The Journal of Chemical Physics

# Reactive SINDy: Discovering governing reactions from concentration data

Cite as: J. Chem. Phys. **150**, 025101 (2019); https://doi.org/10.1063/1.5066099
Submitted: 12 October 2018 . Accepted: 16 December 2018 . Published Online: 08 January 2019

Moritz Hoffmann, Christoph Fröhner, and Frank Noé

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

# Reactive SINDy: Discovering governing reactions from concentration data

Cite as: J. Chem. Phys. 150, 025101 (2019); doi: 10.1063/1.5066099
Submitted: 12 October 2018 • Accepted: 16 December 2018 •
Published Online: 8 January 2019



Moritz Hoffmann,[a),b)] Christoph Fröhner,[b),c)] and Frank Noé[d)]

## AFFILIATIONS

Freie Universität Berlin, Fachbereich Mathematik und Informatik, Arnimallee 6, 14195 Berlin, Germany

[a)]**Electronic mail:** moritz.hoffmann@fu-berlin.de
[b)]M. Hoffmann and C. Fröhner contributed equally to this work.
[c)]**Electronic mail:** christoph.froehner@fu-berlin.de
[d)]**Author to whom correspondence should be addressed:** frank.noe@fu-berlin.de

## ABSTRACT

The inner workings of a biological cell or a chemical reactor can be rationalized by the network of reactions, whose structure reveals the most important functional mechanisms. For complex systems, these reaction networks are not known *a priori* and cannot be efficiently computed with *ab initio* methods; therefore, an important goal is to estimate effective reaction networks from observations, such as time series of the main species. Reaction networks estimated with standard machine learning techniques such as least-squares regression may fit the observations but will typically contain spurious reactions. Here we extend the sparse identification of nonlinear dynamics (SINDy) method to vector-valued ansatz functions, each describing a particular reaction process. The resulting sparse tensor regression method "reactive SINDy" is able to estimate a parsimonious reaction network. We illustrate that a gene regulation network can be correctly estimated from observed time series.



## I. INTRODUCTION

Mapping out the reaction networks behind biological processes, such as gene regulation in cancer,[1] is paramount to understanding the mechanisms of life and disease. A well-known example of gene regulation is the lactose operon whose crystal structure was resolved in Ref. 26 and dynamics were modeled in Ref. 49. The system's "combinatorial control" in *E. coli* cells was quantitatively investigated in Ref. 24, in particular, studying repression and activation effects. These gene regulatory effects often appear in complex networks,[39] and there exist databases resolving these for certain types of cells, e.g., *E. coli* cells[11] and yeast cells.[25] Another example where mapping the active reactions is important is that of chemical reactors,[33] where understanding which reactions are accessible for a given set of educts and reaction conditions is important to design synthesis pathways.[7,21]

The traditional approach to determine a reaction network is to propose the structure of the network based on chemical insight and subsequently fit the parameters given available data.[36] To decipher complex reaction environments such as biological cells, it would be desirable to have a data-driven approach that can answer the question which reactions are underlying a given observation, e.g., the time series of a set of reactants. However, in sufficiently complex reaction environments, the number of reactive species and possible reactions is practically unlimited—as an illustration, consider the vast amount of possible isomerizations and post-translational modifications for a single protein molecule. Therefore, the more specific formulation is "given observations of a set of chemical species, what is the *minimal* set of reactions necessary to explain their time evolution?" This formulation calls for a machine learning method that can infer the reaction network underlying the observation data.

Knowledge about the reaction network can be applied to parameterize other numerical methods to further investigate the processes at hand. Such methods include particle-based approaches derived from the chemical master

equation,[13,19,46,47] as well as highly detailed but parameter-rich methods such as particle-based or interacting-particle reaction dynamics[2,8,10,17,37,44,45] capable of fully resolving molecule positions in space and time; see Refs. 3 and 38 for recent reviews.

Existing methods to infer regulatory networks include ARACNE[28] that uses experimental essay data and information theory, as well as the likelihood approach presented in Ref. 42 that takes the stochasticity of observed reactant time series into account.

The method presented in this work can identify underlying complex reaction networks from concentration time series by following the law of parsimony, i.e., by inducing sparsity in the resulting reaction network. This promotes the interpretability of the model and avoids overfitting. We formulate the problem as data-driven identification of a dynamical system, which renders the method consistent with and an extension of the framework of sparse identification of nonlinear dynamics (SINDy).[5] Specifically, the problem of identifying a reaction network from time traces of reactant concentrations can be solved by finding a linear combination from a library of candidate nonlinear functions (ansatz functions) that each corresponds to a reaction acting on a set of reactants. With this formulation, the reaction rates can be determined via regression. Sparsity is induced by equipping the regression algorithms with a sparsity inducing regularization. SINDy was investigated, generalized, and applied in many different ways, e.g., including control[6] (SINDYc), in the context of partial differential equations,[34] updating already existing models[32] (abrupt-SINDy), and looking into convergence properties.[50]

We extend and apply SINDy to the case of learning reaction networks from non-equilibrium concentration data. Similar approaches make use of SINDy but do not resolve specific reactions,[27] use weak formulations to avoid numerical temporal derivatives,[31] or use compressive sensing and sparse Bayesian learning.[30]

Our extension of the original SINDy method mostly involves estimating parameters which are coupled across the equations of the arising dynamical system. In the context of learning reaction networks, this means that we look for specific reactions and their rate constants that might have led to the observations instead of net flux across species. We demonstrate the algorithm on a gene regulatory network in three different scenarios of measurement: When there is no noise in the data, we can find, given sufficient amounts of data, all relevant processes of the ground truth. If there is noise in the data, we converge to the correct reaction network and rates with decreasing levels of noise. The third scenario generalizes the method to two measurements with different initial conditions, also converging to the correct model with decreasing levels of noise.

We additionally demonstrate the algorithm on time series data of the mitogen activated protein kinases (MAPK) pathway as an example for a bimodal system and on time series data of the Lotka–Volterra system which describes oscillatory predator-prey dynamics subject to social friction. In both systems, reactive SINDy recovers the generating reaction network, whereas non-sparse estimation detects many spurious processes.

