SCIENTIFIC REPORTS

OPEN

Learning (from) the errors of a systems biology model

Benjamin Engelhardt[1], Holger Fröhlich[1] & Maik Kschischo[2]

Received: 13 October 2015
Accepted: 07 January 2016
Published: 11 February 2016

Mathematical modelling is a labour intensive process involving several iterations of testing on real data and manual model modifications. In biology, the domain knowledge guiding model development is in many cases itself incomplete and uncertain. A major problem in this context is that biological systems are open. Missed or unknown external influences as well as erroneous interactions in the model could thus lead to severely misleading results. Here we introduce the dynamic elastic-net, a data driven mathematical method which automatically detects such model errors in ordinary differential equation (ODE) models. We demonstrate for real and simulated data, how the dynamic elastic-net approach can be used to automatically (i) reconstruct the error signal, (ii) identify the target variables of model error, and (iii) reconstruct the true system state even for incomplete or preliminary models. Our work provides a systematic computational method facilitating modelling of open biological systems under uncertain knowledge.

Mathematical models of living systems are increasingly used in systems biology to gain important biological insights and to make testable predictions[1–3]. Ideally, a good model covers the essential features of the system whilst still being simple enough for interpretation and mechanistic understanding. Developing a good model is usually a labour intensive manual effort. In biology, the system to be modelled is often only partially known and the distinction of relevant and irrelevant features and variables can be difficult[4–8]. But, even if the major components of a biological system are well known, the sheer complexity of the system might prevent the development of an accurate mathematical model, either because the quantitative data necessary for modelling are not available or because the model is itself too complex to be useful. Thus, researchers in systems biology are frequently confronted with a paradoxical situation: A model is needed to better understand the system and to design informative experiments, but the system is too large and complex for mathematical modelling given the limited amount of knowledge, data and time.

One strategy for modelling is to start with a simple model, which incorporates the most interesting variables and interactions as well as the known input stimuli to the system (Fig. 1a). For example, to model a biochemical reaction network, we might incorporate the concentrations of a few interesting proteins as dynamic state variables and integrate the knowledge about the reactions into simplified assumptions about the interactions between these states. We refer to this simple draft model as the nominal system.

There are two reasons, why the nominal model might not be in sufficient agreement with the experimental data (Fig. 1b): First, some interactions between the nominal state variables could be missing or misspecified. For a reaction network that means there are missing biochemical reactions, incorrect assumptions about the reaction kinetics or inaccurate parameter estimates. Second, the nominal system is in fact—opposed to the typical situation in many areas of physics—open and embedded into a larger dynamic system[9]. Exogenous variables, which are not incorporated, but interact with the nominal model might act as hidden inputs and thereby alter the dynamics of the nominal system. It is the task of the modeller to first identify the most relevant errors in the nominal model and then compare different model versions in order to achieve a better fit to the available experimental data. This process is labour intensive and in many cases a trial and error exercise, even with the help of innovative software and algorithms assisting modelling and model comparison[5–7,10–14].

Here, we introduce a computational method for ordinary differential equations (ODEs), which automatically estimates the model error from the data. ODEs are frequently used in different areas of biology including biochemical reaction networks, pharmacokinetics, pharmacodynamics and population dynamics.

[1]Rheinische Friedrich-Wilhelms-Universität Bonn, Institute for Computer Science, Algorithmic Bioinformatics, c/o Bonn-Aachen International Center for IT, Dahlmannstr. 2, 53113, Bonn, Germany. [2]University of Applied Sciences Koblenz, RheinAhrCampus, Department of Mathematics and Technology, Joseph-Rovan-Allee 2, 53424 Remagen, Germany. Correspondence and requests for materials should be addressed to M.K. (email: kschischo@rheinahrcampus.de)

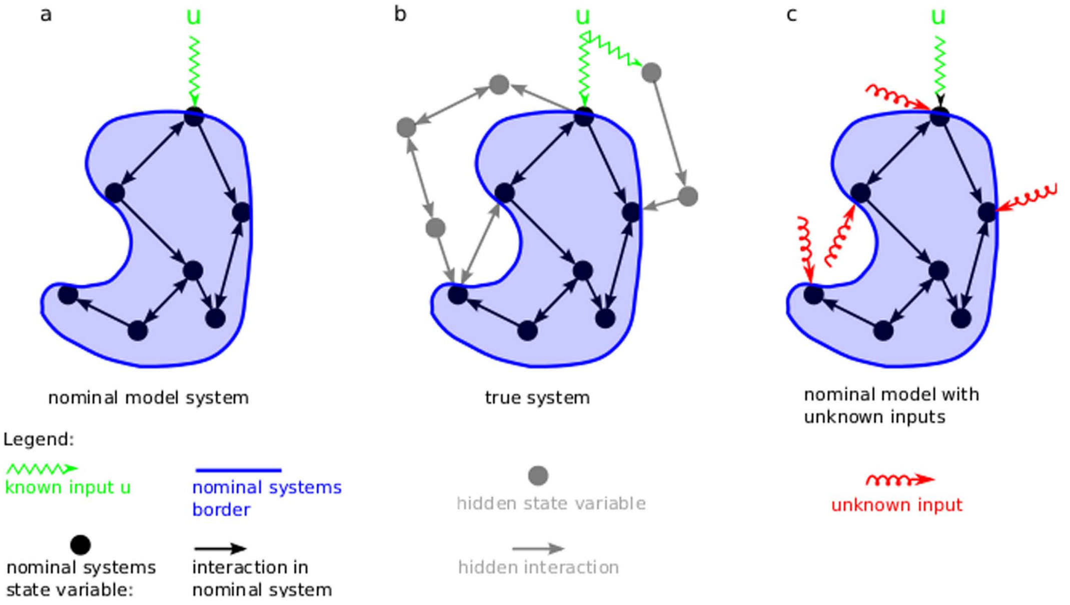


Figure 1. (**a**) The nominal model represents the current assumptions about the true system. The systems model is specified by its dynamic state variables and their interactions, here represented as vertices and edges of a graph. The systems border defines the distinction between internal states and exogenous inputs. The exogenous inputs *u* are assumed to be known. (**b**) In reality, the nominal model is embedded in a larger network outside the nominal systems border. The hidden dynamics of the exosystem interacts with the nominal system. In addition, some interactions between nominal state variables might be missing or misspecified in the nominal model. These model errors can potentially lead to discrepancies between model and experimental data. (**c**) Representation of model errors as hidden inputs to the nominal model. The dynamic elastic-net approach infers the hidden inputs from data and thereby corrects for the bias in the nominal state variables induced by model errors.

