Inferring catalysis in biological systems

Ivan Kondofersky[1, 2], Fabian J. Theis[1, 2], and Christiane Fuchs[1, 2]

[1]*Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany*
[2]*Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Boltzmannstr. 3, 85748 Garching, Germany*

April 25, 2016

Abstract

In systems biology, one is often interested in the communication patterns between several species, such as genes, enzymes or proteins. These patterns become more recognisable when time-resolved experiments are performed under possibly several conditions. The time-resolved communication between such species can often be structured by reaction networks such as gene regulatory networks, biochemical networks or signalling pathways. Mathematical modelling of data arising from such networks often reveals important details, thus helping better to understand the studied system. In many cases, however, corresponding models still deviate from the observed data. This may be due to unknown but present catalytic reactions. From a modelling perspective, the question of whether a certain reaction is catalysed leads to a large increase of model candidates. For large networks the calibration of all possible models becomes computationally infeasible very fast. We propose a method which determines a substantially reduced set of appropriate model candidates and identifies the catalyst of each reaction at the same time. To that end, we propose a multiple-step procedure which first extends the network by additional latent variables and subsequently identifies catalyst candidates using similarity analysis methods. The developed method is applied on several simulated examples. Results suggest a good performance even for non-informative data with few observations. This method is applied on CD95 apoptotic pathway and provides new insights into apoptosis regulation.

1 Introduction

A central objective in systems biology is to derive a mathematical model which is used to explain multivariate readouts and thus serves as a tool for detailed investigation of a given biochemical process [1, 2]. Although there are many ways in constructing such models, they generally share the well-known dilemma of models being always only an approximation of reality. This means that regardless of the quality of the model performance, there always remains uncertainty when explaining a biological phenomenon [3]. This uncertainty may arise from different sources, some of which are: the collected data may be subject to various kinds of noise; parameters of complex models may be unidentifiable and thus lead to equal quality of several competing models; the model topology may be specified in a wrong way, e. g. providing a too extreme simplification of reality.

Describing the connection between several variables can be conveniently done using networks or pathways [4]. This has successfully been applied in the field of biology in past decades [5]. In fact, the number of available biological models and reactions has been extensively growing throughout the last decade [6]. For example, signalling pathways are known to be the core mechanism of numerous biochemical processes, such as cell differentiation, cell death or cell division. Additionally, the intracellular behaviour of small molecules can be described in a detailed manner [7, 8]. Small differences in this behaviour may determine the cell fate and thus are of major importance for the overall understanding of the modelled system. Interactions between single components of such networks can occur in various complexities e. g. linear, higher-order or catalytic reactions. The identification of catalytic reactions can be especially challenging if the catalyst of an interaction is not known.

One way of better understanding the underlying mechanisms is to study these pathways over time. Typically the time-resolved data for such systems provides

concentration time-series for some parts of the signalling pathway. For example, immunoblotting [9] allows the assessment of phosphorylation states of proteins and thus provides a measure for the total sum of all phosphorylated molecules in the studied system. Repeated over time, this method provides a time-series of relative protein concentrations which can be used for further analysis. Mathematical modelling of time-resolved variables in a network arrangement is often done by ordinary differential equations (ODEs) [10]. Depending on the size of the studied system, the calibration of ODEs based on data may be computationally demanding.

Considering the possibility that reactions are catalysed expands the model candidate space in an exponential way. To address this challenge, some established model selection techniques, such as greedy stepwise model selection or full best-subset model selection, can be applied [11]. However, these model selection techniques often fail to find the most appropriate model for given data due to either not taking correlation of network components into account or overfitting to data. This means that reducing the model candidate space often comes at a high price of reduced method performance.

Recently, a novel scheme of catalysis identification has been proposed [12]. Here, the authors suggest a model reduction technique which is a graphical approach, taking into account the network topology of the system. Although their approach is able to vastly reduce the model candidate space, this reduction is mostly achieved by eliminating catalysis from certain reactions due to biological prior knowledge rather than performing a statistical comparative study. Furthermore, their approach needs user input suggesting which reactions should be investigated for catalysis.

This manuscript proposes a novel approach for identification of catalysis in biological systems. We first extend the known network by including hidden components and estimate their time courses with a combination of smoothing splines

and least squares approach. In the next step, we compare those time courses of the hidden components to the time courses of network components. Here, we measure similarity between two time courses based on correlation and L^2-distance and associate each comparison with a score. Subsequently, we choose a threshold and only consider components with high scores to be relevant catalyst candidates. The reduced number of model candidates is finally calibrated and the best model is chosen via maximum likelihood.

We propose this approach as an extension of our previous work [13] where we considered a systematic network extension by by estimating the time courses of latent network components. The work in the present manuscript extends the available method in several ways. First, we now consider multiple latent components as opposed to the identification of a single additional component. Moreover, we formulate the coupling of the additional latent components to the original components in a non-linear way as opposed to the linear network extensions in our previous approach. Most importantly, we compare these latent components to the original network components and replace them accordingly if we find high similarity. The final models do not contain any unspecified components. This is different from having a single final model with an open for interpretation additional component. Overall, our new approach is an application-driven method extension of [13].

Concerning the application, the goal of the proposed method is not a dramatic alteration of the network topology of the studied biological system. In that sense, we do not want to introduce new reactions in the system but only check if an alteration of existing ones (provided e. g. by literature knowledge) in terms of catalysis significantly improves the explanation of a given dataset. General network topology identification is a complex field studied extensively in literature [4, 5, 14, 15] and is beyond the scope of this manuscript.

This paper is organised as follows: In Section 2 we define our modelling ap-

proach and explain how we estimate model parameters. This ultimately leads to building a score for every network component, which describes its affinity to catalyse a certain reaction. Section 3 applies the developed technique to different simulated scenarios and to a real data example - the CD95 apoptosis pathway. Section 4 concludes this paper and discusses strengths and limitations of the proposed method.

2 Methods

In this section, we present the developed method for inferring catalytic reactions in biological systems. We first describe the types of systems we aim to study with this method in the context of catalysis. Then, we introduce the individual steps of the estimation procedure. In brief, we model an extended system with external or hidden catalysts and afterwards compare these external catalysts to observed network components by construction of a similarity score. This allows us to preselect only a small number of model candidates, which we then compare in more detail with a likelihood approach. Overall, this results in obtaining the most appropriate model for the data without wasting computational resources.