## II. REACTIVE SINDY: SPARSE LEARNING OF REACTION KINETICS

We are observing the concentrations of S chemical species in time $t$,

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_S(t) \end{pmatrix} \in \mathbb{R}^S. \tag{1}$$

We assume that their dynamics are governed by classical reaction-rate equations subject to the law of mass action. A general expression for the change in concentration of reactant $s$ as a result of order-0 reactions (creation), order-1 reactions (transitions of other species into $s$, transitions of $s$ into other species, or annihilation), order-2 reactions (production or consumption of $s$ by the encounter of two species), etc, is given by

$$\dot{x}_s = \sum_i \beta^{(i)}_{s,0} + \sum_i \beta^{(i)}_{s,1} x_i + \sum_{i,j} \beta^{(i,j)}_{s,2} x_i x_j + \dots, \tag{2}$$

where the $\beta^{(\dots)}_{s,k}$-values are constants belonging to the reactions of order $k$. These rate constants however can incorporate several underlying reactions at once. For example, the two reactions

$$s_1 \xrightarrow{\xi_1} s_2, \tag{3}$$

$$s_1 \xrightarrow{\xi_2} s_3 \tag{4}$$

both contribute to $\dot{x}_1 = \beta^{(i)}_{1,1} x_1 = -(\xi_1 + \xi_2) x_1$. To disentangle (2) into single reactions, we choose a library of $R$ possible ansatz reactions that each represent a single reaction,

$$\mathbf{y}_r(\mathbf{x}(t)) = \begin{pmatrix} y_{r,1}(\mathbf{x}(t)) \\ \vdots \\ y_{r,S}(\mathbf{x}(t)) \end{pmatrix}, \quad r = 1, \dots, R. \tag{5}$$

With this ansatz, the reaction dynamics (2) becomes a set of linear equations with unknown parameters $\xi_r$ that represent the sought macroscopic rate constants,

$$\dot{\mathbf{x}}_i(t) = \sum_{r=1}^{R} y_{r,i}(\mathbf{x}(t)) \xi_r, \quad i = 1, \dots, S, \tag{6}$$

where $\xi_r$ are the to-be estimated macroscopic rate constants. The two reactions in the previous examples (3) and (4) would be modeled by the functions

$$y_1(\mathbf{x}) = (-x_1, x_1, 0)^\top,$$
$$y_2(\mathbf{x}) = (-x_1, 0, x_1)^\top,$$

illustrating that the values of the coefficients $\xi_1$ and $\xi_2$ can be used to decide whether a single reaction is present and to what degree.

Now suppose that we have measured the concentration vector (1) at $T$ time points $t_1 < \cdots < t_T$. We represent these data as a matrix

$$\mathbf{X} = \left(\mathbf{x}(t_1)\mathbf{x}(t_2)\cdots\mathbf{x}(t_T)\right)^\top \in \mathbb{R}^{T\times S}. \tag{7}$$

Given this matrix, a library $\Theta : \mathbb{R}^{T\times S} \to \mathbb{R}^{T\times S\times R}$, $\mathbf{X} \mapsto \left(\theta_1(\mathbf{X})\theta_2(\mathbf{X})\cdots\theta_R(\mathbf{X})\right)$ of $R$ candidate (ansatz) reactions can be proposed with corresponding reaction functions,

$$\theta_r(\mathbf{X}) = \begin{pmatrix} \mathbf{y}_r(\mathbf{X}_{1*})^\top \\ \vdots \\ \mathbf{y}_r(\mathbf{X}_{T*})^\top \end{pmatrix} \in \mathbb{R}^{T\times S}, \quad r = 1,\ldots,R, \tag{8}$$

where $\mathbf{X}_{i*}$ denotes the $i$th row in $\mathbf{X}$. Applying the concentration trajectory to the library yields $\Theta(\mathbf{X}) \in \mathbb{R}^{T\times S\times R}$.

The goal is to find coefficients $\Xi = \left(\xi_1\,\xi_2\cdots\xi_R\right)^\top$ so that

$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi = \sum_{r=1}^{R} \theta_r(\mathbf{X})\xi_r. \tag{9}$$

In particular, the system is linear in the coefficients $\Xi$, which makes regression tools such as elastic net regularization[52] applicable. To this end, one can consider the regularized minimization problem (reactive SINDy),

$$\hat{\Xi} = \underset{\Xi}{\operatorname{argmin}}\left(\frac{1}{2T}\left\|\dot{\mathbf{X}} - \Theta(\mathbf{X})\Xi\right\|_F^2 + \alpha\lambda\|\Xi\|_1 + \alpha(1-\lambda)\|\Xi\|_2^2\right) \quad \text{subject to } \Xi \geq 0. \tag{10}$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm, $\lambda \in [0,1]$ is a hyperparameter that interpolates linearly between the least absolute shrinkage and selection operator (LASSO)[15,43] and Ridge[16] methods, and $\alpha \geq 0$ is a hyperparameter that, depending on $\lambda$, can induce sparsity and give preference to smaller solutions in the $L_1$ or $L_2$ sense. For $\alpha = 0$, the minimization problem reduces to standard least-squares (LSQ) with the constraint $\Xi \geq 0$. Reactive SINDy (10) is therefore a generalization of the SINDy method to vector-valued ansatz functions.

Since only the concentration data $\mathbf{X}$ is available but not its temporal derivative, $\dot{\mathbf{X}}$ is approximated numerically by second order finite differences with the exception of boundary data. Once the pair $(\mathbf{X}, \dot{\mathbf{X}})$ is obtained, the problem becomes invariant under temporal reordering. Hence, when presented with multiple trajectories, the data matrices $\mathbf{X}_i$ and $\dot{\mathbf{X}}_i$ can simply be concatenated.

In order to solve (10), the numerical sequential least-squares minimizer SLSQP[23] is applied via the software package SciPy.[20] Code related to this paper can be found under https://github.com/readdy/readdy_learn.