The basic idea of the method is to represent errors in the nominal model as hidden inputs to the state variables (Fig. 1c) and to estimate these inputs from the experimental data[8,15–17]. Since this is an inverse problem with potentially many different solutions, we propose a regularised method which provides parsimonious error estimates. Due to its formal similarity to the elastic-net regression approach[18], we term our algorithm the dynamic elastic-net.

The dynamic elastic-net provides important information about the variables in the nominal model, which are targeted by model errors. In addition, the dynamic elastic-net removes the bias in the nominal state variables induced by the model error. This is important for the frequent situation, that not all nominal states (e.g. protein concentrations) can directly be measured. The utility of the dynamic elastic-net is demonstrated here for two established models of the EPO receptor[4] and of the photomorphogenic UV-B signalling network[19]. Further examples including a model for G protein signalling and models for several network motifs as well as some technical details are given in Supplementary text.

Results

The nominal model.

We assume that a nominal ODE model

$$\dot{\tilde{\boldsymbol{x}}}(t) = \widetilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}(t), \boldsymbol{u}(t)) \tag{1a}$$

$$\tilde{\boldsymbol{y}}(t) = \boldsymbol{h}(\tilde{\boldsymbol{x}}(t)) \tag{1b}$$

$$\tilde{\boldsymbol{x}}(0) = \tilde{\boldsymbol{x}}_0 \tag{1c}$$

has been proposed to describe the dynamics of the system of consideration. The state vector $\tilde{\boldsymbol{x}}(t) = (\tilde{x}_1(t), \ldots, \tilde{x}_n(t))^T$ contains the *n* dynamic variables $\tilde{x}_k(t)$, and $\dot{\tilde{\boldsymbol{x}}}(t)$ is the derivative with respect to time *t*. The initial value of the state vector is $\tilde{\boldsymbol{x}}_0$. For a biochemical reaction network, $\tilde{x}_k$ is often the concentration or abundance of the *k*-th species. The function $\boldsymbol{u}(t) = (u_1(t), \ldots, u_m(t))^T$ represents a known external input to the system. The dynamics of the state variables is determined by the function $\widetilde{\boldsymbol{f}} = (\widetilde{f}_1, \ldots, \widetilde{f}_n)^T$ and encodes the model assumptions made in the nominal model. This can be represented as a graph[20], where each node corresponds to one variable and a directed edge from *l* to *k* indicates, that the time derivative of $\tilde{x}_k$ depends on $\tilde{x}_l$ (Fig. 1a). If $\dot{\tilde{x}}_k$ is directly influenced by a known input, we illustrate this by a green zigzag arrow. Typically, not all state variables $\tilde{\boldsymbol{x}}$ can directly be measured. The variables $\tilde{\boldsymbol{y}}(t) = \left(\tilde{y}_1(t), \ldots, \tilde{y}_p(t)\right)^T$ represent all outputs which are experimentally accessible. In equation (1b), we assume that the mapping $\boldsymbol{h}$ from the state $\boldsymbol{x}$ to the output $\boldsymbol{y}$ is known. We use a tilde to highlight, that $\widetilde{\boldsymbol{f}}$ and thus $\widetilde{\boldsymbol{x}}(t)$ are usually not perfectly known due to limited or uncertain knowledge about the true underlying dynamics.

Representation of the model error. The response of the real natural system to a known input stimulus $\boldsymbol{u}(t)$ is usually measured at discrete time points $t_1 \leq \ldots \leq t_N$ and provides experimental observations for the output $\boldsymbol{y}(t_k)$. A part of these data is usually used to estimate the parameters of the model. We consider the initial parameter estimates as part of the nominal model specification $\widetilde{\boldsymbol{f}}$ in equation (1a).

The nominal model is unsatisfactory, when its output $\widetilde{\boldsymbol{y}}(t_1), \ldots, \widetilde{\boldsymbol{y}}(t_n)$ is not in sufficient agreement with the data $\boldsymbol{y}(t_1), \ldots, \boldsymbol{y}(t_N)$. One source of model error comes from hidden inputs to the nominal system, which are caused by dynamical processes exogenous to the nominal system (Fig. 1b). In addition, there might be missing or erroneous interactions between the state variables $\widetilde{\boldsymbol{x}}$ in the nominal model itself. Both types of model error can be represented by hidden inputs $\boldsymbol{w}(t) = (w_1(t), \ldots, w_n(t))^T$ acting on the nodes of the nominal model (Fig. 1c). The "true" dynamics $\boldsymbol{x}(t)$ of the real system can be described by

$$\dot{\boldsymbol{x}}(t) = \widetilde{\boldsymbol{f}}(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \boldsymbol{w}(t) \tag{2a}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t)) \tag{2b}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{2c}$$

Here, the state $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))^T$ represents the same variables as the nominal state $\widetilde{\boldsymbol{x}}(t)$, but we suppress the tilde to distinguish solutions of (2) from that of the the nominal model. The model error is the difference $\boldsymbol{w}(t) = \dot{\boldsymbol{x}}(t) - \widetilde{\boldsymbol{f}}(\boldsymbol{x}(t), \boldsymbol{u}(t))$ between the rate of change of the true system $\dot{\boldsymbol{x}}(t)$ and the nominal system $\widetilde{\boldsymbol{f}}(\boldsymbol{x}(t), \boldsymbol{u}(t))$, evaluated along the true state trajectory $\boldsymbol{x}(t)$. Thus, it incorporates any discrepancy between the true system and the nominal system. The known input $\boldsymbol{u}$ and the output function $\boldsymbol{h}$ are assumed to be identical to the nominal model (1). However, we will also discuss the impact of measurement noise below.