2.1 Mathematical formulation of catalysis

We consider N-dimensional ODEs with m reaction fluxes, which can be formulated as

$$\frac{d\boldsymbol{x}(t)}{dt} = \dot{\boldsymbol{x}}(t) = \boldsymbol{S}\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k}) = \sum_{g=1}^{m} \boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k}) \tag{1}$$

with $N \times m$ stoichiometry matrix $\boldsymbol{S}$ with scalars $s_{i,g} \in \mathbb{Z}$. $\boldsymbol{s}_{\cdot,g}$ is the g-th N-dimensional column of this matrix and $\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k}) = (v_1(\boldsymbol{x}(t);\boldsymbol{k}),\ldots,v_m(\boldsymbol{x}(t);\boldsymbol{k}))^T : \mathbb{R}^N \times \mathbb{R}^p \to \mathbb{R}^m$ is an m-dimensional flux function with arguments $\boldsymbol{x}(t) = (x_1(t),\ldots,x_N(t))^T \in$

$\mathbb{R}^N_{\geq 0}$ as the non-negative network component concentration functions and $\boldsymbol{k} \in \mathbb{R}^p$ as the reaction rate constants. We assume the individual flux functions $v_g(\boldsymbol{x}(t);\boldsymbol{k})$ to be linear combinations of $\boldsymbol{x}(t)$. We further denote $x_i(\boldsymbol{t}) = (x_i(t_0),\ldots,x_i(t_n))^T$ as the values of the i-th component at time points t_l and we choose suitable initial values $x_i(t_0)$, where $t \geq t_0$ represents the time. Furthermore, $\dot{\boldsymbol{x}}(t) = (\dot{x}_1(t),\ldots,\dot{x}_N(t))^T$ is the derivative of $x(t)$ with respect to time. The components $x_i(t)$ may be observed or unobserved. From a biological point of view, the components $\boldsymbol{x}(t)$ describe the time courses of e. g. metabolite concentrations in blood; the stoichiometry matrix $\boldsymbol{S}$ holds the integer-valued stoichiometric coefficients that represent the net change by a particular reaction; the vector $\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k})$ represents the flux through all reactions in the network.

Modelling network dynamics as in (1) presents a general way of describing biological systems. However, this description is often not sufficient to explain the observed network dynamics. To improve the discrepancies between model fit and observed data, one can choose different strategies. Examples of such approaches include more complex interactions, time-varying reaction rates or complex formations introducing external latent variables [13]. Catalysis is an additional way of improving the model fit and at the same time maintaining a low level of model complexity. Furthermore, it represents the modelling of a realistic scenario since catalysis is an often-occurring pattern in many biological systems [16]. A catalytic reaction can be included into (1) by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{S}(\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k}) \circ \boldsymbol{h}(t)) = \sum_{g=1}^{m} \boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k}) h_g(t) = \psi(\boldsymbol{S},\boldsymbol{v},\boldsymbol{x}(t),\boldsymbol{k},\boldsymbol{h}(t)) \quad (2)$$

with $\circ$ denoting the Hadamard product (componentwise multiplication), $\boldsymbol{h}(t) = (h_1(t),\ldots,h_m(t))^T \in \mathbb{R}^m_{\geq 0}$ representing the concentration of the m non-negative catalysts and ψ as a summarizing function for the right-hand side of the ODE. We will

later estimate the unknown catalysts $\boldsymbol{h}(t)$. From a biological point of view the components $\boldsymbol{h}(t)$ can be regarded as components of a network holding the time-resolved concentration measurements in the same way as $\boldsymbol{x}(t)$. The difference between $\boldsymbol{h}(t)$ and $\boldsymbol{x}(t)$ is that $\boldsymbol{h}(t)$ are *unknown* components as opposed to $\boldsymbol{x}(t)$ which represent species which were already used for the construction of the network. We further restrict our models to $\boldsymbol{h}(t)$ having a meaningful effect on ψ and thus require

$$\forall \varepsilon > 0, \forall t \geq t_0, \forall h_2(t) \in U_\varepsilon(h_1(t)) : \psi(h_1(t)) \neq \psi(h_2(t)) \tag{3}$$

with $U_\varepsilon(h_1(t)) = \{h_2(t) \in \mathbb{R}_{\geq 0} :|| \, h_1(t) - h_2(t) \, ||^2 < \varepsilon\}$ and $|| \cdot ||^2$ denoting the L^2 norm. Furthermore, for the rest of the manuscript we assume $\boldsymbol{S}$ and $\boldsymbol{v}$ to be known in parametric form, e. g. from literature. This assumption can be relaxed and $\boldsymbol{S}$ and $\boldsymbol{v}$ can also be estimated with our method, which increases the number of unknown parameters. The assumption seems reasonable since we want to apply our method to well-studied systems where information about $\boldsymbol{S}$ and $\boldsymbol{v}$ is available.

2.2 Estimation of hidden catalysts

In the first part of the proposed method, we estimate $\boldsymbol{h}(t)$ and interaction parameters $\boldsymbol{k}$. Here, we approximate the observed time courses of $\boldsymbol{x}(t)$ by smoothing splines as done e. g. in [13] resulting in an estimtate $\hat{\boldsymbol{x}}(t)$. This also presents an immediate approximation of $\dot{\boldsymbol{x}}(t)$ as $\hat{\dot{\boldsymbol{x}}}(t) = \frac{\partial}{\partial t}\hat{\boldsymbol{x}}(t)$. Subsequently, we plug in these approximations into (2) and estimate the unknowns $\boldsymbol{h}(t)$ and $\boldsymbol{k}$ by minimizing the quadratic distance between $\hat{\dot{\boldsymbol{x}}}(t)$ and the predicted time courses:

$$(\hat{\boldsymbol{k}}, \hat{\boldsymbol{h}}(t)) = \underset{\boldsymbol{k}, \boldsymbol{h}(t)}{\operatorname{argmin}} \left[|| \, \hat{\dot{\boldsymbol{x}}}(t) - \psi(\boldsymbol{S}, \boldsymbol{v}, \hat{\boldsymbol{x}}(t), \boldsymbol{k}, \boldsymbol{h}(t)) \, ||^2 \right] \tag{4}$$

with $|| \cdot ||^2$ denoting the L^2 norm. In general, (4) has more unknown parameters than the ODE dimension and thus some estimated parameters in (4) may not be

identifiable. One possibility to reduce the estimated parameter space is to set non-identifiable parameter entries in $\boldsymbol{k}$ to a constant, e. g. to 1. Such non-identifiable parameters can occur wherever the ODE has entries such as $\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k}) h_g(t)$ where both $\boldsymbol{k}$ and $h_g(t)$ cannot be estimated simultaneously without additional prior information or constraints due to $\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k}) h_g(t) = (a h_g(t)) \cdot \frac{\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k})}{a}$, $\forall a \in R_{\neq 0}$, $\forall t > t_0$. Approximations for the non-identifiable entries in $\boldsymbol{k}$ will be found in the second step of the proposed method. After elimination of such non-identifiable parameters, (4) is numerically optimized e.g. by a gradient descent method. The result of this first step are the approximations of the components of $\boldsymbol{h}(t)$, which can be grouped in a set:

$$\mathcal{H} = \{\hat{h}_g(t_j)\}_{g=1,\ldots,m;j=0,\ldots,n}. \tag{5}$$

Once these approximations are found, we perform similarity analysis to relate them to the network components which we describe in the following.

2.3 Relating hidden catalysts to network components

In the second step, we compare the entries of $\mathcal{H}$ to the set

$$\mathcal{X} = \{X_{ij} \mid X_{ij} = \hat{x}_i(t_j) \text{ if } i \leq N, X_{N+1}(t_j) = 1\}_{i=1,\ldots,N+1;j=0,\ldots,n}, \tag{6}$$

which contains the spline-approximated time courses of the network components $\hat{\boldsymbol{x}}$ with an additional component $x_{N+1}(t)$, which is equated to 1 for all t and is thus comparable to the intercept term in a regression context. This comparison between entries in $\mathcal{H}$ and $\mathcal{X}$ is done in terms of two different measures of similarity. On the one hand, we measure similar time-course shapes of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{h}}$ by calculating the Pearson correlation coefficient between entries in $\mathcal{H}$ and $\mathcal{X}$, resulting in the set of

correlations $\mathcal{C}$:

$$\mathcal{C} = \{C_{ig} \mid C_{ig} = \mathrm{cor}(\mathcal{X}_{i\cdot}, \mathcal{H}_{g\cdot})\}_{i=1,\ldots,N;g=1,\ldots,m}. \tag{7}$$

On the other hand, the proximity between $\mathcal{X}$ and $\mathcal{H}$ is measured by the L^2 distance and these values are collected in a set $\mathcal{L}$:

$$\mathcal{L} = \{\min_{\kappa_{ig}\in\mathrm{R}}(|| \mathcal{X}_{i\cdot} - \kappa_{ig}\mathcal{H}_{g\cdot}) ||^2)\}_{i=1,\ldots,N+1;g=1,\ldots,m} \tag{8}$$

with scaling parameters κ_{ig}, which are used to find the best scaling of $\mathcal{H}_{g\cdot}$ so that the L^2-distance to $\mathcal{X}_{i\cdot}$ is minimized. Recall that while optimizing (4), we set the non-identifiable parameters in $\hat{\boldsymbol{k}}$ equal to 1. With the optimization in (8), these parameters can now be estimated as the minimizers in (8).