## III. RESULTS

We demonstrate the method by estimating the reactions of a gene-regulatory network from time series of concentrations of the involved molecules. Let $S := \{A, B, C\}$ be a set of three species of proteins which are being translated each from their respective mRNA molecule. Each mRNA in turn has a corresponding DNA which it is transcribed from. The proteins and mRNA molecules decay over time, whereas the DNA concentration remains constant. The network contains reactions of the following form:[41]

$$\begin{aligned} \mathrm{DNA}_i &\to \mathrm{DNA}_i + \mathrm{mRNA}_i && \text{(transcription)},\\ \mathrm{mRNA}_i &\to \mathrm{mRNA}_i + i && \text{(translation)},\\ \mathrm{mRNA}_i &\to \emptyset && \text{(decay of mRNA)},\\ i &\to \emptyset && \text{(decay of protein)},\\ i + \mathrm{mRNA}_j &\to i && \text{(regulation of } j \in S), \end{aligned}$$

for each of the species $i \in S$. These reactions model a regulation of species $j$ by virtue of the fact that the transcription product inhibits the transcription processes. In our example, proteins of type A regulate the $\mathrm{mRNA_B}$ molecules, proteins of type B regulate the $\mathrm{mRNA_C}$ molecules, and proteins of type C regulate the $\mathrm{mRNA_A}$ molecules (Fig. 1). Using this reaction model, time series of concentrations are generated using the rates given in Table II under the initial condition described in Table I(a), which were chosen so that all the reactions in the reaction model significantly contribute to the temporal evolution of the system's concentrations. The generation samples the integrated equations equidistantly with a discrete time step of $\tau = 3\cdot10^{-3}$ yielding 667 frames which amounts to a cumulative time of roughly $T = 2$.

The proposed estimation method is applied to analyze these time series of concentrations in order to recover the underlying reaction network from data. To this end, we use

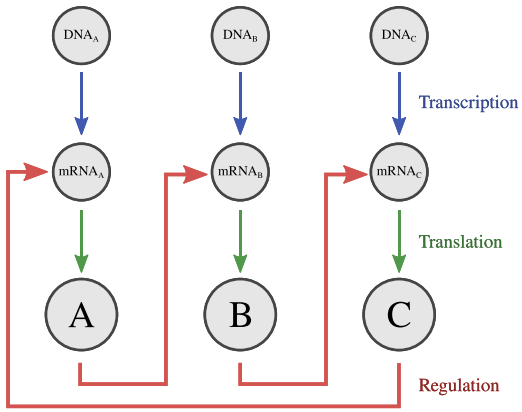


**FIG. 1.** The regulation network example described in Sec. III. Each circle depicts a species, and each arrow corresponds to one reaction. Blue arrows denote transcription from DNA to mRNA, green arrows denote translation from mRNA to protein, and red arrows denote the regulatory network.

**TABLE I**. Initial conditions (a) and (b) used to generate concentration time series. Reaction rates can be found in Table II.

| | $DNA_A$ | $mRNA_A$ | A | $DNA_B$ | $mRNA_B$ | B | $DNA_C$ | $mRNA_C$ | C |
|---|---|---|---|---|---|---|---|---|---|
| (a) | 1 | 2 | 0 | 1 | 0 | 3 | 1 | 0 | 0 |
| (b) | 1 | 1.5 | 0 | 1 | 0 | 2 | 1 | 0 | 1 |

the library of ansatz functions given in Table II, which contains a large number of possible reactions, only few of which are actually part of the model.

### A. Learning the reaction network in the low-noise limit

We first demonstrate that the true reaction network can be reconstructed when using a finite amount of observation data without additional measurement noise, i.e., the observations are reflecting the true molecule concentrations at any given time point. The minimization problem (10) is solved using the concentration time series shown in Fig. 1(b).

We first set the hyperparameter $\alpha = 0$ in the minimization problem (10), which results in constrained least-squares regression without any of the regularization terms. In this case, we estimate a reaction network that can reproduce the observations almost exactly (Fig. 2). However, the result is mechanistically wrong as the sparsity pattern does not match the reaction network used to generate the data. On the one hand, many spurious reactions are estimated that were not in the true reaction scheme and would lead to wrong conclusions about the mechanism, such as A + A ⇀ A and A + C ⇀ C. More dramatically, the reaction responsible for the decay of A particles is completely ignored (Fig. 3).

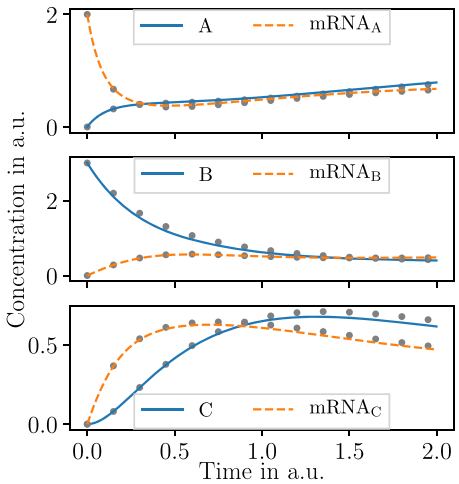


**FIG. 2**. Concentration time series generated from integrating the reaction network shown in Fig. 1(a). The initial condition prescribes positive concentration values only for B protein and $mRNA_A$ species [Table I(a)]. This initial condition is used in Secs. III A – III C for further analysis. Gray dots depict concentration time series yielded from the LSQ rates estimated in Sec. III A.

Next, we sought sparse solutions by using $\alpha > 0$ and additionally eliminating reactions with rate constants smaller than a cutoff value $\kappa$. For a suitable choice of hyperparameters $\alpha \approx 1.91 \cdot 10^{-7}$, $\lambda = 1$, and $\kappa = 0.22$, a sparse solution is obtained that finds the correct reaction scheme and also recovers the decay reaction (Fig. 3).

The value of the cutoff $\kappa$ was determined by comparing the magnitude of estimated rates and finding a gap; see Fig. 8. The hyperparameter pair $(\alpha, \lambda)$ was obtained by a grid search and evaluating the difference $\|\hat{\Xi}_{\alpha,\lambda} - \Xi\|_1$, where $\hat{\Xi}_{\alpha,\lambda}$ is the estimated model under a particular hyperparameter choice and $\Xi$ is the ground truth. If the ground truth is unknown, a hyperparameter pair can be estimated by utilizing cross-validation as in Secs. III B–III E.

### B. Learning the reaction network from data with stochastic noise

In contrast to Sec. III A, we now employ data that include measurement noise. Such noise can originate from uncertainties in the experimental setup or from shot noise in single- or few-molecule measurements. In gene regulatory networks, such noise is commonly observed when only few copy numbers of mRNA are present.[4,9,14] In order to simulate noise from few copies of molecules, the system of Sec. III with initial conditions as given in Table I(a) is integrated using the Gillespie stochastic simulation algorithm (SSA).[12,13] In the limit of many particles and realizations, the Gillespie SSA converges to the integrated reaction-rate equations subject to the law of mass action. As our model is based on exactly these dynamics, the initial condition's concentrations are interpreted in terms of hundreds of particles. Each realization is then transformed back to a time series of concentrations. We define the noise level as the mean-squared deviation of the concentration time series from the integrated reaction-rate equations. Data with different noise levels are prepared by averaging multiple realizations of the time series obtained by the Gillespie SSA.