The typical approach to model improvement is to compensate for the model error $\boldsymbol{w}(t)$ by explicit mathematical expressions, often additional differential equations. This increases the number of variables and parameters in the model. Here, we proceed differently by estimating the model error $\boldsymbol{w}$ from the data, what also enables us to correct for the bias $\boldsymbol{x}(t) - \widetilde{\boldsymbol{x}}(t)$ of the state estimate incurred by the nominal model.

Estimating the unmodelled dynamics. To estimate the model error $\boldsymbol{w}(t)$, we use the observer system

$$\dot{\hat{\boldsymbol{x}}} = \widetilde{\boldsymbol{f}}(\hat{\boldsymbol{x}}(t), \boldsymbol{u}(t)) + \hat{\boldsymbol{w}}(t) \tag{3a}$$

$$\hat{\boldsymbol{y}}(t) = \boldsymbol{h}(\hat{\boldsymbol{x}}(t)), \tag{3b}$$

which is a copy of equations (2a) and (2b). The hat marks estimates of the state $\hat{\boldsymbol{x}}(t)$, of the output $\hat{\boldsymbol{y}}(t)$ and of the model error $\hat{\mathbf{w}}(t)$. The latter is obtained by minimising the error functional

$$J[\hat{\boldsymbol{w}}] = \sum_{k=1}^{N} \|\boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}(t_k)\|^2_{Q(t_k)} + \mathscr{R}[\hat{\boldsymbol{w}}]. \tag{3c}$$

The first term in equation (3c) is the weighted mean square error between the measured outputs $\boldsymbol{y}(t_k)$ and the outputs $\hat{\boldsymbol{y}}(t_k)$ of the observer system in equations (3a) and (3b). The weighted square norm

$$\|\boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}(t_k)\|^2_{Q(t_k)} = (\boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}(t_k))^T Q(t_k)(\boldsymbol{y}(t_k) - \hat{\boldsymbol{y}}(t_k)) \tag{3d}$$

contains the symmetric weighting matrix $Q(t_k)$, which is often chosen to be diagonal and can be used to transform outputs of very different magnitude to a common scale or to incorporate precision estimates of the measurements at the different time points t_k. The regularisation term

$$\mathscr{R}[\hat{\boldsymbol{w}}] = \alpha_1 \|\hat{\boldsymbol{w}}\|_1 + \frac{\alpha_2}{2}\|\hat{\boldsymbol{w}}\|_2^2 \tag{3e}$$

$$\|\hat{\boldsymbol{w}}\|_1 = \int_{t_1}^{t_N} \sum_{i=1}^{n} |\hat{w}_i(t)| dt \tag{3f}$$

$$\|\hat{\boldsymbol{w}}\|_2^2 = \int_{t_1}^{t_N} \sum_{i=1}^{n} |\hat{w}_i(t)|^2 dt. \tag{3g}$$

is necessary to avoid overfitting of the data $\boldsymbol{y}(t_k)$ by overly complex estimates $\hat{\boldsymbol{w}}(t)$. The nonnegative parameters α_1 and α_2 determine the relative contributions of the L_1 norm in equation (3f) and of the L_2 norm in (3g).

Minimisation of equation (3c) under the constraints in equations (3a) and (3b) is an optimal control problem[21–23], which needs be solved numerically (see Methods and Supplementary Text).

The combined $L_1 - L_2$ regularisation in equation (3e) is reminiscent of the elastic-net penalty used in regression models[18]. Therefore, we termed our approach the dynamic elastic-net. In analogy to regression, the L_1 term causes some components $\hat{w}_k(t)$ of the estimated model error to shrink to zero (Supplementary text). The amount of shrinkage is determined by α_1, which can be chosen to suppress small error signals or noise distributed over many components of the estimate $\hat{\boldsymbol{w}}$. The resulting sparse estimate is useful, because it provides information about the states of the system which are targeted by systematic model errors, as represented by hidden inputs.

In contrast to regression, a pure L_1 or Lasso type[24] regularisation is not useful in the dynamic setting, because the solution for $\alpha_2 = 0$ can result in unbounded estimates of $\hat{\boldsymbol{w}}(t)$. Even when additional constraints on $\hat{\boldsymbol{w}}(t)$ are imposed, the resulting solution is not smooth and either zero or at the boundaries of the constraints[25]. These insights about the optimal control problem can be obtained from Pontryagin's minimum principle[21,22], as it is detailed in the Supplementary text together with some strategies to chose suitable regularisation parameters α_1 and α_2. In addition to sparse but smooth estimates of the model error, the dynamic elastic-net automatically provides a state estimate $\hat{\boldsymbol{x}}(t)$. Often this is very interesting information, when not all state variables are experimentally accessible.

The optimal control problem in equations (3a–c) for $\hat{\boldsymbol{w}}(t)$ requires the specification of an initial condition $\hat{\boldsymbol{x}}(0) = \boldsymbol{x}_0$, which is often not known or uncertain. Alternatively, one can add the additional constraint

$$\|\boldsymbol{y}(t_1) - \hat{\boldsymbol{y}}(t_1)\|_{Q(t_1)} \leq \Delta_1 \tag{3h}$$

to (3a–c), where Δ_1 is a preset tolerance given for the fit of $\hat{\boldsymbol{y}}(t_1)$ to $\boldsymbol{y}(t_1)$ at time t_1. Similarly, a tolerance Δ_1 can be prescribed to the fit at the last data point by

$$\|\boldsymbol{y}(t_N) - \hat{\boldsymbol{y}}(t_N)\|_{Q(t_N)} \leq \Delta_N. \tag{3i}$$

The tolerance parameters Δ_1 and Δ_N of these optional constraints can often be obtained from error bars of the measurements.