The two sets, $\mathcal{C}$ and $\mathcal{L}$, measure two different aspects of similarity (shape and proximity), which are suitable for comparing two time series. Furthermore, *smaller* values in $\mathcal{L}$ and *larger* values in $\mathcal{C}$ correspond to higher similarities. Therefore, they are combined and weighted to form a set of scores $\mathcal{S}$, which can be used to easily identify catalysis candidates. To construct such a set, single entries of $\mathcal{C}$ and $\mathcal{L}$ are combined and scaled in the unit interval. Formally, we build

$$\mathcal{S} = \left\{ \alpha \frac{\max(\mathcal{L}_{i\cdot}) - \mathcal{L}_{ig}}{\max(\mathcal{L}_{i\cdot}) - \min(\mathcal{L}_{i\cdot})} + (1-\alpha)\frac{\mathcal{C}_{ig} - \min(\mathcal{C}_{i\cdot})}{\max(\mathcal{C}_{i\cdot}) - \min(\mathcal{C}_{i\cdot})} \right\}_{i=1,\ldots,N+1;g=1,\ldots,m} \tag{9}$$

with $\max(\mathcal{L}_{i\cdot}) := \max_{g'}\{\mathcal{L}_{ig'} \mid g' = 1,\ldots,m\}$ and $\min(\mathcal{L}_{i\cdot})$, $\max(\mathcal{C}_{i\cdot})$ and $\min(\mathcal{C}_{i\cdot})$ defined in the same way. The special cases of $\max(\mathcal{L}_{i\cdot}) = \min(\mathcal{L}_{i\cdot})$ and $\max(\mathcal{C}_{i\cdot}) = \min(\mathcal{C}_{i\cdot})$ can be excluded without loss of generality. If one of those cases occurs, it means that we cannot distinguish between all candidates either on basis of distance or correlation. In this case all candidates describe the data equally and no candidate reduction can be achieved. The scalar weighting parameter $\alpha \in [0,1]$ presents a

possibility to manually adapt the importance of the two different measures. For values $\alpha > 0.5$, the proximity measure $\mathcal{L}$ gets a higher weight in the overall score computation and contrarily for $\alpha < 0.5$ the correlation measure $\mathcal{C}$ becomes more influential on the score calculation. In the examples in this manuscript we used $\alpha = 0.5$ which leads to equal weighting of both measures. Overall, $\mathcal{S}$ has values in the unit interval with a value of 1 in $\mathcal{S}_{i,g}$ meaning that $\mathcal{X}_{i\cdot}$ is best correlated and has the lowest L^2 distance (after scaling) to $\mathcal{H}_{j\cdot}$, making $\mathcal{X}_{i\cdot}$ the most obvious catalyst candidate for the g-th reaction. It is possible that multiple entries of $\mathcal{X}_{i\cdot}$ have a high score close to 1 which may then all be considered as catalyst candidates. We define a threshold τ, which is used to filter components with high associations for catalysts of a given reaction by the rule:

$$\mathcal{S}_{i,g} > 1 - \tau \Rightarrow \mathcal{X}_{i\cdot} \text{ candidate for } g\text{-th reaction.} \tag{10}$$

The index τ can be used in various ways. If τ equals 1, all components are classified as possible catalyst candidates (no model reduction), whereas if τ equals 0, at most one component per reaction is chosen as a possible catalyst (and only if it outperforms all other components in distance *and* correlation measure). Generally, τ can be used to control the trade-off between a large number of acceptable models and a high probability of finding the most appropriate model with the proposed algorithm. In practice and for the examples presented in this manuscript, we found that setting τ to 0.1 presents a reasonable choice.

2.4 Choosing most appropriate model from reduced model candidates with maximum likelihood

After performing the described steps above, a reduced set of models M_τ is obtained as a subset of all possible models, M. Additionally, we obtain the set $\bar{M}_\tau$, which

describes the models which are not considered to be appropriate for the characterization of the studied system. It holds that $M_\tau \subseteq M$ and for large systems and for small τ we usually obtain $\mid M_\tau \mid \ll \mid M \mid$. Without loss of generality, we can assume that $\mid M_\tau \mid > 1$ and we still need to find the most appropriate model from the set M_τ. In this context, we apply a maximum likelihood optimization scheme to determine the model of choice. Therefore, we first specify an error distribution of the observed data:

$$x_i^{\text{obs}}(t_j) = x_i(t_j) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{11}$$

In applications, $x_i^{\text{obs}}(t_j)$ often has a positive domain in which case (11) might be ill-defined. One possible solution for this might be log-normally distributed multiplicative noise as in [13]. Such error model is straightforward here, however, for reasons of notation simplicity, we only consider normally distributed errors in the manuscript and in the example section. The distribution of ε_{ij} immediately propagates to the measurements:

$$x_i^{\text{obs}}(t_j) \big| x_i(t_j) \overset{\text{iid}}{\sim} \mathcal{N}\left(x_i(t_j), \sigma^2\right). \tag{12}$$

While the true time course $x_i(t)$ is unknown, it has already been approximated by smoothing splines and we can plug in this approximation in (12):

$$x_i^{\text{obs}}(t_j) \big| \hat{x}_i(t_j) \overset{\text{iid}}{\sim} \mathcal{N}\Big(\hat{x}_i(t_j), \sigma^2\Big). \tag{13}$$

With this last approximation, we are now able to formulate a likelihood function, which measures the overall agreement between model and data depending on the model parameters

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=0}^{n} f_{\boldsymbol{\theta}}(x_i^{\text{obs}}(t_j)) \tag{14}$$

with $f_{\boldsymbol{\theta}}(x_i^{\text{obs}}(t_j))$ representing the probability density function of the normal distribution in (13) and $\boldsymbol{\theta}$ representing the conglomerate of parameters $(\boldsymbol{k}, \sigma^2)^T$. The maximum of the likelihood function can be found using a gradient descent method and the parameters corresponding to this optimum are then called $\hat{\boldsymbol{\theta}}$. The dimension of $\hat{\boldsymbol{\theta}}$ does not change regardless of the number of reactions catalysed and the different combinations of catalytic reactions. Therefore, comparing models only by comparing likelihoods instead of using e. g. information criteria is possible in this setting.

In the next section, we will test the developed method on several artificial datasets and also apply it on real-world data from a biochemical pathway.