It can be observed that decreasing levels of noise lead to fewer spurious reactions when applying reactive SINDy (10); see Fig. 4(a). Also the estimation error $\|\xi - \hat{\xi}\|_1$ with respect to the ground truth $\xi$ decreases with decreasing levels of noise [Fig. 4(b)]. In both cases, the regularized method with a suitable hyperparameter pair $(\alpha, \lambda)$ performs better than LSQ.

The hyperparameters $(\alpha, \lambda)$ are obtained by shuffling the data and performing a 10-fold cross validation.

### C. Learning the reaction network from multiple initial conditions

Preparing the experiment that generates the data in different initial conditions can help identifying the true reaction

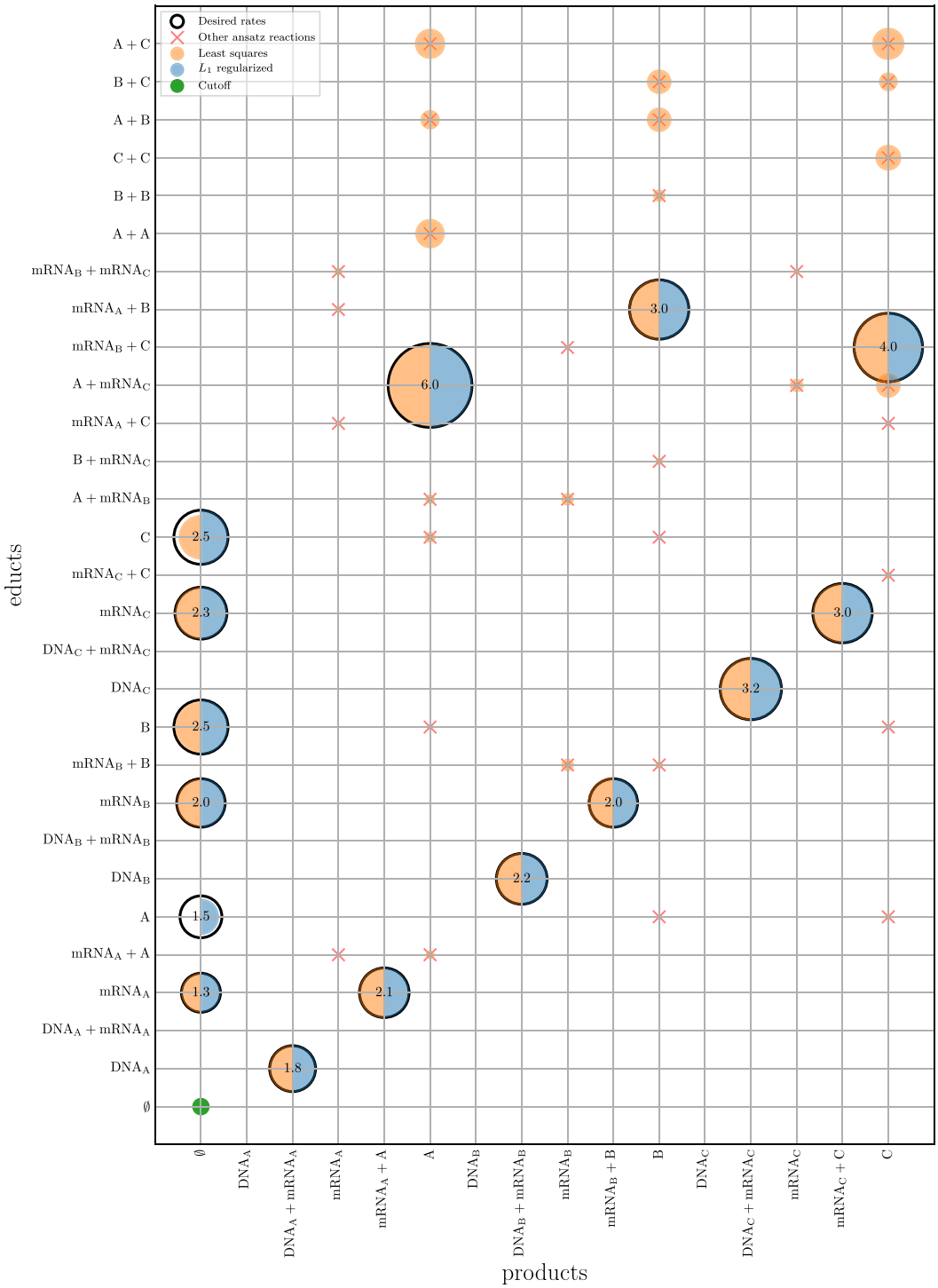


**FIG. 3**. Estimated reaction rates in the system described in Sec. III A. The y and x axes contain reaction educts and products, respectively. A circle at position $(i, j)$ represents a reaction $i \rightharpoonup j$ whose rate has a linear relation with the area of the circle. The black outlines denote the reactions with which the system was generated and contain the respective rate value. Red crosses denote reactions that were used as additional ansatz reactions. Blue circles are estimated by LSQ, and orange circles depict rates which were obtained by solving the minimization problem (10). The latter rates are subject to a cutoff $\kappa = 0.22$ corresponding to the green circle's area under which a sparse solution with the correct processes can be recovered. If a certain rate was estimated in both cases, two wedges instead of one circle are displayed.

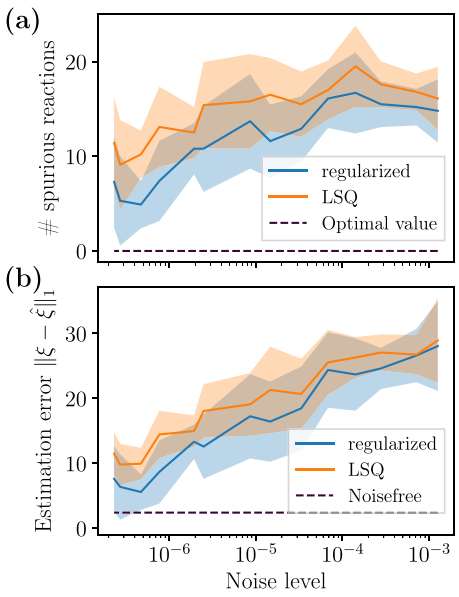


**FIG. 4**. Convergence of the estimation error when estimating the system described in Sec. III A with varying levels of noise by application of reactive SINDy (10) with and without regularization in blue and orange, respectively. The procedure was independently repeated 10 times with different realizations giving rise to the mean and standard deviation depicted by solid lines and shaded areas, respectively. (a) The number of detected spurious reactions up to the cutoff value introduced in Sec. III A over different levels of noise. (b) The estimation error given by the mean absolute error between the generating reaction rates $\xi$ and the estimated reaction rates $\hat{\xi}$ over different levels of noise.

mechanisms as a more diverse dataset makes it easier to confirm or exclude the participation of specific reactions. This section extends the analysis of Sec. III B to two initial conditions, where the first initial condition is identical to the one used previously and the second initial condition is given in Table I(b).