Validation of the dynamic elastic-net. *JAK-STAT signalling example.* To illustrate the dynamic elastic-net estimator for a small and comprehensible model we used established experimental data for the JAK-STAT signal transduction pathway[4]. The four state variables of the system represent unphosporylated cytoplasmatic STAT5 (x_1), phosphorylated monomeric STAT5 (x_2), phosphorylated dimeric STAT5 (x_3) and nuclear dimeric STAT5 (x_4). The nominal model[4]

$$\begin{pmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \\ \dot{\tilde{x}}_4 \end{pmatrix} = \underbrace{\begin{pmatrix} -\theta_1\tilde{x}_1 u \\ \theta_1\tilde{x}_1 u - 2\theta_2\tilde{x}_2^2 \\ \theta_2\tilde{x}_2^2 - \theta_3\tilde{x}_3 \\ \theta_3\tilde{x}_3 \end{pmatrix}}_{\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}},u)} \tag{4a}$$

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} \theta_4(\tilde{x}_2 + 2\tilde{x}_3) \\ \theta_5(\tilde{x}_1 + \tilde{x}_2 + 2\tilde{x}_3) \end{pmatrix}}_{\boldsymbol{h}(\tilde{\boldsymbol{x}})} \tag{4b}$$

describes the phosphorylation of cytoplasmatic STAT5 upon activation of the erythropoietin receptor (known input u), the dimerisation of phosphorylated STAT5 and the export to the nucleus (Fig. 2). Time course data[4] for the amount of cytoplasmatic phosphorylated STAT5 (y_1) and total cytoplasmatic STAT5 (y_2) were used to calibrate the parameters $\theta_1, \ldots, \theta_5$. However, the presence of systematic model error is apparent from the inalterable discrepancy between the experimental data and the nominal model incorporating optimised parameter values (Fig. 2b,c).

To estimate this model error $\boldsymbol{w}(t)$, we numerically fitted the dynamic elastic-net (3) with the nominal model (4) to the output measurements. To quantify the magnitude of the different components, we numerically computed the area under the curve (AUC) of each $|\hat{w}_k(t)|$, i.e. $\mathrm{AUC}_k = \int_{t_1}^{t_N} |\hat{w}_k(t)| dt$. The AUC and the estimated time course $\hat{\boldsymbol{w}}(t) = (\hat{w}_1(t), \ldots, \hat{w}_n(t))^T$ of the model error indicate (Fig. 2e), that the dominant contributions $\hat{w}_1(t)$ and $\hat{w}_4(t)$ of the model error target the states x_1 and x_4, representing the amount of unphosphorylated cytosplasmatic STAT5 and nuclear STAT5. The second component $\hat{w}_2(t)$ of the dynamic elastic-net estimate is identically zero for the whole time interval (Fig. 2e). Apart from the small signal $\hat{w}_3(t)$ initiated after approximately 40 mins, this is consistent with the improved nucleocytoplasmatic cycling model reported in[26], which is based on the same data[4] and incorporates the relocation of dephosphorylated nuclear STAT5 molecules into the cytoplasm. Importantly, the dynamic elastic-net also provides modified estimates for the four STAT5 state variables (Fig. 2f), which are also in good agreement with the nucleocytoplasmatic cycling model (Supplementary Text).

An important problem with regularisation approaches is the choice of the regularisation parameters α_1 and α_2. We used $\alpha_1 = 10$ and $\alpha_2 = 0.2$ in Fig. 2, but we found empirically, that the AUC values clearly indicate the target

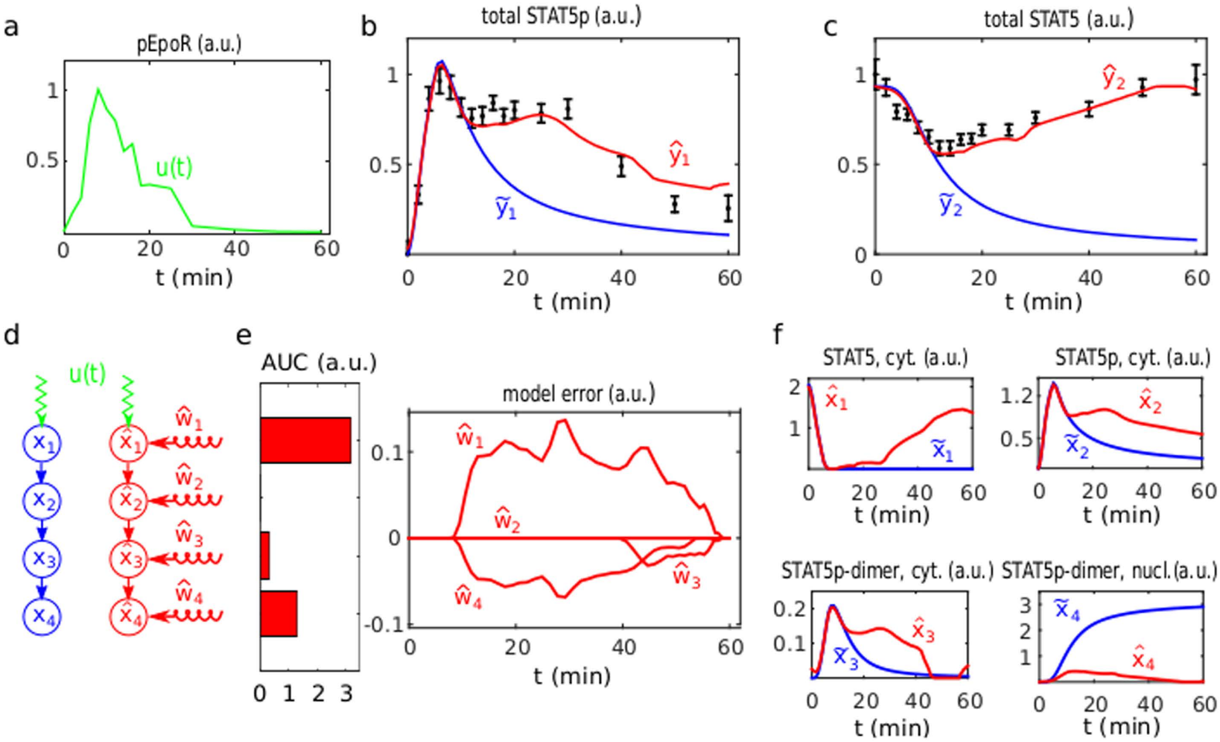


Figure 2. Estimating the model error for the JAK-STAT pathway. (**a**) The known input $u(t)$ is given by linearly interpolated phosphorylation measurements for the erythropoietin receptor[4]. (**b**,**c**) The output measurements[4] (black) for phosphorylated STAT5 (y_1) and total STAT5 (y_2) in the cytoplasm compared to the outputs of the nominal model (blue) and the fit of the dynamic elastic-net (red). (**d**) Graph of the nominal model (blue) and of the observer system (red) with the state variables cytoplasmatic STAT5 (x_1), phosphorylated monomeric STAT5 (x_2), phosphorylated dimeric STAT5 (x_3) and nuclear STAT5 (x_4). (**e**) Dynamic elastic-net estimates $\hat{\boldsymbol{w}}(t) = (\hat{w}_1(t), \ldots, \hat{w}_4(t))^T$ of the model error and the area under the curve (AUC) for the magnitude of each component $|\hat{w}_k(t)|$. (**f**) The state estimates $\tilde{x}_1, \ldots \tilde{x}_4$ obtained from the nominal model (blue) and the dynamic elastic-net observer (see $\hat{x}_1, \ldots \hat{x}_4$ in red).