3 Applications

In this section, we apply our method on artificial and real-world data. We perform two excessive simulations in which we test the applicability and effectiveness of our method. First, we simulate random networks of different sizes and estimate the reaction catalysts in those networks. Second, we fix the network size at $N = 5$ and test our method by comparing it to two other common approaches in model selection – the computationally demanding best subset selection and the greedy forward selection. Finally, as a third application, we apply our method to real-world data collected from the CD95 pathway. All computations were performed using the open source software R [17], version 3.2.1 and associated packages fda [18] for the smoothing spline estimation and deSolve [19] for estimating ODEs.

3.1 Simulation 1: random networks and random catalysts

We use several simulation runs to test the general applicability of the proposed method. To that end, we consider networks consisting of 2 to 10 nodes and sampled

from

$$\dot{x}_i(t) = \sum_{g=1}^{N} \left(k_{ig} x_g(t) h_{ig}(t) - k_{gi} x_i(t) h_{gi}(t) \right), \tag{15}$$

where the reaction rates k_{ig} are chosen randomly from $\mathscr{U}(\frac{-1}{N}, \frac{1}{N})$ and the catalysts $h_{gi}(t)$ are chosen randomly to equal one of $(\mathbb{1}, x_1, \ldots, x_N)$ with equal probability, where $\mathbb{1}$ is a constant function equal to 1. Furthermore, the initial values $\boldsymbol{x}(0)$ are chosen randomly from $\mathscr{U}(1, 100)$. To achieve more realistic sparse networks, we randomly delete approximately a fraction of $\frac{2}{N}$ of the possible reactions by setting the corresponding reaction rates k_{ig} to 0. After forward simulation of the randomly chosen network, we add normally distributed measurement noise $\varepsilon \sim \mathscr{N}(0, \sigma^2)$ to the simulated time snapshots and arrive at the observed measurement points used for further analysis. The number of observed time points per component $\boldsymbol{x}_i(t)$ and the noise parameter σ^2 are also chosen at random for each simulation run from $\mathscr{U}(10, 30)$ and $\mathscr{U}(1, 15)$, respectively. In the described setting, we run 100 simulation runs per fixed network size and estimate the catalyst of each reaction. The results are shown in Fig. 1.

Fig. 1A shows violin plots of the fraction of models to be estimated after applying the latent catalyst method depending on the network size. Additionally, the average time needed to compute either all possible combinations for a given network size or the reduced set of models is shown with solid lines. Here, we observe that computing all possible combinations for networks of size 2 or 3 is faster than computing only a reduced number of models. This can be explained by the computational time needed to fit the splines as shown in (13) and the relatively low number of possible candidate models for such small network sizes. With increasing network size, this relationship switches very fast and already at network size 5 the computational time needed to identify the correct catalysts with our method is

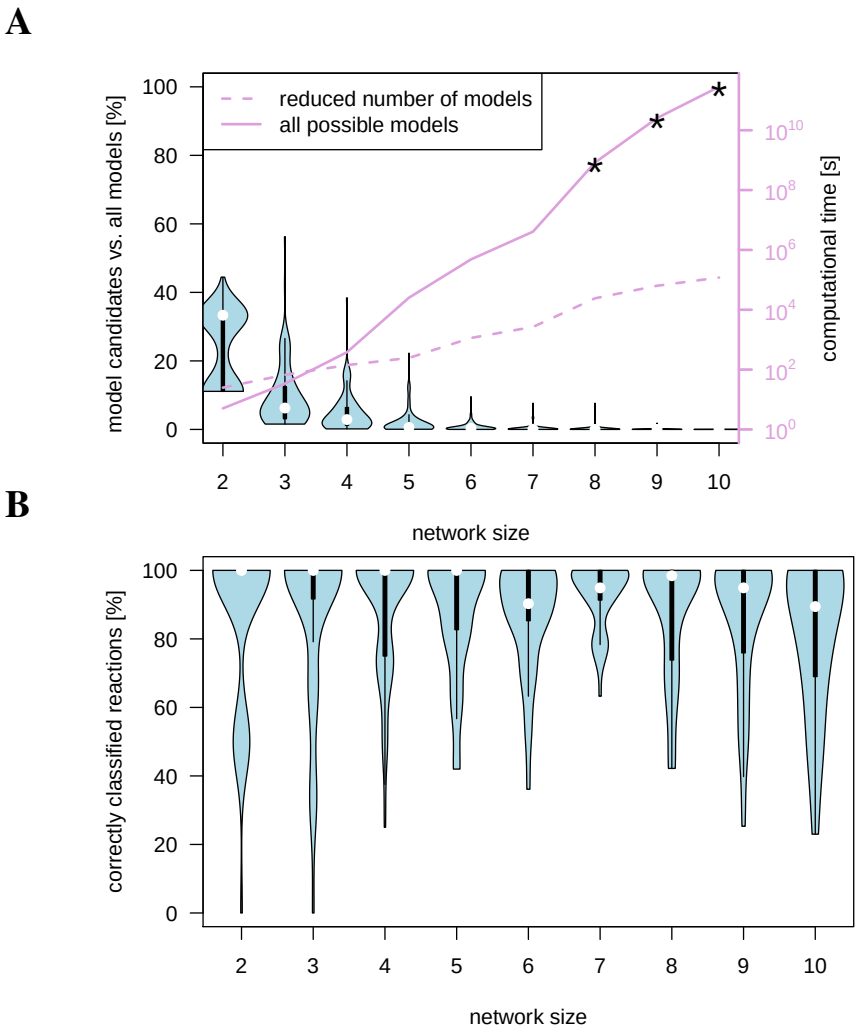

Figure 1: Results of simulation 1. **A**: Violin plots show fraction of model candidates chosen by application of the latent catalysts method compared to all possible models. The reduction of model candidates becomes more pronounced for larger networks. Additionally, lines show the average computational time in (log-scale) needed to estimate either all possible models (solid line) or the reduced set of model candidates (dashed line). Stars mark computational times that were estimated based on a fraction of completed model computations. **B**: Violin plots of the fraction of correctly classified catalysts. The reduced model candidates include the correct model that was used to generate the data in almost all simulation runs. This is consistent for all studied network sizes.

less than 0.1% of the time needed to compute all possible models. For networks larger than $N = 7$, we stopped computations after two weeks of parallel computation of 120 models and estimated the computational times based on the fraction of finished models. In the violin plots, a value of 100% means that no reduction of model candidates was achieved with our method and all possible models have to be computed. We observe a dramatic decrease of model candidates for networks consisting of more than 4 nodes. This shows the efficiency of our method, which potentially allows a reduction of computational time from days to minutes depending on the studied system.

This efficiency would not be meaningful if the reduced number of models did not include the correct model, which was used to generate the data. However, as Fig. 1B suggests, in most simulation runs the correct model is part of the reduced model candidates. More specifically, this is true on average for 85.1% of the simulation runs. This is consistent for all studied network sizes as can be observed in Table 1.

Table 1: Fraction of simulation runs for which the correct model is part of the reduced space of model candidates depending on network size N. Data is simulated based on (15).

network size N	2	3	4	5	6	7	8	9	10
correct classification [%]	92	91	86	91	80	85	86	78	77

Although the method may also miss the correct model in certain simulation scenarios with e. g. large noise or many similarly shaped component dynamics, we observe a median of above 90% correctly identified catalysts by applying our method. Additionally, we note that we used a threshold parameter $\tau = 0.1$ for all simulations. If we set this parameter to a higher value, we will capture more correct models in the model candidates, however at the cost of lower efficiency and higher computational demand.