The corresponding time series are depicted in Fig. 5(a). The gray graph corresponds to a sample trajectory generated by the Gillespie SSA. For both initial conditions, the same time step of $\tau = 3{\cdot}10^{-3}$ has been applied, amounting to $2{\cdot}667 = 1334$ frames. Once the data matrices

$$\mathbf{X}_1 = \left(\mathbf{x}_1(t_1)\cdots\mathbf{x}_1(t_{667})\right),\ \mathbf{X}_2 = \left(\mathbf{x}_2(t_1)\cdots\mathbf{x}_2(t_{667})\right)$$

and the corresponding derivatives $\dot{\mathbf{X}}_1$, $\dot{\mathbf{X}}_2$ have been obtained, the frames are concatenated so that

$$\mathbf{X} = \left(\mathbf{x}_1(t_1)\cdots\mathbf{x}_1(t_{667})\mathbf{x}_2(t_1)\cdots\mathbf{x}_2(t_{667})\right),$$

analogously for $\dot{\mathbf{X}}$.

Similarly to Sec. III B, decreasing levels of noise lead to fewer spurious reactions [Fig. 5(b)] and a smaller $L_1$ distance to the ground truth [Fig. 5(c)]. Again applying the optimization

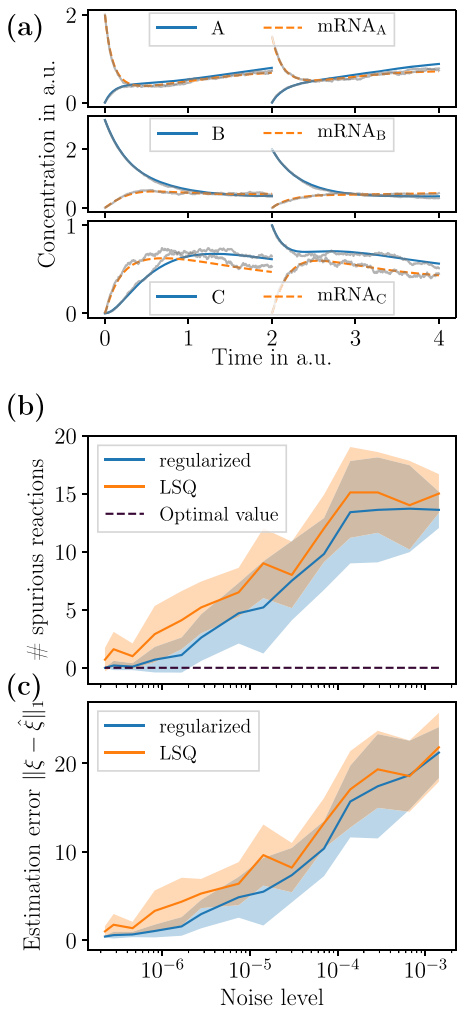


**FIG. 5**. Convergence of estimation error of reaction schemes from noisy gene-regulation data starting from two different initial conditions under decreasing levels of noise. The minimization problem (10) was solved for $\alpha = 0$ (LSQ) and with regularization. This was repeated 10 times on different sets of observation data generated by Gillespie SSA, giving rise to mean and standard deviation (solid lines and shaded areas, respectively). (a) Concentration time series corresponding to the initial conditions, generated by integrating the reaction-rate equations. The first initial condition is identical to the one used in Secs. III A and III B. The second initial condition [Table I(b)] prescribes positive initial concentrations for $\mathrm{mRNA_A}$, B, and C species. The gray graphs are sample realizations of integration using the Gillespie SSA. [(b) and (c)] Analogous to Fig. 4 with the difference that 20-fold cross validation was used for hyperparameter estimation.

problem with a suitable set of parameters $(\alpha, \lambda, \kappa)$ performs better than LSQ. Compared to Sec. III B, the convergence is better due to twice as much available data. At noise levels of smaller than roughly $10^{-6}$, the model can reliably be recovered when using the regularized method.

The hyperparameters $(\alpha, \lambda)$ are obtained by shuffling the data and performing a 20-fold cross validation.

### D. Application to MAPK cascade

The reactive SINDy method is applied to the mitogen activated protein kinases (MAPK) pathway[48] which is an important regulatory mechanism of biological cells to respond to stimuli and is involved in proliferation, differentiation, inflammation, and apoptosis.[51] Single-cell MAPK kinetics can be observed experimentally.[35] Mathematically MAPK kinetics are often modelled using reaction rate equations[22,29] which enables analysis using reactive SINDy.

Generally a MAPK pathway consists of multiple stages of kinases that are either inactive or active, denoted by "$*$". Their activation occurs due to phosphorylation catalyzed by the upstream kinase of the previous stage, and dephosphorylation is catalyzed by phosphatases. When the kinase is active, it can activate other downstream kinases of the next stage. The initial activation is often due to an external stimulus. The response of the whole cascade is the amount of activated substrate after the final stage, typically measured as a function of the initial stimulus.

Here the MAPK pathway is modeled with three stages of kinases MAPK, MAPKK, and MAPKKK. The initial stimulus is called S, and the final substrate to be activated is a transcription factor TF. The ground truth reaction network consists of activation/phosphorylation reactions,

$$\begin{aligned}
\mathrm{S} + \mathrm{MAPKKK} &\rightharpoonup \mathrm{S} + \mathrm{MAPKKK}*,\\
\mathrm{MAPKKK}* + \mathrm{MAPKK} &\rightharpoonup \mathrm{MAPKKK}* + \mathrm{MAPKK}*,\\
\mathrm{MAPKK}* + \mathrm{MAPK} &\rightharpoonup \mathrm{MAPKK}* + \mathrm{MAPK}*,\\
\mathrm{MAPK}* + \mathrm{TF} &\rightharpoonup \mathrm{MAPK}* + \mathrm{TF}*,
\end{aligned}$$

and deactivation/dephosphorylation reactions

$$\begin{aligned}
\mathrm{MAPKKK}* &\rightharpoonup \mathrm{MAPKKK},\\
\mathrm{MAPKK}* &\rightharpoonup \mathrm{MAPKK},\\
\mathrm{MAPK}* &\rightharpoonup \mathrm{MAPK},\\
\mathrm{TF}* &\rightharpoonup \mathrm{TF}.
\end{aligned}$$

For simplicity, we assume phosphatase to be abundant such that deactivations effectively become first order reactions. The external stimulus S is not consumed such that time integration of these reactions yields a steady state in which the response, i.e., the concentration [TF$*$] can be measured as a function of the stimulus concentration [S]. Using the rate constants given in Table III, we obtain the response curve given in Fig. 6(a).