points of the model error for a wide range of α_1-values (Supplementary Fig. S2). The L_2 parameter α_2 was chosen to balance the smoothness of $\hat{\boldsymbol{w}}$ and the accuracy of the fit to the output measurements. In addition, the bias induced by the double regularisation[18] can be compensated by a simple thresholding strategy: Given an initial estimate $\hat{\boldsymbol{w}}(t) = (\hat{w}_1(t), \ldots, \hat{w}_n(t))^T$ of the model error, we refit the dynamic elastic-net by constraining all the components with small AUC to zero. Thresholding is known in the regression context[27] and we found it to improve the state estimates as well as the time course estimates of the remaining model errors (Supplementary Fig. S3).

The impact of measurement noise and parameter uncertainties. To explore the robustness of the dynamic elastic-net against measurement noise, we added random perturbations to the experimental data[4]. For a given noise level, we generated 500 perturbed data sets by adding Gaussian random numbers with mean zero and standard deviation scaled by a multiple of the empirical standard deviation (see the error bars in Fig. 2b,c) to each experimental data point. Thus, the noise level is defined as a multiple of the empirical standard deviation. The dynamic elastic-net was then fitted to each output sample and the corresponding area under the curve $|w_k(t)|$ for each component of the estimated model error $\hat{\boldsymbol{w}}(t)$ was computed. The plots for these AUC values versus the noise level are shown in Fig. 3a. The median values of the AUC for the components $\hat{w}_1, \ldots \hat{w}_4$ are largely independent of the noise level, but the variability of the AUC estimates increases with measurement noise. Nevertheless, the AUC values for $\hat{w}_1$ and $\hat{w}_4$ are always much larger than zero, whereas the AUC of $\hat{w}_2$ and $\hat{w}_3$ is close or even equal to zero for many samples with higher noise level. This increases the confidence that the nodes $\tilde{x}_1$ and $\tilde{x}_4$ (Fig. 2d) of the nominal JAK-STAT model (4) are the main target points of the model error.

The impact of parameter uncertainty in the nominal model was assessed in a similar way. Parameter estimation algorithms[4,10,26] applied to the nominal model using the experimental data (Fig. 2b,c) provide point estimates and confidence intervals for each component of the parameter vector. These confidence intervals were again scaled by the noise level, yielding an interval for each parameter from which uniform random samples were drawn. Again, we generated 500 modified parameter vectors per noise level. For each parameter sample, the system (4) was taken as the nominal model and the AUC of the resulting estimates $\hat{w}_1, \ldots \hat{w}_4$ was recorded (Fig. 3b). Again, there is no systematic trend for the AUC of the different components of the estimated error $\hat{\boldsymbol{w}}$. However, the variation of the AUC increases much faster than in Fig. 3a. Apart from the different sampling distributions

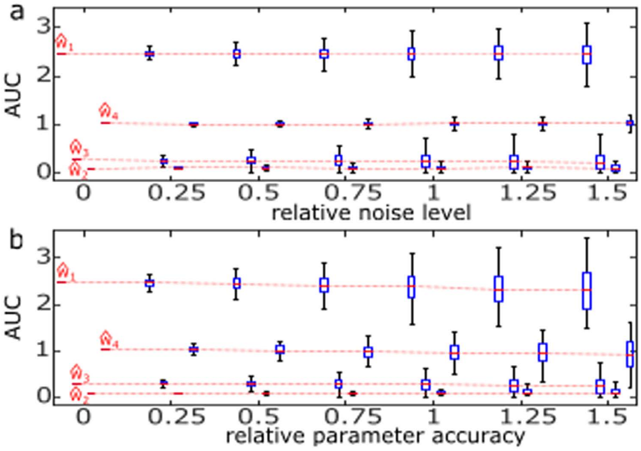


Figure 3. The impact of simulated measurement noise and parameter uncertainty to the dynamic elastic-net estimate in the JAK-STAT model. (**a**) Box plots visualising the variation of the AUC of $|\hat{w}_1(t)|, \ldots |\hat{w}_4(t)|$ for the dynamic elastic-net estimates caused by measurement noise (see main text for details). To ease visualisation, box plots at a given noise level are slightly offset. (**b**) The variation of the AUC caused by parameter uncertainty.

used, this effect is related to the definition of the model error $\boldsymbol{w}$, which is always defined with respect to the nominal model (confer eqution 2a). Hence, the estimated model error $\hat{\boldsymbol{w}}$ contains contributions from both structural and parameter misspecifications in the nominal model. Nevertheless, it is still possible to infer the dominant components $\hat{w}_1$ and $\hat{w}_4$ with high confidence. Similar results were found for the sensitivity against the number of measurement time points (Supplementary text, Fig. S6).

Photomorphogenic UV-B signalling example. As a test case for a larger system, we used a recent model for the coordination of photomorphogenic UV-B signalling in plants[19]. The model consists of 11 ODEs describing the dynamics of protein concentrations $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_{11})^T$ coupled by 10 chemical reactions (Fig. 4). We considered this model as the nominal model in order to test the dynamic elastic-net method for a situation, where the ground truth is known. The model error was simulated by adding the hidden inputs $w_3(t) = w_9(t) = 1000 \cdot [1 - 1/(1 + t)]$ to the nodes $\tilde{x}_3$ and $\tilde{x}_9$. The output function $\boldsymbol{h}(\tilde{\boldsymbol{x}})$ is a linear combination of 7 different state variables (see Supplementary text for all equations). Synthetic data were sampled at discrete time points from the outputs of the true model and Gaussian random perturbations were added to simulate measurement noise (Fig. 4b–f). The dynamic elastic-net with the nominal model was used to reconstruct the model error $\boldsymbol{w}(t)$ and the true state $\boldsymbol{x}(t)$ from these simulated data. The absolute area under the curve for each component of the model error estimate $\hat{\boldsymbol{w}}(t)$ clearly indicates that the states $\tilde{x}_3$ and $\tilde{x}_9$ are targeted by hidden inputs (Fig. 4g), whereas all other components are either very small ($|\hat{w}_6|$) or even zero. This illustrates the sparsity of the dynamic elastic-net estimate, which is a clear advantage over pure L_2 regularisation. The discrepancy $\boldsymbol{w}(t) - \hat{\boldsymbol{w}}(t)$ between the model error and the corresponding estimate relative to the amplitude $A = \max_{t \in (t_1, \ldots, t_N)} w_{3,9}(t)$ of the true model error is at most 10% (Fig. 4h) and mainly caused by numerical inaccuracies. Most importantly, the discrepancy $\boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$ between the true and the estimated state trajectory is almost zero (Fig. 4i), indicating the excellent performance of the dynamic elastic-net as a state observer.

Testing the limitations. As for any inverse method, there are limitations of the dynamic elastic-net method. Some model errors $\boldsymbol{w}(t)$ are unobservable, because there exists a different hidden input function $\boldsymbol{w}^{\dagger}(t)$ which generates an output $\boldsymbol{y}(t)$ which is identical to the output obtained for $\boldsymbol{w}(t)$, see the Supplementary text for a simple example. Other model errors might be practically unobservable, because the output for another hidden input function might not be distinguishable within the measurement errors. A special case are model errors which have no or almost no effect on the output at all. These will not be noticed during modelling and the nominal model will be accepted.

To further test the ability of the dynamic elastic-net to infer the states targeted by the model error, i.e. the non-zero components of the true model error $\boldsymbol{w}(t)$, we systematically simulated perturbations to different nodes and node pairs. First, we simulated model errors $w_k(t) = 1000 \cdot [1 - 1/(1 + t)]$, $w_l(t) = 0$ for $l \neq k$ targeting a single node k in the same way as before. For the nodes $k = 2$, $k = 10$ and $k = 11$ there was no effect on the output (see again Fig. 4b–f) and thus these nodes were omitted from further analysis. In addition, we simulated hidden inputs for all remaining two node combinations. For each of these 36 simulated true models we tested the ability of the dynamic elastic-net to recover the correct target nodes from the AUC of the estimated $|\hat{w}_i(t)|$. We considered a node or a node pair to be correctly recovered, if their AUC was at least 85% of the total AUC over all nodes. By this stringent criterion, we found that two single node errors targeting $\tilde{x}_1$ or $\tilde{x}_4$ were not correctly detected and another single node was predicted to be the target of the model error (Fig. 5a). This indicates, that these model errors are unobservable and the observed output data can be explained by different inputs to different

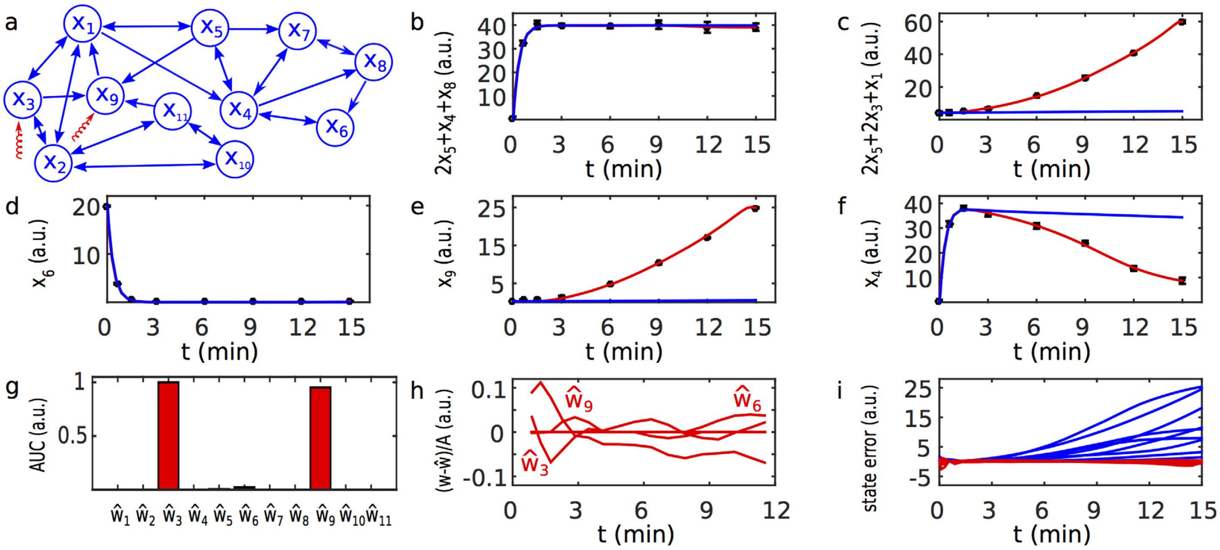
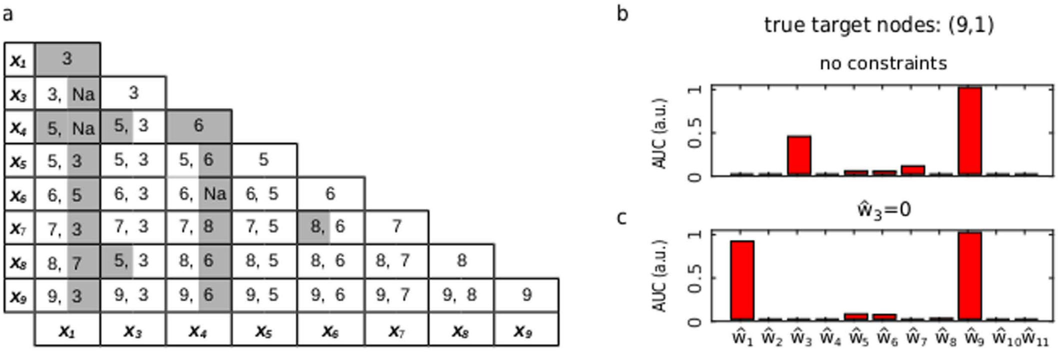


Figure 4. The Photomorphogenic UV-B signalling example. (**a**) The graph (without self loops) of the model states[19]. The target points of the simulated model errors are indicated by the red arrows. (**b–f**) The simulated output $y_1, \ldots, y_5$ with error bars (black), the output of the nominal model (blue) and the output of the dynamic elastic-net (red). (**g**) The AUC of the absolute model errors $|\hat{w}_1(t)|, \ldots, |\hat{w}_{11}(t)|$. (**h**) The components of $\boldsymbol{w} - \hat{\boldsymbol{w}}$ relative to the amplitude A of the true model error. (**i**) The discrepancy $\boldsymbol{x}(t) - \tilde{\boldsymbol{x}}(t)$ between the true state and the nominal model state (blue) compared to the discrepancy $\boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$ of the dynamic elastic-net (red).



x_1	3							
x_3	3, Na	3						
x_4	5, Na	5, 3	6					
x_5	5, 3	5, 3	5, 6	5				
x_6	6, 5	6, 3	6, Na	6, 5	6			
x_7	7, 3	7, 3	7, 8	7, 5	8, 6	7		
x_8	8, 7	5, 3	8, 6	8, 5	8, 6	8, 7	8	
x_9	9, 3	9, 3	9, 6	9, 5	9, 6	9, 7	9, 8	9
	x_1	x_3	x_4	x_5	x_6	x_7	x_8	x_9

Figure 5. Detecting the target nodes of simulated model errors in the UV-B signalling network. (**a**) All nodes and all pairs of nodes were perturbed by a simulated model error. Nodes x_2, x_{10} and x_{11} are omitted, since the simulated error signal had no effect on the output. The rows and the columns correspond to the true target nodes of the model error and the numbers in the cells are the nodes found by the dynamic elastic-net (NA means that no second node was assigned). Gray cells indicate errors made by the dynamic elastic-net for unobservable model errors. (**b**) An example for an unobservable model error. The true target nodes of the model error are (9,1), but the dynamic elastic-net predicts the target nodes (9,3). (**c**) Refitting the dynamic elastic-net under the constraint $\hat{w}_3 = 0$ provides an alternative solution. The other two combinations $\hat{w}_9 = 0$ and $\hat{w}_3 = \hat{w}_9 = 0$ of the nodes (9,3) did not fit the output data.

nodes. With two exceptions ((8, 3) and (7, 6)), the mistakes made by the algorithm for simulated pairwise model errors involve these two state nodes 1 and 4. However, with exception of the combination (1, 4), at least one node is correctly predicted.

These results demonstrate the inherent limitations of any attempt to recover the model error from observed outputs. For an unobservable model error, the true model error $\boldsymbol{w}(t)$ might correspond to a slightly larger value of the error functional (3c) than the minimum $\hat{\boldsymbol{w}}(t)$ obtained by the dynamic elastic-net. A heuristic approach to explore some of these slightly suboptimal solutions is to rerun the dynamic elastic-net with some of the estimated target nodes (from the first run) excluded and to check, whether the output data can satisfactory be fitted with the same level of sparsity. This is illustrated in Fig. 5b for the node pair (9, 1), which was predicted to be (9, 3) by our criterion. Refitting the dynamic elastic under the constraint $\hat{w}_3 = 0$ identifies the correct nodes (9, 1), see Fig. 5c. The two other combinations $\hat{w}_9 = 0$ and $\hat{w}_3 = \hat{w}_9 = 0$ do not provide a satisfactory fit to the data (Supplemental Fig. S9). For the UVB-signaling network we find, that the slightly suboptimal solutions identified by this heuristics always contain the correct target node configuration. The combinatorial explosion of this strategy should typically not be a problem, thanks to the sparsity of the dynamic elastic-net predictions. The decision, which of the predicted target node sets, (9,3) or (9,1), is the correct one can in practice only be made when additional states are measured. However, this example shows, how the dynamic elastic-net provides useful information to select further states for experimental observation[20,28].

Discussion

Efficient computational methods to learn from incomplete model drafts and to direct model improvement are urgently needed. Our proposed dynamic elastic-net approach provides suggestions for the location of these model errors in the network and estimates their dynamic time courses from measured output data. The sparsity of the proposed target points for the model error promotes model improvements in the most parsimonious way. Even for an incomplete nominal model the algorithm can provide estimates for the system states which are not experimentally accessible. This is in stark contrast to many other state estimators including the Kalman Filter[29] for linear systems and its various extensions for nonlinear systems[30,31], which usually require a correctly specified model.

Not all model errors can uniquely be determined from the output. For such unobservable model errors, our strategy to explore alternative, slightly suboptimal solutions might indicate alternative explanations for observed discrepancies between the data and the nominal model. In addition, this approach can also be informative for selecting additional nodes required for observing the state from output measurements[20,28]. Further research is needed to establish the relationship between the network topology and the observability of a model error.

Model errors arising in kinetic reaction systems can originate from erroneous rate equations or lacking reactions. The dynamic elastic-net can detect both types of errors as hidden inputs to the corresponding nodes of the network, but it can not discriminate between these errors. However, knowing the nodes affected by a model error might already be very informative for systematic model improvement.

In view of the rapid progress of technologies to monitor biological dynamics, our approach could have implications for many fields including metabolic engineering, synthetic biology and and pharmacokinetics/pharmacodynamics. As our method is designed for generic ODE models, it can also be applied to challenging modelling tasks in engineering, robotics and in the earth sciences. Our work also raises fundamental questions regarding successful modelling strategies. The approach to manually include more and more details into the model to compensate the initial model errors is often not practical or at least very time consuming. The dynamic elastic-net hence paves the way towards a more principled and systematic way, in which models could be adapted based on experimental data.

Methods

Software.

Simulations were performed in MATLAB (R2014a, The MathWorks, Inc.) using TOMLAB v8.0 with SQOPT 7.2–5 QP and SNOPT 7.2–5 NLP (Tomlab Optimization AB) for solving the optimal control problems. MATLAB scripts are provided as Supplementary material. The computing time for a single run of the dynamic elastic-net on a laptop (Intel CoreTM i5-4200M CPU with 4 × 2.50 GHz and 16 GB RAM) was between 3 seconds and 1 min.

Data and models.

Data for the JAK-STAT system[4] were downloaded from http://webber.physik.uni-freiburg.de/~jeti/PNAS_Swameye_Data. Model equations for the UV-B signaling network[19] were obtained from the Biomodels data base[3], see BIOMD0000000545. For parameter values and mathematical details see the Supplementary text.

References

1. Gunawardena, J. Models in biology: 'accurate descriptions of our pathetic thinking'. *BMC Biol* **12,** 29 (2014).
2. Cvijovic, M. *et al.* Bridging the gaps in systems biology. *Mol Genet Genomics* **289,** 727–734 (2014).
3. Li, C. *et al.* BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* **4,** 92 (2010).
4. Swameye, I., Müller, T. G., Timmer, J., Sandra, O. & Klingmüller, U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. USA* **100,** 1028–1033 (2003).
5. Sunnaker, M. *et al.* Automatic Generation of Predictive Dynamic Models Reveals Nuclear Phosphorylation as the Key Msn2 Control Mechanism. *Sci Signal* **6,** ra41 (2013).
6. Sunnaker, M. *et al.* Topological augmentation to infer hidden processes in biological systems. *Bioinformatics* **30,** 221–227 (2014).
7. Babtie, A. C., Kirk, P. & Stumpf, M. P. H. Topological sensitivity analysis for systems biology. *Proc. Natl. Acad. Sci. USA* **111,** 18507–18512 (2014).
8. Kahm, M. *et al.* Potassium Starvation in Yeast: Mechanisms of Homeostasis Revealed by Mathematical Modeling. *PLoS Comput Biol* **8,** e1002548 (2012).
9. Von Bertalanffy, L. The theory of open systems in physics and biology. *Science* **111,** 23–29 (1950).
10. Balsa-Canto, E., Alonso, A. A. & Banga, J. R. An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst Biol* **4,** 11 (2010).
11. Bachman, J. A. & Sorger, P. New approaches to modeling complex biochemistry. *Nat Methods* **8,** 130–131 (2011).
12. Melas, I. N., Samaga, R., Alexopoulos, L. G. & Klamt, S. Detecting and Removing Inconsistencies between Experimental Data and Signaling Network Topologies Using Integer Linear Programming on Interaction Graphs. *PLoS Comput Biol* **9,** e1003204 (2013).
13. Rodriguez-Fernandez, M., Rehberg, M., Kremling, A. & Banga, J. R. Simultaneous model discrimination and parameter estimation in dynamic models of cellular systems. *BMC Syst. Biol.* **7,** 76 (2013).
14. de Hijas-Liste, G. M., Klipp, E., Balsa-Canto, E. & Banga, J. R. Global dynamic optimization approach to predict activation in metabolic pathways. *BMC Syst. Biol.* **8,** 1 (2014).
15. Mook, D. J. & Junkins, J. L. Minimum model error estimation for poorly modeled dynamic systems. In *AIAA Aerospace Sciences Meeting, Reno, NV* (American Institute of Aeronautics and Astronautics, 1987).
16. Kolodziej, J. R. & Mook, J. D. A novel approach to model determination using the minimum model error estimation. In *Proceedings of the American Control Conference, 2005.* (IEEE Service Center, Portland, OR, USA, 2005).
17. Schelker, M., Raue, A., Timmer, J. & Kreutz, C. Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinformatics* **28,** i529–i534 (2012).
18. Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. B* **67,** 301–320 (2005).

19. Ouyang, X. *et al.* Coordinated photomorphogenic UV-B signaling network captured by mathematical modeling. *Proc. Natl. Acad. Sci. USA* **111,** 11539–11544 (2014).
20. Liu, Y.-Y., Slotine, J.-J. & Barabasi, A.-L. Observability of complex systems. *Proc. Natl. Acad. Sci. USA* **110,** 2460–2465 (2013).
21. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V. & Mishchenko, E. F. *The mathematical theory of optimal processes.* No. v. 1 in Classics of Soviet mathematics (Gordon and Breach Science Publishers, New York, 1986), english ed.
22. Fleming, W. H. & Rishel, R. W. *Deterministic and stochastic optimal control.* Applications of mathematics 1 (Springer-Verlag, Berlin; New York, 1975).
23. Gerdts, M. *Optimal control of ODEs and DAEs.* De Gruyter textbook (De Gruyter, Berlin; Boston, 2012).
24. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective: Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. B* **73,** 273–282 (2011).
25. Vossen, G. & Maurer, H. On L1-minimization in optimal control and applications to robotics. *Optim Control Appl Methods* **27,** 301–321 (2006).
26. Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25,** 1923–1929 (2009).
27. van de Geer, S., Bühlmann, P. & Zhou, S. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5,** 688–749 (2011).
28. Raue, A., Becker, V., Klingmüller, U. & Timmer, J. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos* **20,** 045105 (2010).
29. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *J Basic Eng-T ASME* **82,** 35 (1960).
30. Julier, S. J. & Uhlmann, J. K. New extension of the Kalman filter to nonlinear systems. In *AeroSense'97* vol. 3068, 182–193 (International Society for Optics and Photonics, 1997).
31. Crassidis, J. L. & Junkins, J. L. *Optimal estimation of dynamic systems.* No. 24 in Chapman & Hall/CRC applied mathematics & nonlinear science (CRC Press, Boca Raton, Fla., 2012), 2. ed.

Acknowledgements

B.E. was supported by the Deutsche Forschungsgemeinschaft (DFG) funded Research Training Group 1873.

Author Contributions

B.E. performed the simulations, B.E. and M.K. developed the method, B.E. and M.K. analysed and interpreted the data, B.E. and M.K. draft the manuscript, H.F. citically revised the manuscript, H.F. and M.K. designed the research. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Engelhardt, B. *et al.* Learning (from) the errors of a systems biology model. *Sci. Rep.* **6**, 20772; doi: 10.1038/srep20772 (2016).