3.2 Simulation 2: catalysis in common network motifs in systems biology

Fig. 2 shows artificial networks with and without catalytic interactions. This network consists of 5 nodes $x_1 - x_5$ and a total of 7 regulatory interactions between those nodes. In Fig. 2A, we first show a version of the network with no catalytic

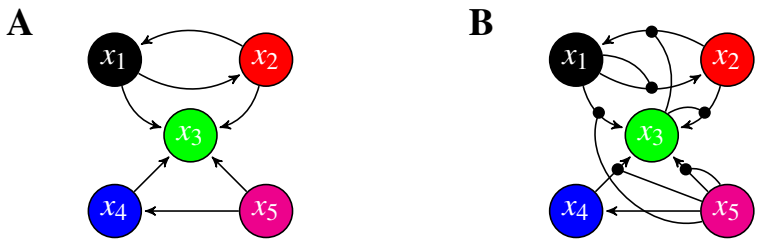

Figure 2: Network of interest for simulation 2. **A**: "core network" with no catalysis involved. **B**: network with catalysis from which data is sampled.

reactions in order to demonstrate the general connection between the nodes. In Fig. 2B, we include catalysis in the network structure and use this network to simulate artificial data. We chose this network to further investigate our method performance because it captures several patterns which are commonly observed in systems biology. First, x_1 and x_2 are engaged in a mutual activation pattern and whichever of the two dominates this pattern also dominates the flow to x_3. Second, a typical motif is presented by the conversion from x_5 to x_3 for which there is either a direct way and also an indirect or lagged reaction through x_4. Finally, we observe both layers to be connected by the key node x_3 and several catalytic connections which contribute to the overall pattern of the studied network.

We sampled data from the network shown in Fig. 2B by choosing random reaction rate parameters from $\mathscr{U}(-1,1)$, random initial values $x_l(0) \sim \mathscr{U}(0,100)$, $l = 1,\ldots,5$ at equidistant time points between $t_0 = 0$ and $t_n = 1$ with $t_{i+1} - t_i = 0.1$. We also added normally distributed measurement noise $\varepsilon \sim N(0,10)$ to each

simulated data point as shown in Fig. 3.

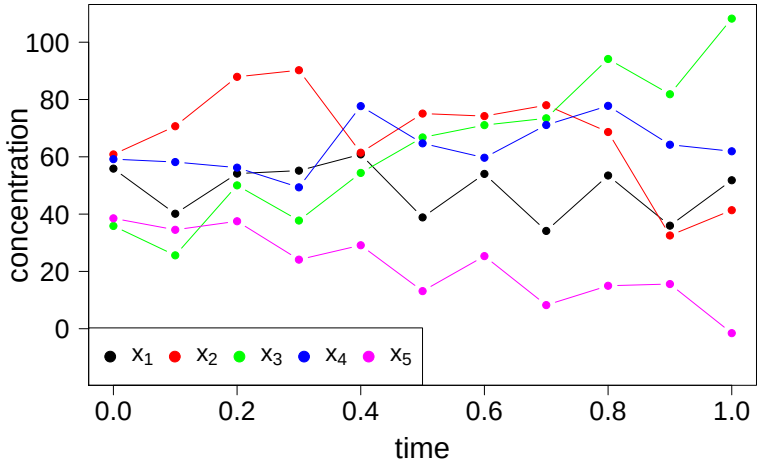

Figure 3: Simulated data from network shown in Fig. 2B and used for simulation study in section 3.2.

The next step in this simulation was to apply three techniques in order to estimate the correct catalyst for each reaction. First, we applied a very extensive search for the best model in which we fitted all possible models. In this case there are $6^7 \approx 300000$ different models (5 network components which may act as catalyst as well as no catalysis as a sixth possibility for each one of the 7 reactions). For each model, we calculated a loglikelihood function as described in (14). The models were then ordered by the loglikelihood values with the most appropriate model having the highest loglikelihood value. Results of selected models are shown in Table 2.

Second, we applied a greedy forward selection method. Here, the idea is to start from the null model with no catalysis in the network (Fig. 2A) and subsequently allow for one catalytic reaction after another. For the studied network, it means that we calculate the loglikelihood of only one model in the first step, then 35 models in the second step (we have 5 possible catalysts for 7 different reactions) with 35 corresponding loglikelihoods. Subsequently, we choose *one* catalyst for *one*

Table 2: Results of fitting all possible models for the network shown in Fig. 2. Here, we see the best six models and the worst model with respect to negative loglikelihood value. The model used to generate the data is highlighted in red.

rank	$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$	-(loglikelihood)
1	x_3	x_3	x_4	x_1	x_4	x_4	x_2	324.34
2	x_3	x_5	x_4	x_3	x_4	x_5	x_2	325.37
3	x_3	x_3	x_1	x_1	x_5	x_2	1	326.50
4	x_3	x_4	x_1	x_1	x_5	x_2	x_2	326.57
5	x_3	x_1	x_5	x_3	x_5	x_5	1	326.75
6	x_3	x_1	x_5	x_1	x_5	x_4	x_2	327.65
279936	1	1	1	x_2	x_1	1	x_5	1331.40

reaction corresponding to the model with the highest loglikelihood value and move on to the third step where another catalyst is selected from 30 different models in the same manner. We stop when the loglikelihood is not longer increased by a subsequent step. The results of this procedure are shown in Table 3.

Finally, we applied our method and selected model candidates with threshold $\tau = 0.1$. With our approach we select 288 model candidates and compute the corresponding loglikelihood. The component candidates for each model are shown in Table 4.

Table 3: Results of applying a forward model selection to the network shown in Fig. 2. The best model, corresponding to the highest loglikelihood value, is achieved in step 2. The data-generating ("true") model does not equal the chosen one.

steps	$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$	-(log likelihood)
step 0	1	1	1	1	1	1	1	436.57
step 1	x_3	1	1	1	1	1	1	394.10
step 2	x_3	1	1	1	x_5	1	1	386.92
step 3	x_3	1	1	1	x_5	1	x_2	411.28

Table 4: Results of application of our latent catalyst approach to the network shown in Fig. 2. Each reaction has a different number of possible components which may act as a catalyst. The data generating model is highlighted in red.

$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$
$\{x_3\}$	$\{x_1, x_3, x_4, x_5\}$	$\{x_1, x_4, x_5\}$	$\{x_1, x_3\}$	$\{x_4, x_5\}$	$\{x_2, x_4, x_5\}$	$\{1, x_2\}$

Application of the three different model selection techniques revealed different aspects. On the one hand, the forward selection is very fast due to the low number of models being fitted, however it fails in detecting a model that can fit the data

reasonably well. On the other hand, the best subset selection does not only find the correct model which was used to generate the data shown in Fig. 3 but it also finds four models which fit the data more appropriately. This can be explained by the fact that we added a high amount of measurement noise to the true ODE solutions and thus created data situations where the data generating model is not anymore the model that best fits the data. Nevertheless, we believe that this represents a scenario which is much more realistic for real-world applications than looking at the true ODE solutions as measurements where the data generating model will fit the data best by a large margin. The computational cost of this procedure is very high even for this medium-sized example as it runs a total of 1818.963 hours on a single core machine (faster with parallelisation). Finally, our approach with modelling latent catalysts also reveals the model that fits the data best'. This is achieved in a very efficient way by reducing the possible model candidates to 288, which is a reduction by 99.897%.