We generate concentration time series data of the MAPK reactions above at three different initial conditions, each differing in the amount of stimulus [S]. The response yielded by the chosen initial conditions is marked in Fig. 6(a) by vertical dashed lines. The concatenated time series is a dataset of 300 frames in total. We use the library $\Theta$ of ansatz reactions, Table III. The hyperparameter $\alpha = 6.6 \times 10^{-9}$ was determined by shuffling the data and performing 15-fold cross validation. The estimated rate constants were obtained by solving the minimization problem (10) with $\lambda = 1$. The results are given in Fig. 6(b). Least-squares estimation detects 5 of the 8 reaction processes that belong to the ground truth model. However it also detects 12 spurious reaction processes ($\theta_{18}$–$\theta_{29}$). Reactive SINDy estimation detects all reactions of the ground truth, and two processes ($\theta_4$ and $\theta_8$) show deviations in rate constants. Generally reactive SINDy yields a sparse model which allows further simplification of the reaction network by dropping out reaction processes that lie beneath a certain cutoff. In this case, for example, a cutoff of $\kappa = 0.25$ would directly recover the ground truth reaction network. Quantitatively, one may consider the $L_1$ norm of the relative distance of estimated rate constants $\hat{\xi}_r$ to the non-zero rate constants of the ground truth $\xi_r$,

$$\sum_{r=1}^{8} \left| (\hat{\xi}_r - \xi_r)/\xi_r \right|,$$

which yields 167% error for least-squares and 21% error for reactive SINDy.

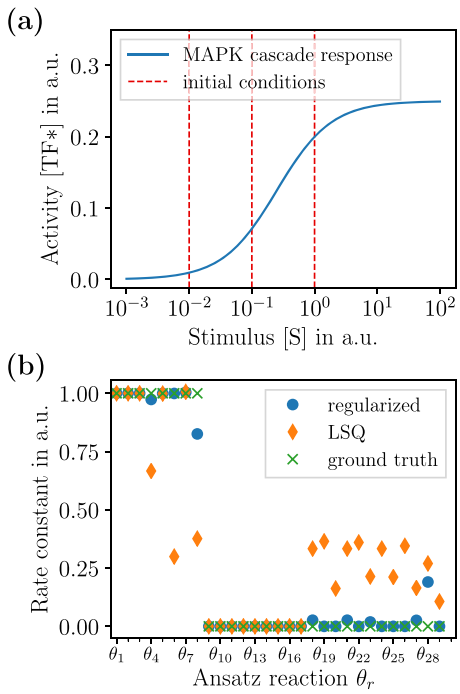


**FIG. 6.** Application of reactive SINDy to the MAPK pathway system. (a) The response curve of the MAPK cascade as a function of external stimulus given as a constant concentration [S]. The activity is the steady state concentration of activated transcription factors [TF$*$]. Dashed lines show the values of [S] at which concentration time series data were generated. (b) Estimated rate coefficients of candidate reactions (see Table III) after application of reactive SINDy (regularized) to the time series data. Least-squares estimation (LSQ) and the ground truth model for comparison.

### E. Application to Lotka–Volterra system

As biological pathways often exhibit oscillatory behavior[18] which can stem from positive or negative feedback loops,[40] we apply reactive SINDy to an idealized oscillatory system, namely, the Lotka–Volterra system. The predator-prey dynamics of two species A (prey) and B (predator) is defined by the reaction network

$$\begin{aligned} A &\rightharpoonup A + A && \text{(prey growth)},\\ A + B &\rightharpoonup B + B && \text{(predator eats prey)},\\ B &\rightharpoonup \emptyset && \text{(predator decay)},\\ A + A &\rightharpoonup \emptyset && \text{(prey friction)},\\ B + B &\rightharpoonup \emptyset && \text{(predator friction)}. \end{aligned}$$

From this model, we generated concentration time series data with 200 frames which is displayed in Fig. 7(a). The library of ansatz reactions $\Theta$ is given in Table IV. The hyperparameter $\alpha = 2.7 \times 10^{-7}$ was determined by shuffling the data and performing 5-fold cross validation. The estimated rate constants were obtained by solving the minimization problem (10) with $\lambda = 1$. The results are depicted in Fig. 7(b). Least-squares estimation detects all reactions of the ground truth model but also two spurious processes ($\theta_6$ and $\theta_7$) with a higher rate than the first two underlying processes ($\theta_1$ and $\theta_2$). Reactive SINDy recovers the true reaction network with minor deviations in rate constants. As in Sec. III D, considering the $L_1$ norm of the relative distance to the ground truth for non-zero rate constants

$$\sum_{r=1}^{5} \left| (\hat{\xi}_r - \xi_r)/\xi_r \right|$$

yields 75% error for least-squares and 7% error for reactive SINDy.

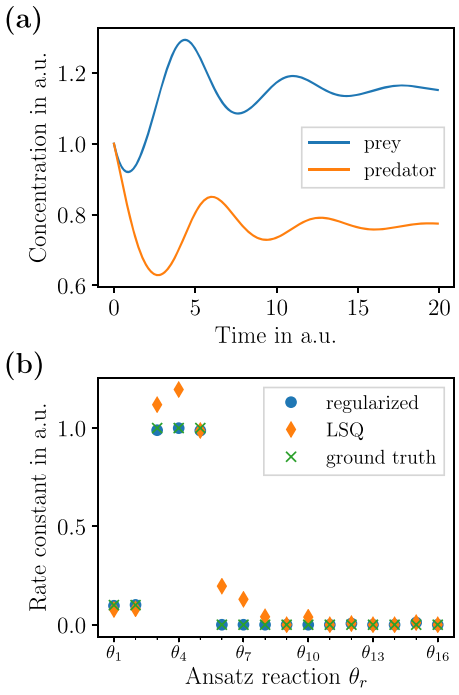


**FIG. 7**. Application of reactive SINDy to the Lotka–Volterra system with social friction. (a) Concentration data as a function of time for predator and prey species. (b) Estimated rate coefficients of candidate reactions (see Table IV) after application of reactive SINDy (regularized) to the time series data. Least-squares (LSQ) estimation and the ground truth model for comparison.

## IV. CONCLUSION

In this work, we have extended the SINDy method to reactive SINDy, not only parsimoniously detecting potentially nonlinear terms in a dynamical system from noisy data, but also yielding, in this case, a sparse set of rates with respect to generating reactions (8). Mathematically this has been achieved by permitting vector-valued basis functions and obtaining a tensor linear regression problem. We have applied this method on data generated from a gene regulation network, a MAPK pathway, and a Lotka–Volterra system and could successfully recover the underlying reaction networks.

The studies of Secs. III B and III C have shown that the applied regularization terms can mitigate noise up to a certain degree compared to the unregularized method so that identification of the reaction network is more robust and closer to the ground truth. Potentially, this method could be used to identify reaction networks from time series measurements even if the initial conditions are not always exactly identical, as was demonstrated in Sec. III C.

One apparent limitation is that the method can only be applied if the data stem from the equilibration phase, as the concentration-based approach has derivatives equal to zero in the equilibrium, which precludes the reaction dynamics to be recovered. Thus, in the case of oscillatory systems, the reaction network can be recovered robustly.

In future work, we will consider the identification of reaction schemes from instantaneous fluctuations of particle numbers in equilibrium.

## ACKNOWLEDGMENTS

The authors are grateful to the Center for Theoretical Biological Physics (CTBP, supported by NSF Grant No. PHY-1427654) at Rice University for hosting their sabbatical visit, during which part of this work was performed.

We gratefully acknowledge funding from Deutsche Forschungsgemeinschaft (SFB 958/Project A04, TRR 186/Project A12, SFB 1114/Project C03), the Einstein Foundation Berlin (ECMath Project CH17), and the European Research Council (ERC CoG 772230 "ScaleCell"). We are grateful for inspiring discussions with Simon Olsson, Mohsen Sadeghi, Felix Höfling, and Christof Schütte.

## APPENDIX: ANSATZ REACTIONS AND CUTOFF

**TABLE II**. Full set of ansatz reactions Θ used in Sec. III for the gene-regulatory network. The given rate constants define the ground truth reaction model.

| Reaction | Rate | Description |
|---|---|---|
| $DNA_A \rightarrow DNA_A + mRNA_A$ | $k_1 = 1.8$ | Transcription of $mRNA_A$ |
| $mRNA_A \rightarrow mRNA_A + A$ | $k_2 = 2.1$ | Translation of A proteins |
| $mRNA_A \rightarrow \emptyset$ | $k_3 = 1.3$ | $mRNA_A$ decay |
| $A \rightarrow \emptyset$ | $k_4 = 1.5$ | Decay of A proteins |
| $DNA_B \rightarrow DNA_B + mRNA_B$ | $k_5 = 2.2$ | Transcription of $mRNA_B$ |
| $mRNA_B \rightarrow mRNA_B + B$ | $k_6 = 2.0$ | Translation of B proteins |
| $mRNA_B \rightarrow \emptyset$ | $k_7 = 2.0$ | $mRNA_B$ decay |
| $B \rightarrow \emptyset$ | $k_8 = 2.5$ | Decay of B proteins |
| $DNA_C \rightarrow DNA_C + mRNA_C$ | $k_9 = 3.2$ | Transcription of $mRNA_C$ |
| $mRNA_C \rightarrow mRNA_C + C$ | $k_{10} = 3.0$ | Translation of C proteins |
| $mRNA_C \rightarrow \emptyset$ | $k_{11} = 2.3$ | $mRNA_C$ decay |
| $C \rightarrow \emptyset$ | $k_{12} = 2.5$ | Decay of C proteins |
| $mRNA_A + A \rightarrow A$ | $k_{13} = 0$ | Self-regulation of A proteins |
| $mRNA_B + B \rightarrow B$ | $k_{14} = 0$ | Self-regulation of B proteins |
| $mRNA_C + C \rightarrow C$ | $k_{15} = 0$ | Self-regulation of C proteins |
| $mRNA_B + A \rightarrow A$ | $k_{16} = 0$ | Regulation of $mRNA_B$ |
| $mRNA_C + B \rightarrow B$ | $k_{17} = 0$ | Regulation of $mRNA_C$ |
| $mRNA_A + C \rightarrow C$ | $k_{18} = 0$ | Regulation of $mRNA_A$ |
| $mRNA_C + A \rightarrow A$ | $k_{16} = 6.0$ | Regulation of $mRNA_C$ |
| $mRNA_B + C \rightarrow C$ | $k_{17} = 4.0$ | Regulation of $mRNA_B$ |
| $mRNA_A + B \rightarrow B$ | $k_{18} = 3.0$ | Regulation of $mRNA_A$ |
| $mRNA_A + A \rightarrow mRNA_A$ | $k_{19} = 0$ | Artificial fusion |
| $mRNA_B + B \rightarrow mRNA_B$ | $k_{20} = 0$ | Artificial fusion |
| $mRNA_A + B \rightarrow mRNA_A$ | $k_{21} = 0$ | Artificial fusion |
| $mRNA_B + C \rightarrow mRNA_B$ | $k_{22} = 0$ | Artificial fusion |
| $mRNA_C + A \rightarrow mRNA_C$ | $k_{23} = 0$ | Artificial fusion |
| $mRNA_A + C \rightarrow mRNA_A$ | $k_{24} = 0$ | Artificial fusion |
| $mRNA_B + A \rightarrow mRNA_B$ | $k_{25} = 0$ | Artificial fusion |
| $A + A \rightarrow A$ | $k_{26} = 0$ | A regulates A |
| $B + B \rightarrow B$ | $k_{27} = 0$ | B regulates B |
| $C + C \rightarrow C$ | $k_{28} = 0$ | C regulates C |
| $B + A \rightarrow A$ | $k_{29} = 0$ | Artificial fusion |
| $C + B \rightarrow B$ | $k_{30} = 0$ | Artificial fusion |
| $A + C \rightarrow C$ | $k_{31} = 0$ | Artificial fusion |
| $C + A \rightarrow A$ | $k_{32} = 0$ | Artificial fusion |
| $B + C \rightarrow C$ | $k_{33} = 0$ | Artificial fusion |
| $A + B \rightarrow B$ | $k_{34} = 0$ | Artificial fusion |
| $A \rightarrow B$ | $k_{35} = 0$ | Artificial conversion |
| $B \rightarrow C$ | $k_{36} = 0$ | Artificial conversion |
| $C \rightarrow A$ | $k_{37} = 0$ | Artificial conversion |
| $A \rightarrow C$ | $k_{38} = 0$ | Artificial conversion |
| $C \rightarrow B$ | $k_{39} = 0$ | Artificial conversion |
| $B \rightarrow A$ | $k_{40} = 0$ | Artificial conversion |
| $mRNA_B + mRNA_C \rightarrow mRNA_A$ | $k_{41} = 0$ | Artificial fusion |
| $mRNA_C + mRNA_B \rightarrow mRNA_C$ | $k_{42} = 0$ | Artificial fusion |
| $mRNA_C + A \rightarrow C$ | $k_{43} = 0$ | Artificial fusion |

**TABLE III**. Full set of ansatz reactions Θ used in Sec. III D for the MAPK system. The given rate constants define the ground truth reaction model.

| Reaction | Rate | Description |
|---|---|---|
| S + MAPKKK ⟶ S + MAPKKK∗ | $k_1 = 1$ | External stimulus activates MAPKKK |
| MAPKKK∗ ⟶ MAPKKK | $k_2 = 1$ | Dephosphorylation |
| MAPKKK∗ + MAPKK ⟶ MAPKKK∗ + MAPKK∗ | $k_3 = 1$ | Phosphorylation of MAPKK |
| MAPKK∗ ⟶ MAPKK | $k_4 = 1$ | Dephosphorylation |
| MAPKK∗ + MAPK ⟶ MAPKK∗ + MAPK∗ | $k_5 = 1$ | Phosphorylation of MAPK |
| MAPK∗ ⟶ MAPK | $k_6 = 1$ | Dephosphorylation |
| MAPK∗ + TF ⟶ MAPK∗ + TF∗ | $k_7 = 1$ | Phosphorylation of the transcription factor |
| TF∗ ⟶ TF | $k_8 = 1$ | Dephosphorylation |
| MAPKKK + MAPKK ⟶ MAPKKK + MAPKK∗ | $k_9 = 0$ | Artificial reaction |
| MAPKKK + MAPK ⟶ MAPK∗ | $k_{10} = 0$ | Artificial reaction |
| MAPKKK + TF ⟶ MAPKKK + TF∗ | $k_{11} = 0$ | Artificial reaction |
| MAPKKK∗ + MAPK ⟶ MAPKKK∗ + MAPK∗ | $k_{12} = 0$ | Artificial reaction |
| MAPKKK∗ + TF ⟶ MAPKKK∗ + TF∗ | $k_{13} = 0$ | Artificial reaction |
| MAPKK + TF ⟶ MAPKK + TF∗ | $k_{14} = 0$ | Artificial reaction |
| MAPKK∗ + TF ⟶ MAPKK∗ + TF∗ | $k_{15} = 0$ | Artificial reaction |
| MAPK + TF ⟶ MAPK + TF∗ | $k_{16} = 0$ | Artificial reaction |
| MAPKK + MAPK ⟶ MAPKK + MAPK∗ | $k_{17} = 0$ | Artificial reaction |
| MAPKKK + MAPKK∗ ⟶ MAPKKK + MAPKK | $k_{18} = 0$ | Artificial reaction |
| MAPKKK + MAPK∗ ⟶ MAPKKK + MAPK | $k_{19} = 0$ | Artificial reaction |
| MAPKKK + TF∗ ⟶ MAPKKK + TF | $k_{20} = 0$ | Artificial reaction |
| MAPKKK∗ + MAPKK∗ ⟶ MAPKKK∗ + MAPKK | $k_{21} = 0$ | Artificial reaction |
| MAPKKK∗ + MAPK∗ ⟶ MAPKKK∗ + MAPK | $k_{22} = 0$ | Artificial reaction |
| MAPKKK∗ + TF∗ ⟶ MAPKKK∗ + TF | $k_{23} = 0$ | Artificial reaction |
| MAPKK + MAPK∗ ⟶ MAPKK + MAPK | $k_{24} = 0$ | Artificial reaction |
| MAPKK + TF∗ ⟶ MAPKK + TF | $k_{25} = 0$ | Artificial reaction |
| MAPKK∗ + MAPK∗ ⟶ MAPKK∗ + MAPK | $k_{26} = 0$ | Artificial reaction |
| MAPKK∗ + TF∗ ⟶ MAPKK∗ + TF | $k_{27} = 0$ | Artificial reaction |
| MAPK + TF∗ ⟶ MAPK + TF | $k_{28} = 0$ | Artificial reaction |
| MAPK∗ + TF∗ ⟶ MAPK∗ + TF | $k_{29} = 0$ | Artificial reaction |

**TABLE IV**. Full set of ansatz reactions Θ used in Sec. III E for the Lotka–Volterra system. The given rate constants define the ground truth reaction model.

| Reaction | Rate | Description |
|---|---|---|
| A + A ⟶ ∅ | $k_1 = 0.1$ | Social friction of the prey |
| B + B ⟶ ∅ | $k_2 = 0.1$ | Social friction of the predator |
| A ⟶ A + A | $k_3 = 1$ | Prey growth |
| A + B ⟶ B + B | $k_4 = 1$ | Predator eats the prey |
| B ⟶ ∅ | $k_5 = 1$ | Predator decays |
| A + B ⟶ A + A | $k_6 = 0$ | Artificial reaction |
| A ⟶ ∅ | $k_7 = 0$ | Artificial reaction |
| B + B ⟶ B | $k_8 = 0$ | Artificial reaction |
| B ⟶ B + B | $k_9 = 0$ | Artificial reaction |
| A + A ⟶ A | $k_{10} = 0$ | Artificial reaction |
| A + B ⟶ A | $k_{11} = 0$ | Artificial reaction |
| A + B ⟶ B | $k_{12} = 0$ | Artificial reaction |
| A + A ⟶ B | $k_{13} = 0$ | Artificial reaction |
| A ⟶ B | $k_{14} = 0$ | Artificial reaction |
| B ⟶ A | $k_{15} = 0$ | Artificial reaction |
| A ