambda_{K,\min}}}. \tag{18}$$

where $\bar{S}_k$ is the sensitivity matrix built from the k sensitivity vectors, $\bar{S}_K =[\bar{s}_{K_1}, \bar{s}_{K_2} \ldots \bar{s}_{K_k}]$, and $\lambda_{K,\min}$ is the smallest eigenvalue of $\bar{S}_K^T \bar{S}_K$. The larger the collinearity index is, the more dependent the corresponding parameters are. Brun and co-authors [29] proposed to classify a subset of parameters as identifiable if their collinearity index is smaller than a threshold which they chose as $\mathrm{CI}_K < 20$. Roughly speaking, a value of 20 means that 95% of the variation in the model output caused by changing one of the parameters in the subset can be compensated by changing the other parameters in the set.

Other approaches for finding parameter correlations using sensitivity-based measures have been previously proposed in the literature. Li and Vu presented two methods [50, 51] that search for relationships among parameters in the context of a priori identifiability analysis (i.e. with noise-free, continuous data). The method in [50] provides a necessary but not sufficient condition for identifiability of nonlinear systems, which need to be fully observed (i.e. they must satisfy $y = x$). The method in [51] removes the requirement of measuring all the system states, replacing it with the restriction that the model must be linear. We remark that the method proposed in the present manuscript does not have these limitations: it can be applied to partially observed, nonlinear systems with noisy, discrete-time measurements.

Largest identifiable subset

As explained in the previous subsections, a subset of parameters is considered identifiable if its elements are influential and their sensitivity vectors are not collinear. We are interested in finding the largest set of parameters for which the collinearity of the corresponding sensitivity vectors is below the chosen threshold, $\mathrm{CI}_K < 20$. Such a set of parameters represents all the degrees of freedom in the model. This means that perturbing a parameter *not* included in this set has an effect in the model predictions that can be compensated (at least by 95%) by changing other parameters in the set. However, a perturbation in a parameter belonging to the set cannot be compensated by changes in the remaining parameters.

Several methods have been developed for finding the group of identifiable parameters [30, 52, 53]. *Iterative* selection methods apply a step-wise procedure to select one parameter at a time, until no more parameters can be added to the identifiable set. In each step the parameter to be included is selected based on an optimality criteria. For example, the modified Gram-Schmidt orthogonalization method [54] projects all the remaining sensitivity vectors to the subspace spanned by the already selected sensitivity vectors, and includes the parameter corresponding to the one with the largest projection value. This step is repeated until the largest projection value falls below a threshold, which means that the next parameter would significantly interplay with the parameters already included. The computational cost of this method scales up well with the number of sensitivity vectors. However, the drawback of iterative procedures such as this one is that the solution might not be the global optimum, that is, it might fail to find the largest identifiable subset.

Alternatively, we propose to solve the problem of finding the largest identifiable subset of all the estimated parameters using *combinatorial optimization*. To this end we formulate it as a (nonlinear) integer optimization problem, where the goal is to maximize the number of sensitivity vectors included in the set, with the constraint that the corresponding collinearity index is below a threshold CI*.

This algorithm can be stated as

$$\underset{i\in\{0,1\}^{N_\theta}}{\text{maximize}} \sum_{k=1}^{N_\theta} i_k \tag{19}$$

$$\text{subject to } S_i = \text{cat}(\{s_k \mid i_k = 1, \text{ for } k = 1,\ldots,N_\theta\}) \tag{20}$$

$$\text{CI}(S_i) < \text{CI}^* \tag{21}$$

$$i_k \text{ is a binary variable for } k = 1,\ldots,N_\theta \tag{22}$$

where the binary variable i_k indicates if the k-th parameter is included ($i_k = 1$) or not included ($i_k = 0$) in the identifiable group of parameters. The sensitivity matrix corresponding to the selected parameters is S_i, and 'cat' stands for the concatenation of the column vectors in the constraint (20). The collinearity index of this matrix is CI(S_i) and it is determined by computing the minimum eigenvalue as in (17).

This combinatorial optimization problem has an exponentially scaling computational cost, and thus its solution requires an efficient algorithm. We chose the Variable Neighbourhood Search (VNS) technique [55], which is a heuristic global optimization method for integer optimization problems. We used the version of VNS included in the MEIGO Toolbox [56], which is implemented in MATLAB.

We modified this initial formulation of the problem described in Eqs. (19)–(21) after finding that its solution is often not unique: even after maximizing the number of parameters in the subset, there may be multiple subsets that yield a collinearity index below the threshold CI*. Indeed, we found large variability in the solutions if no initial guess was specified. Therefore, we reformulated the optimization problem in two ways, as described in the following paragraphs.

As a first modification, we transformed the collinearity requirement (21) from a 'hard' to a 'soft' constraint (or penalty). The modified optimization problem reads as

$$\underset{i\in\{0,1\}^{N_\theta}}{\text{maximize}} \sum_{k=1}^{N_\theta} i_k - P_1(i) - P_2(i) \tag{23}$$

$$\text{subject to } S_i = \text{cat}\left(\{s_k \mid i_k = 1, \text{ for } k = 1,\ldots,N_\theta\}\right) \tag{24}$$

$$P_1(i) = \frac{1}{2}\text{CI}(S_i)/\text{CI}^* \tag{25}$$

$$P_2(i) = \begin{cases} 0 & \text{if } \text{CI}(S_i) < \text{CI}^* \\ \alpha\left(\text{CI}(S_i) - \text{CI}^*\right)^\beta & \text{otherwise} \end{cases} \tag{26}$$

$$i_k \text{ is a binary variable for } k = 1,\ldots,N_\theta \tag{27}$$

As above, the binary variable i_k indicates if the k-th parameter is included (1) or not included (0) in the selected group of parameters. The penalty P_1 is a monotone increasing (linear) function of the collinearity index CI(S_i), such that P_1 is 0.5 when the collinearity equals to the threshold. Due to this small value, P_1 does not influence the size of the largest subset below the threshold. In this way, when multiple sets of the same size co-exist, the set with smaller collinearity index is always favoured. This results in an unique solution of the optimization problem if there are no sets with identical collinearity index. The second penalty function P_2 represents a soft constraint that is active when the collinearity exceeds the threshold. The steepness of this constraint is tuned by the values of α and β, which we set to $\alpha = 1$ and $\beta = 2$.

Our second improvement of the formulation of the optimization problem consists in providing a good initial guess of the solution using QR decomposition. The rank revealing QR decomposition algorithm, or rrqr [57], rewrites a matrix S as

$$\Pi S = QR, \tag{28}$$

where Q is an orthogonal matrix, R is an upper triangular matrix, and Π is a permutation matrix. Due to the properties of this decomposition, the permutation matrix defines a reordering of the columns of S. In this re-ordered matrix $S_{\text{ro}} = \Pi S$, the most orthogonal columns are located in the left. In other words, the first n columns of the reordered matrix define a linear subspace, and the $(n+1)$-th column has the largest projection value on this subspace among the remaining $N_\theta - n$ columns located to the right of the n-th column. The outcome of the rrqr technique is similar to that of the aforementioned Gram-Schmidt orthogonalization method, but its implementation is more efficient.

We applied rank revealing QR decomposition to the sensitivity matrix, following the procedure described in Algorithm 2. Then, we used the resulted ordering of the sensitivity vectors to initialize the global optimizer. In this way we improved the performance of the global optimizer, which often found larger sets with collinearity index below the threshold value. The whole procedure for identifying the largest non-collinear subset of parameters is summarized in Algorithm 3.

Algorithm 2 Finding the largest identifiable subset of parameters by rank revealing QR decomposition (rrqr)

```
Require: Find sensitive parameters by Algorithm 1 → $\mathcal{I}_{\text{sensitive}}$
Require: Define collinearity threshold: $\text{CI}^*$
Number of sensitive parameters: $N_{\text{sp}} = \text{cardinality}(\mathcal{I}_{\text{sensitive}})$
for all $i \in \mathcal{I}_{\text{sensitive}}$ do
    Normalize the sensitivity columns: $\bar{s}_i \leftarrow \frac{\hat{s}_i}{||\hat{s}_i||}$ (Eq. (16))
end for
Form $\bar{S} \leftarrow \text{cat}(\{\bar{s}_i \mid i \in \mathcal{I}_{\text{sensitive}}\})$, where 'cat' stands for concatenation of a set of column vectors.
$[Q, R, p, r] \leftarrow$ rrqr$(\bar{S})$, where vector $p$ contains the permutation vector
for $i = 2$ to $N_{\text{sp}}$ do
    $S_{ss} \leftarrow \bar{S}(:, p(1:i))$
    $\text{CI}_{ss} = \text{collinearity}(S_{ss})$
    if $\text{CI}_{ss} > \text{CI}^*$ then
        indexLargestIdSetQR $= p(1:i-1)$
        break
    end if
end for
return indexLargestIdSetQR
```

Algorithm 3 Finding the largest identifiable subset of parameters by VNS

```
Require: Find sensitive parameters by Algorithm 1 → $\mathcal{I}_{\text{sensitive}}$
Require: Find largest set by Algorithm 2 → indexLargestIdSetQR
Require: Define collinearity threshold: $\text{CI}^*$
number of sensitive parameters: $N_s = \text{cardinality}(\mathcal{I}_{\text{sensitive}})$
$x_{\text{init}} = \text{zeros}(1, N_s)$
$x_{\text{init}}(\text{indexLargestIdSetQR}) = 1$
solve optimization (23)–(27) using $x_{\text{init}}$ as initial guess
```

The procedure presented in this subsection has similarities with the one proposed by Chu and Hahn [54]. One difference is that we maximize the subset size for a given collinearity threshold, whereas Chu and Hahn adopted the opposite approach, i.e., maximizing parametric identifiability for a pre-specified subset size. Additionally, both methods differ in the optimization technique: we use Variable Neighbourhood Search, which has better scalability than the genetic algorithm chosen in [54]. Recently, Nienałtowski et al. [58] have proposed a method for finding clusters of correlated parameters using so-called canonical correlation analysis (CCA). CCA is an extension of Pearson correlation for measuring multidimensional correlations between groups of parameters. Given two groups of parameters of sizes m and n, with $m < n$, calculation of the canonical correlations provides m measures, which are summarized in a single measure, called MI-CCA. This similarity measure represents the mutual information between the two groups, although it should be noted that average mutual information is equivalent to canonical correlation only if the random variables follow an elliptically symmetric probability model. Nienałtowski et al. use MI-CCA to cluster parameters until an identifiable subset is reached. This approach is sequential and yields a single parameter subset, which is possibly not maximal. In contrast, the methodology described here combines an initial sequential phase with a subsequent combinatorial optimization procedure. The second phase yields several identifiable parameter subsets and usually improves the initial solution.

Finding all largest subsets

As mentioned above, the largest non-collinear subset of parameters is not unique. To realize this, imagine that we have a non-collinear set of the parameters, and consider an additional pair of highly collinear parameters. Since we may add either of these two parameters to the set, but not both of them, we have two potential solutions. The optimization algorithm described above would choose the option with a lower collinearity index.

However, we may also be interested in enumerating *all* the possible sets, instead of only one. Finding all the largest subsets is a combinatorial problem too, which is computationally expensive. A naive approach for solving it could be to generate all possible sets of parameters and compute the corresponding collinearity index. However, note that if two parameters θ_1 and θ_2 are collinear, then any sets including the pair $\{\theta_1, \theta_2\}$ are highly collinear. Using this fact, we developed an incremental procedure for the systematic determination of the sets. We start by considering all possible pairs of parameters and determining their collinearity. Then we extend only those pairs which have a small collinearity index, by considering all possible combinations of a third parameter. This procedure is repeated until either all the sets are highly collinear, or there is only one set containing all the parameters. In this way, summarized in Algorithm 4, we can find all the largest subsets of non-collinear parameters.

Algorithm 4 Finding all the largest identifiable subset of parameters

```
Require: Sensitivity matrix $S$ at the optimal parameters
Require: Define collinearity threshold: $\mathrm{CI}^*$
Given the sensitivity matrix $S = [s_1, \dots s_{N_\theta}]$ and a subset of column indexes $K \subseteq \mathcal{I} = \{1, \dots N_\theta\}$ of $S$. Then let $S_K$ the sub-matrix of $S$ containing the columns specified by indices in $K$, i.e. $S_K := \mathrm{cat}(\{s_i \mid i \in K\})$
Generate all combinations of pairs of parameter indexes: $\mathcal{I}_2 = \{(i,j) \mid i,j \in \mathcal{I}, i < j)\}$
Compute the collinearity index for each element of the set: $\mathrm{CI}_2 = \{\mathrm{CI}(S_K) \mid K \in \mathcal{I}_2\}$
Find sets with small collinearity: $\mathcal{I}_2^* = \{K \mid K \in \mathcal{I}_2, \mathrm{CI}(S_K) < \mathrm{CI}^*\}$
for setSize = 3 to $n_\theta$ do
    generate all the extension sets $\mathcal{I}_{\mathrm{setSize}} = \{K \cup i \mid K \in \mathcal{I}^*_{\mathrm{setSize}-1},\ i \in \mathcal{I},\ i \notin K\}$
    Compute the collinearity index for each element: $\mathrm{CI}_{\mathrm{setSize}} = \{\mathrm{CI}(S_K) \mid K \in \mathcal{I}_{\mathrm{setSize}}\}$
    Find sets with small collinearity: $\mathcal{I}^*_{\mathrm{setSize}} = \{K \mid K \in \mathcal{I}_{\mathrm{setSize}}, \mathrm{CI}(S_K) < \mathrm{CI}^*\}$
    if cardinality($\mathcal{I}^*_{\mathrm{setSize}}$) = 0 then
        report $\mathcal{I}^*_{\mathrm{setSize}-1}$ and $\mathrm{CI}_{\mathrm{setSize}-1}$
        break;
    end if
end for
```

Partitioning the non-identifiable parameters

The two procedures presented above can be used for finding (i) the largest, least collinear subset of parameters, and (ii) all the largest subsets; in both cases, restricted to those subsets whose collinearity falls below a threshold. However, it is often important to understand why certain parameters are *not* identifiable. For example, a parameter may be unidentifiable because the model output has very low sensitivity to changes in its value. But it could also be because it is highly correlated with another parameter, even when both parameters have high sensitivities. Finding small groups of highly collinear parameters can be helpful in determining the exact source of unidentifiability.

The collinearity of a subset always increases when a new parameter is added to the set. For example, considering three parameters, the collinearity of the triplet is always higher than the collinearity of any pairs. Therefore, if a larger set of parameters contains a collinear pair, then the collinearity index of the large set is also large.

If we are interested in *finding the smallest groups of highly collinear parameters*, we can proceed as follows. First we generate all possible pairs of parameters, and compute the collinearity of the corresponding sensitivity vectors. Then we evaluate all possible triplets. The procedure can be extended for the analysis of larger sets. However, due to the combinatorial explosion of the computational cost, this method can be applied only to models of moderate size (with a maximum of roughly 20 parameters).

Visualization of identifiable subsets

It can be useful to represent the identifiability results graphically, because such visualization can provide modellers with insight about how to reformulate their models and/or design new experiments in order to avoid non-identifiable parameters.

With this aim, we display the model structure in the natural network visualization technique. An example is shown in Fig. 1c. The model structure is represented as a graph whose nodes are state variables, observables, stimuli, and model parameters. The edges – which can be directed (arrows) or undirected – have the following meaning: an arrow from node A to node B indicates that node B appears in the equation of A. For example, if the dynamic equation of a state x_1 is $\dot{x}_1 = p_1 \cdot x_2$, the corresponding graph would show two arrows $x_2 \to x_1$ and $p_1 \to x_1$.

More formally, we determine how the state, input variables, stimuli and parameters are connected and influence each other through symbolic manipulation of the model Eq. (1). For this purpose we compute: (i) the Jacobian matrix with respect to the states: $J^{ss}_{i,j} = \frac{\partial f_i}{\partial x_j}$, (ii) the Jacobian of the observation functions with respect to the states: $J^{so}_{i,j} = \frac{\partial g_i}{\partial x_j}$, (iii) the Jacobian of the systems dynamics with respect to the stimuli $J^{si}_{i,j} = \frac{\partial f_i}{\partial u_j}$, and (iv) the Jacobian with respect to the parameters $J^{sp}_{i,j} = \frac{\partial f_i}{\partial \theta_j}$. All these matrices are evaluated symbolically, and then the expressions are converted to a logical 1 (if the symbolic expression is non zero) or 0 when the symbolic result is zero.

Additionally, we can connect parameters by undirected edges if their collinearity is larger than the collinearity threshold.

Implementation: the VisId software tool

We implemented the techniques proposed in subsections "Practical identifiability analysis" and "Visualization of identifiable subsets" as a MATLAB software package called VisId, which is provided as Additional file 1 and can also be downloaded from GitHub (https://github.com/gabora/visid). It is free software, made available under the terms of the GNU General Public License version 3. The VisId toolbox relies on three other MATLAB toolboxes, which are also freely available: AMIGO2 [59] (https://sites.google.com/site/amigo2toolbox/download), which is used to to store, simulate and calibrate the models; MEIGO [56] (http://www.iim.csic.es/~gingproc/meigo.html), which implements the Variable Neighbouring

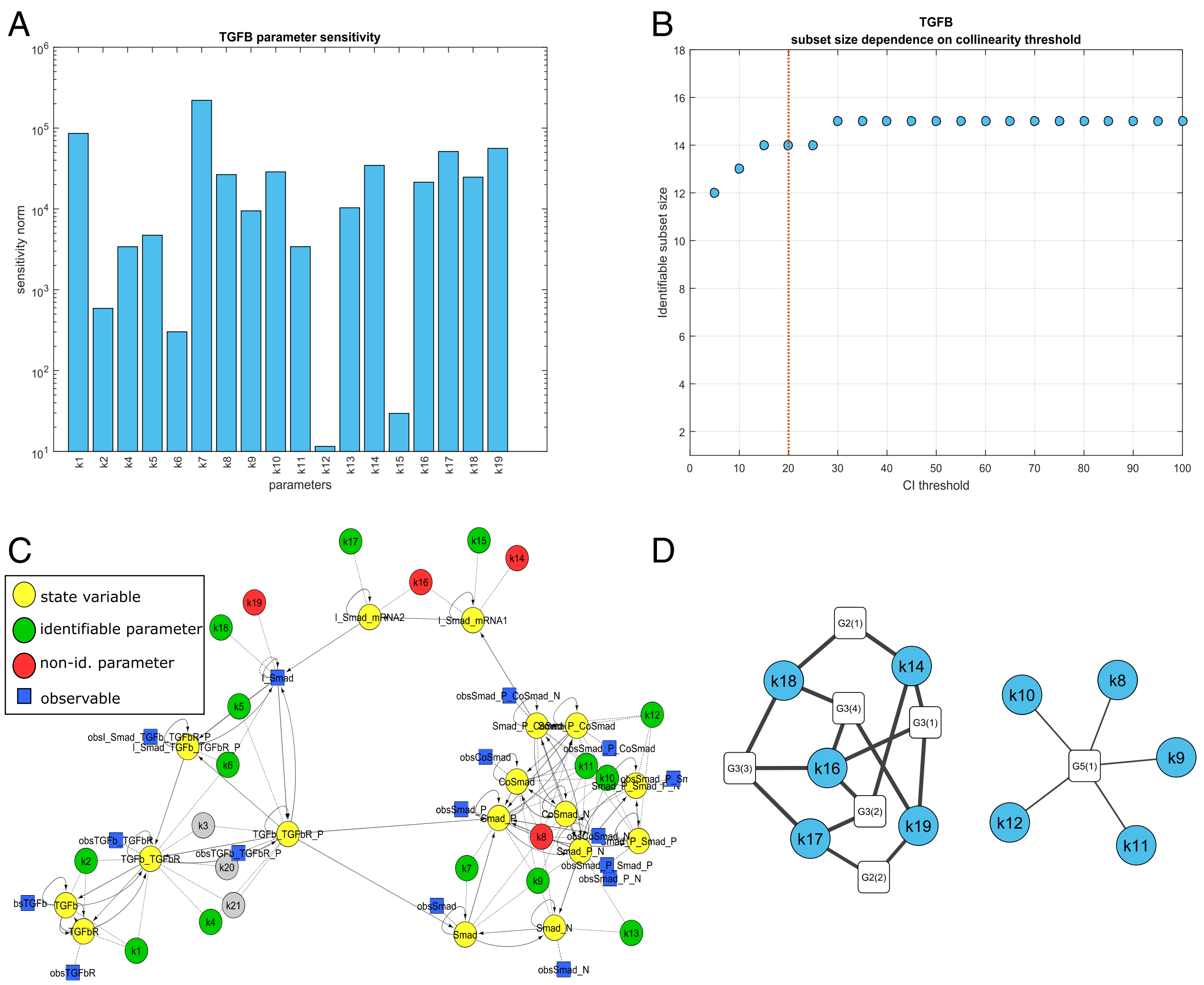

Fig. 1 TGF-β model: panel **a** shows the model sensitivities with respect to the parameters in logarithmic scale. In panel **b** the maximum number of identifiable parameters is depicted depending on the collinearity threshold. The group-size does not change much by the threshold. The *red vertical line* indicates our choice (CI= 20) for the further analysis. In panel **c** the mathematical model is depicted. Nodes indicate states (*yellow*), identifiable (*green*) and not identifiable (*red*) parameters, and observables (*blue*). In panel **d** the interplay among the collinear parameters are indicated. There are 5 groups of parameters: in the first 4 groups triplets of parameters show large collinearity, while in the fifth group 5 parameters are collinear

Search (VNS) optimization method; and (optionally) RRQR (https://www.mpi-magdeburg.mpg.de/1094756/rrqr), which performs the rank revealing QR decomposition used to initialize the global optimizer. Network visualization is performed with Cytoscape [60] (http://www.cytoscape.org/). Further details can be found in Section 4 of Additional file 2.

Results

In this section we demonstrate the application of the methodology presented in the previous section using several dynamic systems biology models of different type and complexity. Their main characteristics are given in Table 1. First we present detailed results of identifiability analysis and visualization for a model of the TGF-β signalling pathway. We also provide similar results for the genetic network that controls the circadian clock in *Arabidopsis thaliana*. Due to their complexity and yet relatively moderate size, these models are well suited as case studies for illustrating the identifiability methodology in depth.

Then we study two large scale benchmark problems included in the BioPreDyn-bench collection [23]. Since the analysis of these latter models is more challenging due to their larger size, we start by demonstrating the performance improvements that can be achieved during parameter estimation using the model calibration procedure proposed in Section "Parameter estimation with regularization and global optimization". Then we perform identifiability analysis and report the corresponding

Table 1 List of models used as case studies and their characteristics

	TGF-β	Circadian	B2	B4
Description	TGF-β signaling pathway	Gene network, *A. thaliana*	Central Carbon Metabolism, *E. coli*	Metabolic model, Chinese Hamster Ovary
Reference	[61]	[62]	[23, 63]	[23, 64]
Parameters	18	27	116	117
States	21	7	18	34
Outputs	16	2	9	13

results, including the graphical representation of the identifiable subset using the natural network visualization.

TGF-β signalling pathway

The dynamic model of the TGF-β signaling pathway was presented in [61] as a tutorial example for model calibration. It has 18 dynamic states and 21 kinetic parameters (k_1–k_{21}), of which 18 need to be estimated. Following [61], we assumed that all the concentrations, except the Smad RNAs ($C_{\mathrm{I\_Smad\_mRNA1}}$ and $C_{\mathrm{I\_Smad\_mRNA2}}$), can be measured in the experiments. The algebraic Equations of the reaction kinetics and the dynamic equations are provided in the Additional file 2.

For the purpose of testing the methodology we generated a training dataset by simulating the model equations using the nominal values of the parameters k_1–k_{21} (numerical values are listed in Additional file 2: Table S1). Then we sampled the simulated trajectories at equidistant time points, and added normally distributed random numbers to the data to mimic measurement errors. Finally, we estimated the model parameters from the generated data set. This approach is widely used for testing calibration methods and assessing the extent to which they recover the nominal parameters. It should be noted that, as the amount of noise in the dataset increases, the information/signal ratio decreases, making the estimation problem more ill-conditioned. This makes it more difficult to recover the correct value of the parameters, but has a small effect in computation times. The numerical values of the estimated parameters are reported in Additional file 2: Table S2.

We started the identifiability analysis by computing the sensitivities of the observations with respect to the estimated model parameters, according to Algorithm 1. We found that all the parameters have a non-negligible influence on the model outputs, thus there are no individually non-identifiable parameters (see Fig. 1a).

Next, following Algorithm 2, we applied QR decomposition and ranked the parameters according to their orthogonality. We then solved the optimization problem (23)–(27) by initializing the variable neighboring search method with the results of the QR decomposition (Algorithm 3). Setting the threshold level for the collinearity index to $\mathrm{CI} = 20$ yielded 14 identifiable parameters, which are shown as green nodes in the network in Fig. 1c. Parameters not present in the identifiable subset are shown as red nodes. Parameters are connected by arrows to state variables (represented by yellow nodes) if they appear in the equation of the corresponding dynamic equation. States which directly influence each other are also connected by directed edges in the same manner. Blue squares represent measurements; a state is connected to a blue square if it appears in the corresponding observation function.

To see how the size of the identifiable subset is influenced by the choice of the collinearity index threshold (CI), we solved the optimization problem for a range of threshold values. The results are depicted in Fig. 1b. As the collinearity index threshold decreases, less parameters are considered identifiable. We can see that the identifiability results are quite robust to the choice of threshold level: the number of identifiable parameters is always between 12 and 15, and it is constant ($= 14$) for a very wide range of CI, $15 \leq \mathrm{CI} \leq 25$.

The results presented so far tell us that the 14 parameters are not correlated. However, they do not inform of the relationships among identifiable and non-identifiable parameters. To investigate this point, we computed the smallest correlated subsets as described in Section "Partitioning the non-identifiable parameters", up to groups of 6 parameters. Figure 1d shows such groups; parameters are depicted as blue circles connected with group identifying nodes (white squares). These nodes are labeled as GX(Y), where X indicates the number of parameters in the group and Y is the group index for a given number of parameters (e.g. G3(2) stands for the second group of three correlated parameters). We found that the large pairwise collinearity between $k_{14} - k_{18}$ and $k_{17} - k_{19}$ explains the non-identifiability of the model parameters only partially. There are 4 groups of triplets and a group of 5 parameters which are highly correlated. The members of the groups and the corresponding collinearity index are reported in Table 2.

It is important to note that collinearity might arise among multiple parameters, even if they are pairwise independent. For example, despite the fact that none of

Table 2 TGF-β model: highly collinear parameter sets. A set ID indicates the number of parameters involved in the collinearity group. They are also depicted in Fig. 1d

Set ID.	CI	Parameters				
G2(1)	2.87e+07	k14	k18			
G2(2)	41.4	k17	k19			
G3(1)	110	k14	k16	k19		
G3(2)	1.37e+03	k14	k16	k17		
G3(3)	1.37e+03	k16	k17	k18		
G3(4)	110	k16	k18	k19		
G5(1)	22.9	k8	k9	k10	k11	k12

the pairs in the group of k_{14}, k_{16} and k_{17} has a high pairwise collinearity, the collinearity index of the triplet is extremely large.

Algorithm 4 found 40 different sets of identifiable parameters with collinearity index ranging between 12.4 and 16, less than the threshold (CI = 20). The sets are reported with the corresponding collinearity index in Additional file 2: Table S3. We can see that parameters $\{k_1 - k_7,\ k_{13},\ k_{15}\}$ are members of all the groups, and they do not participate in any of the small correlated groups in Fig. 1d. From each correlated group of size K, only $K - 1$

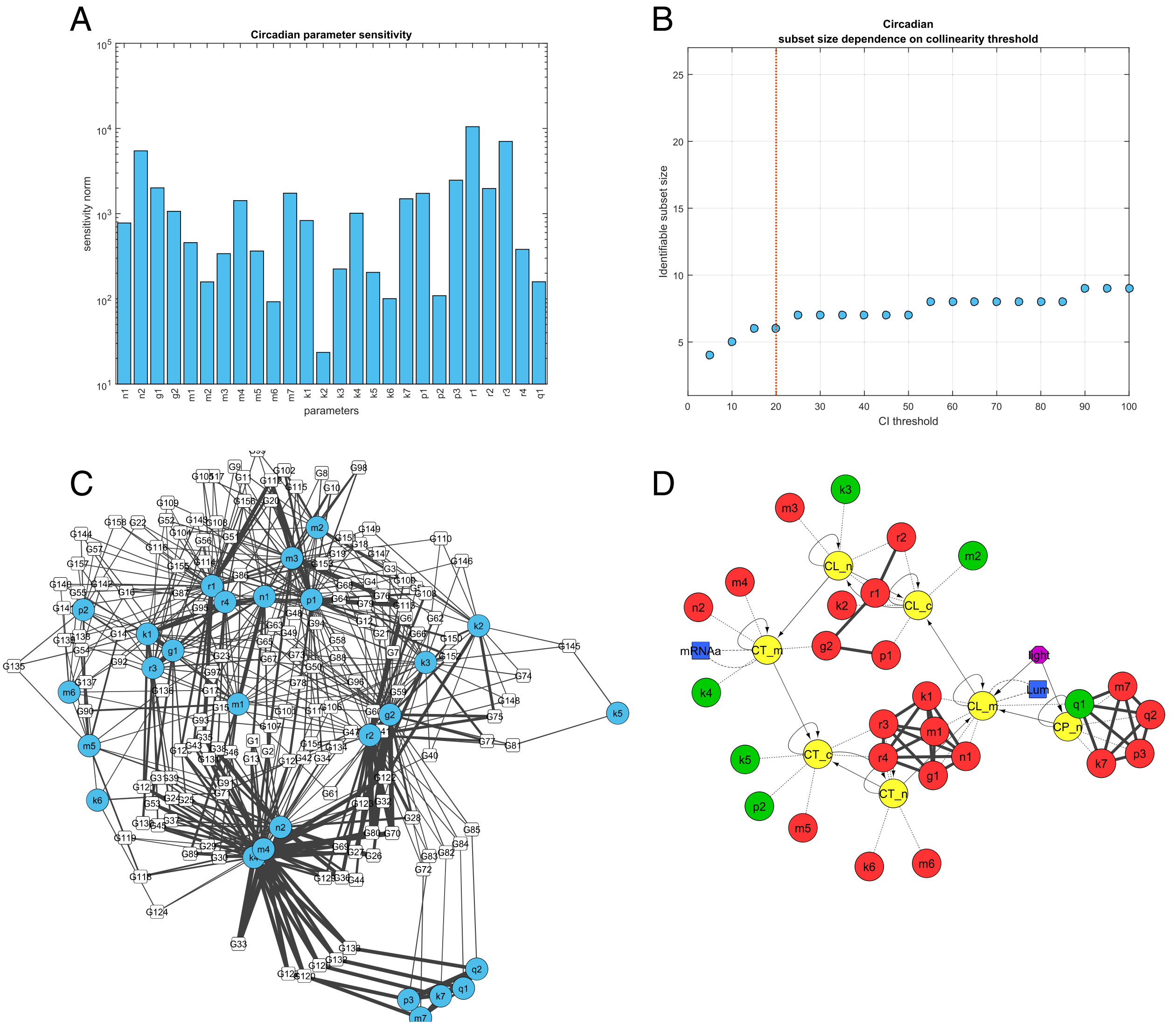

Fig. 2 Circadian clock in *Arabidopsis thaliana*: panel **a** shows the model sensitivities with respect to parameters in logarithmic scale. In panel **b** the maximum number of identifiable parameters is depicted as a function of the collinearity threshold. The *red vertical line* indicates our choice of threshold (CI = 20) used for further analysis. In panel **c** the interplay among the collinear parameters is indicated up to groups of 3 parameters. In panel **d** the mathematical model is depicted. Nodes indicate states (*yellow*), identifiable (*green*) and not identifiable (*red*) parameters, and observables (*blue*). Highly correlated parameter pairs are connected by undirected edges

parameters can participate in the largest set of identifiable parameters.

The aforementioned identifiability procedures can be carried out in a few seconds. Detailed computational costs are shown in Table S6 of the Additional file 2 for all the case studies considered in this paper.

Circadian clock in *Arabidopsis thaliana*

Locke and co-authors [62] described the genetic network controlling the circadian clock in *Arabidopsis thaliana*; the dynamic equations of this model are provided in the Additional file 2.

We generated training data by simulating the model equations with the nominal parameters (Additional file 2: Table S4) in two experimental conditions. In the first one, the model input was kept constant ($\theta_{\text{light}} = 1$), representing continuous light stimulation of the plant. In the second experiment the input was changed pulse-wise in 12-hour cycles, repeated 5 times. As in the previous example, the trajectories were sampled at equidistant time-points and disturbed by pseudo-random noise. Only two states, CT_m and CL_m, were observed. The estimated model parameters are collected in Additional file 2: Table S4.

Although the model outputs showed sensitivity to all the parameters (Fig. 2a), i.e. there were no zero sensitivity vectors, we found that most of the model parameters are non-identifiable due to heavy collinearities. The largest identifiable subset contains only 6 of the 27 parameters, depicted in Fig. 2d by green nodes. The enumeration of the largest sets of identifiable parameters by Algorithm 4 identified 1331 parameter sets.

Benchmarks B2 and B4 from the BioPreDyn-bench collection

In this subsection we analyze two large scale benchmark problems taken from the BioPreDyn-bench collection

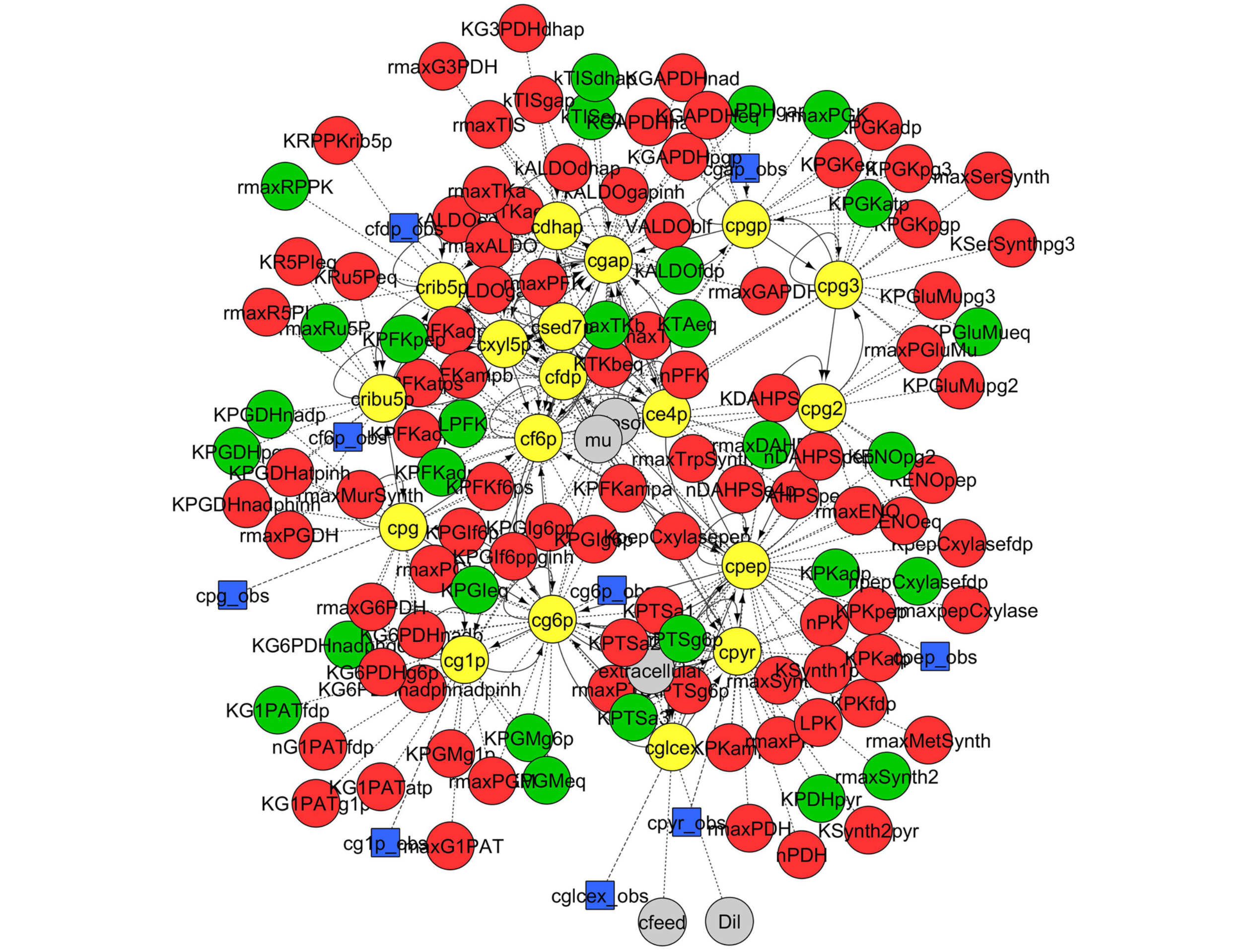

Fig. 3 Representation of the connections in the B2 model using the network diagram formalism. Nodes indicate states (*yellow*), identifiable (*green*) and not identifiable (*red*) parameters, observables (*blue*), and inputs (*grey*). The source file of this figure is provided with the VisId toolbox; using Cytoscape, the user can navigate through it and zoom on different areas to improve the visibility

[23]: the metabolic models of *Escherichia coli* (B2) and Chinese Hamster Ovary cells (B4). They are highly non-linear, partially observed systems with more than 100 unknown parameters, which pose serious challenges for parameter identification. In B4 the calibration data was generated by model simulation and disturbed by random noise, while in B2 it was experimentally measured. Further details about the models and the parameter estimation challenge can be found in [23].

First we use these benchmarks to illustrate the benefits of the parameter estimation strategy proposed in subsection "Parameter estimation with regularization and global optimization", comparing it with the one used in [23]. Both approaches use a hybrid method, eSS [21], which combines a global optimization algorithm (scatter search) with a local search. In [23] the local method of choice was FMINCON; here we compare that configuration with NL2SOL (with and without regularization). Global optimization algorithms use pseudo-random numbers. Hence

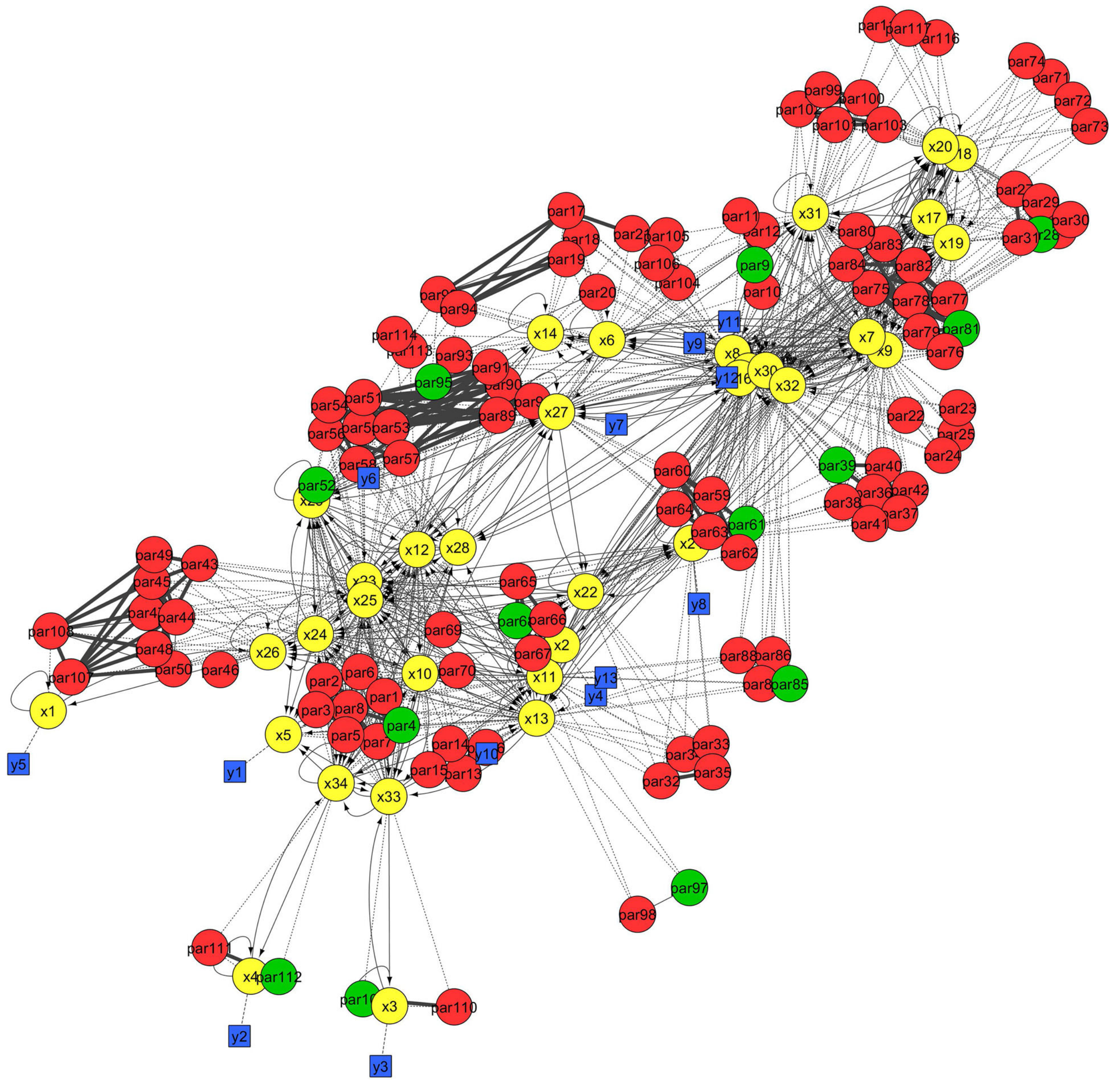

Fig. 4 Representation of the connections in the B4 model using the network diagram formalism. Nodes indicate states (*yellow*), identifiable (*green*) and not identifiable (*red*) parameters, observables (*blue*), and inputs (*grey*)

their performance changes at every run, and the calibration problem should be solved several times to obtain more robust results. Since each optimization takes several hours we limited the number of runs to five for each problem. We used the approximate computation time (CPU time) reported in [23] as the stopping criterion for the model calibration. Convergence curves depict the best objective function value found versus CPU time, and can be used to compare the performance of different algorithms. An optimization method is preferred if it achieves a lower objective function value at earlier CPU time. The best convergence curves (out of 5) corresponding to B2 and B4 are shown in the Additional file 2 for 3 algorithms: (1) eSS-FMINCON, as reported in [23]; (2) eSS-NL2SOL; and (3) eSS-NL2SOL using regularization as recommended in Section "Parameter estimation with regularization and global optimization". From those curves we see that the algorithm (3) proposed here converged earlier than the others to the optimal objective function value (note that log-log scale is used in these curves). We stress that the main purpose of regularization is to avoid overfitting: we do not wish to obtain an excessively good fit, which would indicate that we are reproducing noise instead of the true dynamics. Therefore, regularization should *not* achieve a smaller objective function value.

Next, we apply the identifiability analysis procedures presented in subsections "Practical identifiability analysis" and "Visualization of identifiable subsets" to these two models. For B2 they yield an identifiable subset of size 29, and for B4 of size 13 (recall that both models have a total of 116 parameters). The corresponding networks are shown in Figs. 3 and 4, respectively. It is also possible to find small groups of highly correlated parameters for models of this size; e.g. for B4 we obtained those depicted in Fig. 5.

The aforementioned results show that both models are poorly identifiable in practice for the considered datasets; more informative data would be needed in order to obtain accurate estimates of their parameters.

Discussion and conclusions

In this paper we have presented a workflow to efficiently estimate the parameters of dynamic models and analyze their practical identifiability. Our approach combines an advanced optimization technique, which reduces computation times in parameter estimation, and several identifiability analysis procedures, which can find subsets of identifiable and unidentifiable parameters. Results are visualized using network diagrams, which provide an intuitive representation of the findings and facilitate their analysis and understanding.

Many approaches have been applied to study identifiability of kinetic models, but they suffer from lack of scalability. An advantage of the integrated method presented here is its moderate computational cost, which enables its application to large-scale models; complete results can be obtained in a few hours for models of more than a hundred parameters. Another important aspect is the integration of identifiability analysis with visualization, which presents the results in a way that is easily interpretable for

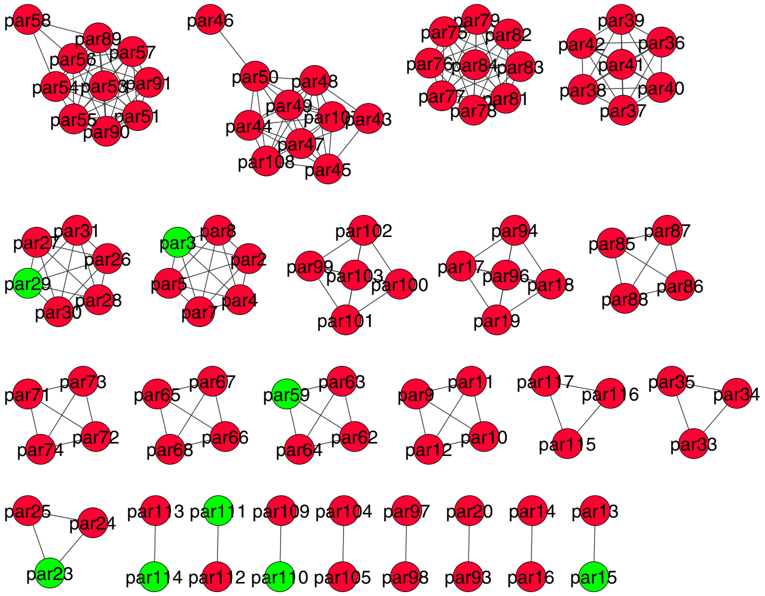

Fig. 5 Visualization of the relationships among highly collinear parameters in the B4 model. The figure shows small groups, whose sizes range between 2 and 10 parameters. Unidentifiable parameters are shown in *red*; identifiable parameters in *green*. Highly correlated pairs are connected by lines

modelers and experimentalists. Currently, its main limitation arises when trying to find *all* the different existing groups of highly correlated parameters: the combinatorial explosion of this particular task makes it feasible only for models of moderate size, i.e. of a few dozens of parameters. However, all the remaining steps of the workflow presented in this manuscript scale up well up to several hundred parameters.

The usefulness of the methodology and workflow presented here goes beyond basic parameter identifiability analysis. The procedure not only (i) determines the largest subset of identifiable parameters, but also (ii) informs about the characteristics of the space of non-identifiable parameters, reporting small groups of highly correlated parameters, and (iii) presents all these results in a coherent and scalable way using visualization techniques, facilitating the understanding of the underlying complex interactions. Uncovering these higher order relationships helps in determining the causes of unidentifiability and provides guidelines for remedying them, e.g. by reformulating the model or by collecting new data through a new experimental design. All this information can be readily used to improve the iterative model-building cycle.

A MATLAB implementation of the identifiability and visualization methodology, which we have called the VisId software package (Additional file 1), is available from GitHub (https://github.com/gabora/visid) as free, open source software. This distribution includes the case studies discussed above.

Additional files

Additional file 1: VisId toolbox. This compressed folder contains the VisId MATLAB toolbox. (ZIP 1030 KB)

Additional file 2: Supplementary material. This document contains detailed descriptions of the case studies and of the VisId toolbox, as well as additional details about the results. (PDF 306 KB)

Abbreviations

CCA: Canonical correlation analysis; CI: Collinearity index; CPU: Central processing unit; eSS: Enhanced scatter search; MI: Mutual information; ODE: Ordinary differential equation; QR: Decomposition of a matrix into an orthogonal matrix (Q) and an upper triangular matrix (R); RNA: Ribonucleic acid; RRQR: Rank revealing QR decomposition; TGF-β: Transforming growth factor beta; VNS: Variable neighbourhood search

Acknowledgements

Not applicable.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 686282 ("CANPATHPRO"), from the EU FP7 project "NICHE" (ITN Grant number 289384), and from the Spanish MINECO project "SYNBIOFACTORY" (grant number DPI2014-55276-C5-2-R).

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the GitHub repository, https://github.com/gabora/visid/tree/master/case_studies.

Authors' contributions

JRB and AG conceived of the study. JRB coordinated the study. AG implemented the methods and carried out all the computations. AFV assisted in the development of the methodology. All authors analysed the results, drafted the manuscript, and read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Author details

[1] BioProcess Engineering Group, IIM-CSIC, Eduardo Cabello 6, 36208 Vigo, Spain. [2] JRC-COMBINE, RWTH Aachen University, Photonics Cluster, Level 4, Campus-Boulevard 79, 52074 Aachen, Germany.

Received: 4 January 2017 Accepted: 25 April 2017
Published online: 05 May 2017

References

1. Wiechert W, Noack S. Mechanistic pathway modeling for industrial biotechnology: challenging but worthwhile. Curr Opinion Biotechnol. 2011;22(5):604–10.
2. Menolascina F, Siciliano V, Di Bernardo D. Engineering and control of biological systems: a new way to tackle complex diseases. FEBS Lett. 2012;586(15):2122–8.
3. Smallbone K, Mendes P. Large-scale metabolic models: From reconstruction to differential equations. Ind Biotechnol. 2013;9(4):179–84.
4. Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnol J. 2013;8(9):1043–57.
5. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology–improving cell factory performance. Metab Eng. 2014;24:38–60.
6. Link H, Christodoulou D, Sauer U. Advancing metabolic models with kinetic information. Curr Opin Biotechnol. 2014;29:8–14. doi:10.1016/j.copbio.2014.01.015.
7. Miskovic L, Tokic M, Fengos G, Hatzimanikatis V. Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes. Curr Opinion Biotechnol. 2015;36:146–53.
8. Srinivasan S, Cluett WR, Mahadevan R. Constructing kinetic models of metabolism at genome-scales: A review. Biotechnol J. 2015;10(9):1345–59.
9. Evangelista PT. Novel approaches for dynamic modelling of e. coli and their application in metabolic engineering PhD thesis, Universidade do Minho. 2016.
10. Vasilakou E, Machado D, Theorell A, Rocha I, Nöh K, Oldiges M, Wahl SA. Current state and challenges for dynamic metabolic modeling. Curr Opin Microbiol. 2016;33:97–104.
11. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150(2):389–401.
12. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO. Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. Cell Syst. 2015;1(4): 283–92.
13. Karr JR, Takahashi K, Funahashi A. The principles of whole-cell modeling. Curr Opin Microbiol. 2015;27:18–24.
14. Macklin DN, Ruggero NA, Covert MW. The future of whole-cell modeling. Curr Opin Biotech. 2014;28:111–5.
15. Villaverde AF, Banga JR. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J R Soc Interface. 2014;11(91):20130505.

16. Jaqaman K, Danuser G. Linking data to models: data regression. Mol Cell Biol. 2006;7(11):813–9. doi:10.1038/nrm2030.
17. Banga JR, Balsa-Canto E. Parameter estimation and optimal experimental design. Essays Biochem. 2008;45:195–210.
18. Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA, Blom JG. Systems biology: parameter estimation for biochemical models. FEBS J. 2009;276(4):886–902.
19. Chou IC, Voit EO. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math Biosci. 2009;219(2):57–83. doi:10.1016/j.mbs.2009.03.002.
20. Gábor A, Banga JR. Robust and efficient parameter estimation in dynamic models of biological systems. BMC Syst Biol. 2015;9(1):74.
21. Egea JA, Martí R, Banga JR. An evolutionary method for complex-process optimization. Comput Oper Res. 2010;37(2):315–24.
22. Dennis JE, Gay DM, Welsch RE. An Adaptive Nonlinear Least-Squares Algorithm. ACM Trans Math Softw. 1981;7(3):348–68.
23. Villaverde AF, Henriques D, Smallbone K, Bongard S, Schmid J, Cicin-Sain D, Crombach A, Saez-Rodriguez J, Mauch K, Balsa-Canto E, Mendes P, Jaeger J, Banga JR. BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. BMC Syst Biol. 2015;9(1):8. doi:10.1186/s12918-015-0144-4.
24. Villaverde AF, Barreiro A. Identifiability of large nonlinear biochemical networks. MATCH Commun Math Comput Chem. 2016;76(2):259–96.
25. Walter E, Pronzato L. Identification of Parametric Models from Experimental Data. Communications and Control Engineering Series. London, UK: Springer; 1997.
26. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev. 2011;53(1):3–39.
27. Chiş O-T, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One. 2011;6(11): 27755.
28. Villaverde AF, Barreiro A, Papachristodoulou A. Structural identifiability of dynamic systems biology models. PLOS Comput Biol. 2016;12(10): 1005153.
29. Brun R, Reichert P, Künsch HR. Practical identifiability analysis of large environmental simulation models. Water Resour Res. 2001;37(4):1015–30.
30. López D, Barz T, Körkel S, Wozny G. Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. Comput Chem Eng. 2015;77:24–42.
31. Ljung L. System Identification: Theory for User. New Jersey: PTR Prentice Hall; 1987. doi:10.1016/0005-1098(89)90019-8.
32. Hawkins DM. The problem of overfitting. J Chem Inform Comput Sci. 2004;44(1):1–12. doi:10.1021/ci0342472.
33. Bauer F, Lukas MA. Comparingparameter choice methods for regularization of ill-posed problems. Math Comput Simul. 2011;81(9): 1795–841. doi:10.1016/j.matcom.2011.01.016.
34. Schittkowski K, Vol. 77. Numerical Data Fitting in Dynamical Systems: a Practical Introduction with Applications and Software. Dordrecht: Springer; 2002, pp. 1–405.
35. Moles CG, Mendes P, Banga JR. Parameter Estimation in Biochemical Pathways : A Comparison of Global Optimization Methods. Genome Res. 2003;13:2467–74. doi:10.1101/gr.1262503.
36. Chen WW, Niepel M, Sorger PK. Classic and contemporary approaches to modeling biochemical reactions. Genes Dev. 2010;24(17):1861–75.
37. Ljung L, Chen T. Convexity issues in system identification. In: 10th IEEE International Conference on Control and Automation. IEEE; 2013. p. 1–9. doi:10.1109/ICCA.2013.6565206, http://ieeexplore.ieee.org/document/6565206/.
38. Esposito WR, Floudas CA. Global optimization for the parameter estimation of differential-algebraic systems. Ind Eng Chem Res. 2000;39: 1291–310.
39. Papamichail I, Adjiman CS. Global optimization of dynamic systems. Comput Chem Eng. 2004;28:403–15.
40. Singer AB, Taylor JW, Barton PI, Green Jr WH. Global dynamic optimization for parameter estimation in chemical kinetics. J Phys Chem. 2006;110(3):971–6.
41. Chachuat B, Singer AB, Barton PI. Global methods for dynamic optimization and mixed-integer dynamic optimization. Ind Eng Chem Res. 2006;45(25):8373–92.
42. Miró A, Pozo C, Guillén-Gosálbez G, Egea JA, Jiménez L. Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems. BMC Bioinformatics. 2012;13(1):90.
43. Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Bio Syst. 2006;83(2–3):248–65. doi:10.1016/j.biosystems.2005.06.016.
44. Sun J, Garibaldi JM, Hodgman C. Parameter estimation using metaheuristics in systems biology: a comprehensive review. Comput Biol Bioinformatics, IEEE/ACM Trans. 2012;9(1):185–202.
45. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. Global Sensitivity Analysis: the Primer. Chichester: John Wiley & Sons; 2008. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470059974.html.
46. Rodriguez-Fernandez M, Banga JR. Senssb: a software toolbox for the development and sensitivity analysis of systems biology models. Bioinformatics. 2010;26(13):1675–6.
47. Turányi T. Sensitivity analysis of complex kinetic systems. Tools and applications. J Math Chem. 1990;5(3):203–48.
48. Saltelli A, Tarantola S, Campolongo F. Sensitivity analysis as an ingredient of modeling. Statist. Sci. 2000;15(4):377–95.
49. Weijers SR, Vanrolleghem PA. A procedure for selecting best identifiable parameters in calibrating Activated Sludge Model No. 1 to full-scale plant data. Water Sci. Technol. 1997;36(5):69–79. ISSN:0273-1223, http://dx.doi.org/10.1016/S0273-1223(97)00463-0.
50. Li P, Vu QD. Identification of parameter correlations for parameter estimation in dynamic biological models. BMC Syst Biol. 2013;7:91. doi:10.1186/1752-0509-7-91.
51. Li P, Vu QD. A simple method for identifying parameter correlations in partially observed linear dynamic models. BMC Syst Biol. 2015;9(1):92.
52. McLean KA, Wu S, McAuley KB. Mean-squared-error methods for selecting optimal parameter subsets for estimation. Ind Eng Chem Res. 2012;51(17):6105–15.
53. Kravaris C, Hahn J, Chu Y. Advances and selected recent developments in state and parameter estimation. Comput Chem Eng. 2013;51:111–23. doi:10.1016/j.compchemeng.2012.06.001.
54. Chu Y, Hahn J. Parameter Set Selection for Estimation of Nonlinear Dynamic Systems. AIChE J. 2007;53(11):2858–70. doi:10.1002/aic.
55. Mladenović N, Hansen P. Variable neighborhood search. Comput Oper Res. 1997;24(11):1097–100.
56. Egea J, Henriques D, Cokelaer T, Villaverde A, MacNamara A, Danciu DP, Banga J, Saez-Rodriguez J. Meigo: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. BMC Bioinf. 2014;15:136.
57. Bischof CH, Quintana-Ortí G. Algorithm 782 : Codes for Rank-Revealing QR Factorization of Dense Matrices. ACM Trans Math Softw. 1998;24(2):254–7.
58. Nienałtowski K, Włodarczyk M, Lipniacki T, Komorowski M. Clustering reveals limits of parameter identifiability in multi-parameter models of biochemical dynamics. BMC Syst Biol. 2015;9(1):65.
59. Balsa-Canto E, Henriques D, Gábor A, Banga JR. Amigo2, a toolbox for dynamic modeling, optimization and control in systems biology. Bioinformatics. 2016;32(21):3357–9. doi:10.1093/bioinformatics/btw411.
60. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
61. Geier F, Fengos G, Felizzi F, Iber D. Analyzing and Constraining Signaling Networks: Parameter Estimation for the User In: Liu X, Betterton MD, editors. Computational Modeling of Signaling Networks. Methods in Molecular Biology. Totowa, NJ:Humana Press;2012. p.23–40. doi:10.1007/978-1-61779-833-7. http://www.springerlink.com/index/10.1007/978-1-61779-833-7.
62. Locke JCW, Millar aJ, Turner MS. Modelling genetic networks with noisy and varied experimental data: the circadian clock in Arabidopsis thaliana. J Theor Biol. 2005;234(3):383–93. doi:10.1016/j.jtbi.2004.11.038.
63. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnol Bioeng. 2002;79(1):53–73. doi:10.1002/bit.10288.
64. Villaverde AF, Bongard S, Mauch K, Müller D, Balsa-Canto E, Schmid J, Banga JR. A consensus approach for estimating the predictive accuracy of dynamic models in biology. Comput Methods Programs Biomed. 2015;119(1):17–28.