3.3 CD95 signalling model for apoptosis

In this section, we apply our method to real-world data collected from the cluster of differentiation 95 (CD95) signalling pathway [20]. This pathway is relevant for regulation of cell death decisions and is mediated via proteins FADD and procaspase-8 as well as its cleavage products p43/p41 and p18 (see Fig. 4). The pathway can be summarized in the following steps: after extracellular binding of the CD95 ligand to its receptor, FADD is recruited to CD95. This creates the death inducing signalling complex (DISC), and procaspase-8, c-FLIP long (c-$FLIP_L$) and c-FLIP short (c-$FLIP_S$) can bind to it. This results in the formation of different types of dimers: procaspase-8 homodimers (p8hod), procaspase-8 heterodimer (p8hed) and c-$FLIP_L$ heterodimers. Next, the procaspase-8 part of the dimers is split and is in its active form of p43 homodimer (p43hod) and p43 heterodimer

(p43hed). Finally, p43hod is processed to form the cleavage product p18. Next, procaspase-8 homo- and hetero- dimers undergo autocatalytic processing resulting in the formation of the p43 homodimer (p43hom) and p43 heterodimer (p43hod), respectively. The latter along with the cleavage product of procaspase-8, p43 comprises the cleavage product of c-FLIP, p43-FLIP. All of the steps described above have been reported in literature [20, 21, 22, 23]. The last three reactions denoted by the green box in Fig. 4 are known to be possibly catalysed [12]. Therefore, the focus of our work lies in the analysis of a small core motif containing 5 species and 3 reactions. The experimental data used in this manuscript provides measurements of the total p43, total p18 and total procaspase-8 concentration for two time-resolved experiments over a total of 6 time points each. The two experiments differ from each other in the amount of ligand used. We modelled both experiments separately thus obtaining two sets of results for the present data.

Our approach resulted in a very strong reduction of model candidates. Without our approach, a total of 216 models need to be computed and compared for each of the two experiments. With our approach, we are able to narrow down the number of model candidates to 3 and 12 for experiment 1 and 2, respectively. These models are presented in Table 5 and illustrated in Fig. 5. Here, we ranked the models by their corresponding negative loglikelihood. For both experiments, we have models that perform better than others in this measure (candidate 1 for experiment 1 and candidates 1–3 for experiment 2). However, we are not able to quantify whether the best-performing models in terms of loglikelihood are significantly better in explaining the data. The more interesting observation we make is the large difference of the number of model candidates which were identified with our method in both experiments. Intriguingly, when more ligand is present in the system (experiment 2), non-catalysed splitting of procaspase-8 heterodimer and non-catalysed processing of p43 homodimer are emerging as reactions contributing to the increase of model

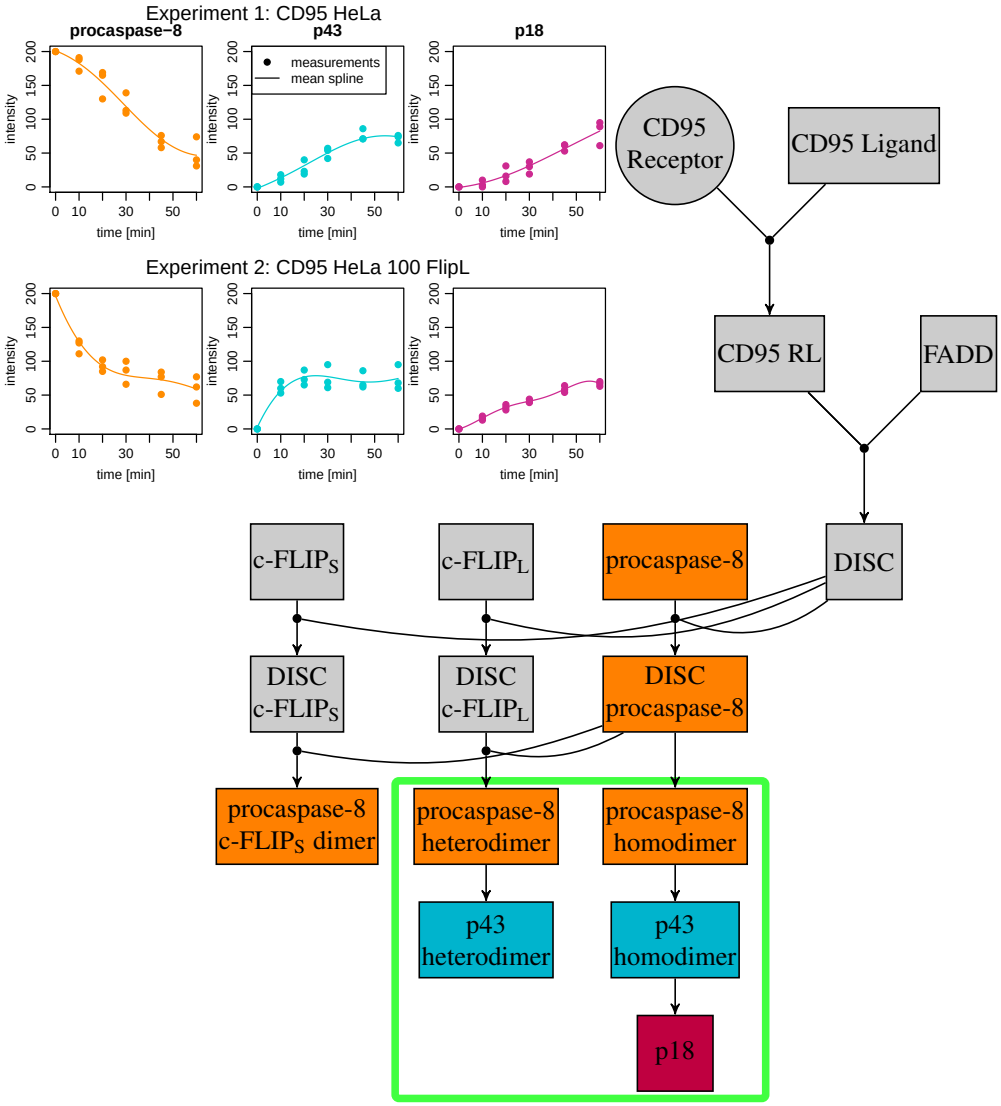

Figure 4: Data and schematic representation of CD 95 pathway. Dots show three replicates of two experiments at each time point of procaspase-8 (orange), p43 (blue) and p18 (red). Corresponding lines show the mean spline approximations. In the pathway, rectangles denote proteins, the extracellular receptor is denoted by a circle. Coloured rectangles indicate the different experimental measurements: total amount procaspase-8 (orange), total amount p43 (blue) and total amount p18 (red). The green box indicates the core motif, which is analysed by our method.

Table 5: Results of application of the latent catalyst approach to CD95 pathway. The three possibly catalysed reactions are shown on top of the table and the possible catalysts associated with the respective reactions are shown in the rows. All suggested model candidates are shown for experiment 1 (low amount of ligand) and experiment 2 (high amount of ligand). Models are ranked by their negative loglikelihood.

experiment 1				
	p8hed → p43hed	p8hod → p43hod	p43hod → p18	-(loglikelihood)
candidate 1	p8hed	1	p8hed	66.95
candidate 2	p8hed	p43hed	p8hed	71.05
candidate 3	p8hed	p43hod	p8hed	71.09
experiment 2				
	p8hed → p43hed	p8hod → p43hod	p43hod → p18	-(loglikelihood)
candidate 1	p8hed	p43hod	p8hed	66.53
candidate 2	p8hed	1	1	68.23
candidate 3	p8hed	p43hod	1	69.45
candidate 4	p8hed	1	p8hed	76.00
candidate 5	1	p43hed	1	78.06
candidate 6	p8hed	p43hed	1	78.42
candidate 7	1	p43hod	1	78.58
candidate 8	1	p43hed	p8hed	78.92
candidate 9	p8hed	p43hed	p8hed	83.42
candidate 10	1	1	1	84.67
candidate 11	1	1	p8hed	97.86
candidate 12	1	p43hod	p8hed	98.13

candidates. This is intuitively understandable because the more ligand is present at the beginning of the experiment, the more procaspase-8 and p43 will be produced and thus a catalysis appears less necessary in the system. The results further suggest four possible catalysts: procaspase-8 heterodimer as a possible catalyst of the splitting of procaspase-8 heterodimer (autocatalysis), procaspase-8 homodimer and p43 heterodimer as catalysts for the splitting of procaspase-8 homodimer and finally procaspase-8 heterodimer as catalyst for the processing of p43 homodimer. These results are in good agreement with previous analysis of the data [12], where three of the four proposed catalysts are suggested by the authors. The additional catalysed reaction (procaspase-8 heterodimer catalysing processing of p43 homodimer) was excluded prior to application of the proposed model reduction scheme. Additionally, we also see model candidates where the reactions are not catalysed especially for the experiment with high amount of ligand. We can therefore con-

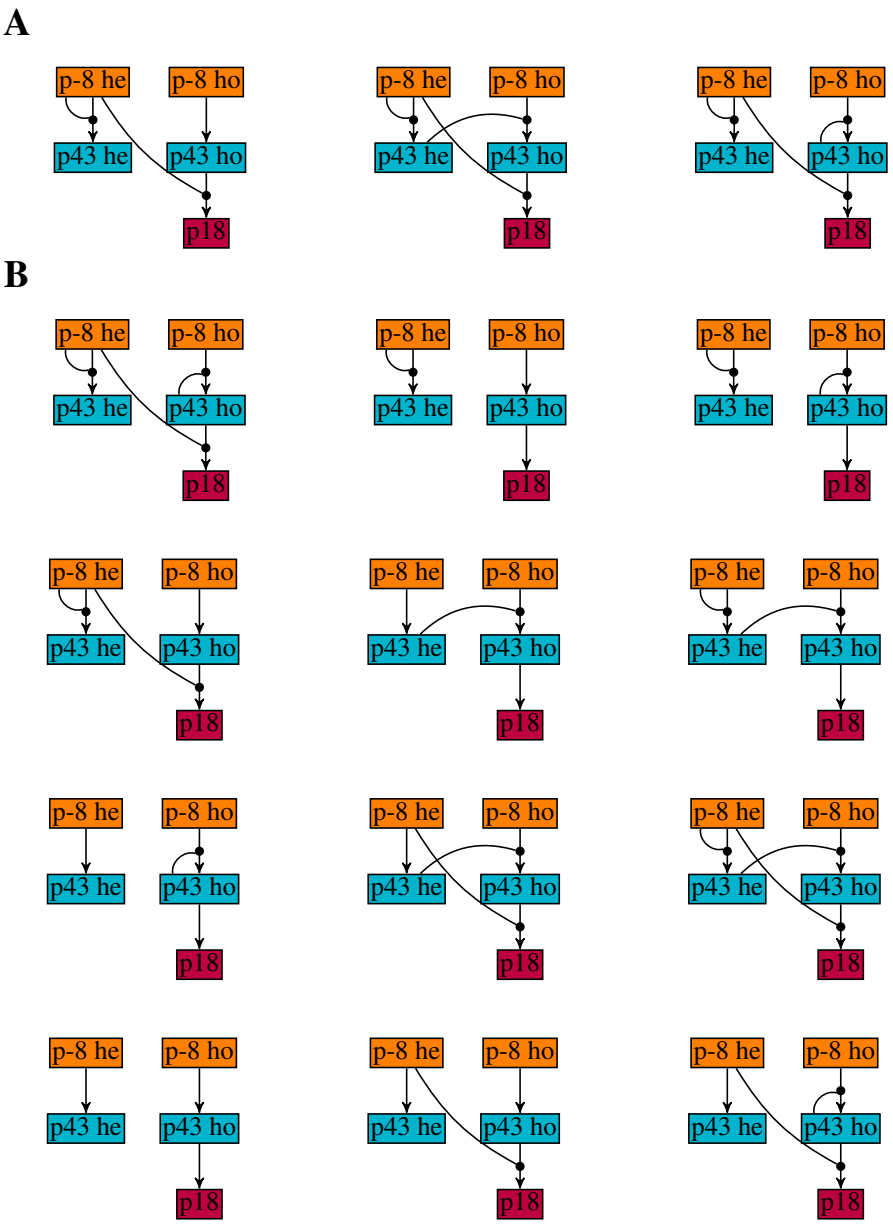

Figure 5: Results of the application of our approach to the CD95 apoptosis pathway. Each combination of five boxes denotes one candidate model for the core network motif highlighted with a green box in Fig. 4. This visualisation is in concordance with results presented in Table 5 and candidate models per experiment are sorted by likelihood from top left to bottom right. **A:** three model candidates for experiment 1. **B:** twelve model candidates for experiment 2.

clude that by adding more ligand at the beginning of the experiment catalytic reactions play only a minor role in the CD95 pathway. The presence of a low amount of ligand, however, enforces catalysis in the studied system. Thus, our approach indicates that low amount of stimuli can drive the cell into apoptosis. Upon overcoming the apoptotic threshold and upon high stimulation, the additional catalytic reactions are not required any more. These results hold true only if the assumptions (e. g. known stoichiometry of the system) made in the formulation of our method in Section 2 are valid. Furthermore, our results only provide a *hint* of the role that the amount of ligand might play in catalysis regulation in the CD95 pathway. Other biological impact factors which were not in the scope of our approach may additionally play an important role or even be more dominant for the explanation of the differences in the two experiments.

4 Discussion

Modelling of biological systems in computational systems biology is generally connected to high complexity. Especially if every reaction is considered to be of non-linear nature, as is true for catalytic reactions, the model candidate space significantly increases. Available approaches such as best subset model selection or stepwise selection schemes become infeasible for large biological systems due to the high amount of computational time involved. Additionally, as it is well known from regression-type models [24], an increase of the number of model candidates leads to an increase in the hazard of overfitting. This can be explained with the fact that if one considers possibly billions of model candidates the risk of choosing a model that follows the data too closely and at the same cannot generalize to other data is increasing and one then ends up fitting not only the data dynamics of interest but also the noise contained in the measurements.

We propose a novel and efficient method to incorporate catalysis into the modelling of biological systems. With our approach, we efficiently reduce the number of model candidates to a manageable number with a low risk of missing the most appropriate model for given data. We do this by extending the studied system by latent catalyst components. Next, we estimate those components by first approximating the available time courses with smoothing splines and subsequently exploiting the structure of the ODEs of the modelled system. Finally, we compare those estimates to all other components of the studied system and each comparison is associated with a score. This score can then be used to identify relevant components which may act as catalysts for a given reaction. Another byproduct of our approach is the automatic identification of model parameters during the estimation steps.

We studied the proposed method on several simulation settings and noted a substantial decrease of model candidates and at the same time we were able to recover the true models in almost all performed simulations. The application of our method to the CD95 apoptosis pathway confirmed previous results in literature and additionally identified different catalysts for some reactions. We could also conclude that the presence of ligand at the beginning in this system is a factor which seems to drive the importance of catalysis at later stages.

As we already stated in the introductory section of this manuscript, with our approach we look only at established reactions in well-studied biological systems and, if we identify the need of introducing catalysis, we alter only this specific reaction. Therefore, our approach is not comparable to other methods concerning the network topology [4, 14]. Furthermore, we assume both $\boldsymbol{S}$ and $\boldsymbol{v}$ to be known from literature. This assumption does not generally hold and therefore a combination of these network topology identification methods with our catalysis approach holds future research potential.

We classify our approach as a model selection method as we select only models which are based on reactions with high similarity scores for further investigation. Closely connected to model selection is the field of model reduction in biochemical networks [25, 26, 27, 28]. Here, the goal is to construct a minimal complex model which is still able to represent the data sufficiently. Contrarily, with our approach we do not aim at reducing the complexity of the model by e. g. dramatically altering its dimensionality or its topology but rather aim at a valid preselection of possible models which we investigate in more detail.

Although we tested our method in numerous simulations and also got interpretable and logical results out of the application example, other aspects of the approach pose challenging questions for future research. To begin with, we did not look into missing data or dependent measurement errors. Furthermore, sensitivity analysis and parameter identification analysis were not performed due to the estimation of parameters being only a byproduct of the whole method. Likewise, we restricted all analyses and applications to a linear catalysis in the form of $x(t) \cdot h(t)$. This can, however, be extended in future work to more general, non-linear settings in the form of $f(x_i(t), h_g(t))$ with a possibly non-linear function f. Next, we require a known network structure to apply our method on. We can imagine combining our approach with a network or motif identification and thus make it completely automatic.

We assume that all possible catalysts are already *part of the network*. This assumption can be relaxed and we could allow for *external catalysts*, which were previously not part of the modelled species. We studied this network extension in another manuscript [13] and aim at combining both methods in the future.

Overall, we are confident that our method is a valuable tool in practice which can be used to gain additional knowledge out of time-resolved measured data and allows for different conclusions regarding catalysis. Based on such findings, we

expect additional hypotheses for future research to be generated and thus lead to a better understanding of biochemical models.

Acknowledgements

The research leading to these results has received funding from the European Research Council under grant agreement number 259294 (Starting grant Latent Causes) and by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, subproject A17. The authors are grateful to Inna Lavrik and Nicolai Fricker for providing the CD95 data for analysis and helping with result interpretation as well as to Atefeh Kazeroonian and Dennis Rickert for critical comments and helpful discussions.

References

[1] Patrick Aloy and Robert B Russell. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, 7(3):188–197, 2006.

[2] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[3] D Fernández Slezak, Cecilia Suárez, Guillermo A Cecchi, Guillermo Marshall, and Gustavo Stolovitzky. When the optimal is not the best: parameter estimation in complex biological models. *PLoS ONE*, 5(10):e13283, 2010.

[4] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[5] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[6] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie I Stefan, et al. Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4(1):92, 2010.

[7] Spyros Artavanis-Tsakonas, Matthew D Rand, and Robert J Lake. Notch signaling: cell fate control and signal integration in development. *Science*, 284(5415):770–776, 1999.

[8] Viola Vogel and Michael P Sheetz. Cell fate regulation by coupling me-

chanical cycles to biochemical signaling pathways. *Current Opinion in Cell Biology*, 21(1):38–46, 2009.

[9] Sean Gallagher et al. Immunoblotting and immunodetection. *Current protocols in cell biology*, pages 6–2, 2011.

[10] David Gilbert, Hendrik Fuß, Xu Gu, Richard Orton, Steve Robinson, Vladislav Vyshemirsky, Mary Jo Kurth, C Stephen Downes, and Werner Dubitzky. Computational methodologies for modelling, analysis and simulation of signalling networks. *Briefings in Bioinformatics*, 7(4):339–353, 2006.

[11] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. Model assessment and selection. In *The Elements of Statistical Learning*, pages 193–224. Springer, 2001.

[12] Dennis Rickert, Nicolai Fricker, Inna N Lavrik, and Fabian J Theis. Systematic complexity reduction of signaling models and application to a CD95 signaling model for apoptosis. In *Systems Biology of Apoptosis*, pages 57–84. Springer, 2013.

[13] Ivan Kondofersky, Christiane Fuchs, and Fabian J Theis. Identifying latent dynamic components in biological systems. *IET Systems Biology*, 9(5):193–203, October 2015.

[14] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC systems biology*, 1(1):1, 2007.

[15] Jin Zhou and Jun-an Lu. Topology identification of weighted complex dynamical networks. *Physica A: Statistical Mechanics and Its Applications*, 386(1):481–491, 2007.

[16] Richard I Masel et al. *Chemical kinetics and catalysis*. Wiley-Interscience New York, 2001.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[18] J. O. Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2009. R package version 2.1.2.

[19] KER Soetaert, Thomas Petzoldt, and R Woodrow Setzer. Solving differential equations in R: package deSolve. *Journal of Statistical Software*, 33, 2010.

[20] Inna N Lavrik, Alexander Golks, Dagmar Riess, Martin Bentele, Roland Eils, and Peter H Krammer. Analysis of CD95 threshold signaling triggering of CD95 (FAS/APO-1) at low concentrations primarily results in survival signaling. *Journal of Biological Chemistry*, 282(18):13664–13671, 2007.

[21] FC Kischkel, S Hellbardt, Iris Behrmann, M Germer, M Pawlita, PH Krammer, and ME Peter. Cytotoxicity-dependent APO-1 (Fas/CD95)-associated proteins form a death-inducing signaling complex (DISC) with the receptor. *The EMBO journal*, 14(22):5579, 1995.

[22] Nicolai Fricker, Joel Beaudouin, Petra Richter, Roland Eils, Peter H Krammer, and Inna N Lavrik. Model-based dissection of CD95 signaling dynamics reveals both a pro-and antiapoptotic role of c-FLIPL. *The Journal of Cell Biology*, 190(3):377–389, 2010.

[23] Leo Neumann, Carina Pforr, Joel Beaudouin, Alexander Pappa, Nicolai Fricker, Peter H Krammer, Inna N Lavrik, and Roland Eils. Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. *Molecular Systems Biology*, 6(1):352, 2010.

[24] Michael A Babyak. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, 2004.

[25] Sune Danø, Mads F Madsen, Henning Schmidt, and Gunnar Cedersund. Reduction of a biochemical model with preservation of its basic dynamic properties. *Febs Journal*, 273(21):4862–4877, 2006.

[26] M Mahfouf, DA Linkens, and D Xue. A new generic approach to model reduction for complex physiologically based drug models. *Control Engineering Practice*, 10(1):67–81, 2002.

[27] Ovidiu Radulescu, Alexander N Gorban, Andrei Zinovyev, and Vincent Noel. Reduction of dynamical biochemical reaction networks in computational biology. *arXiv preprint arXiv:1205.2851*, 2012.

[28] Nishith Vora and Prodromos Daoutidis. Nonlinear model reduction of chemical reaction systems. *AIChE Journal*, 47(10):2320–2332, 2001.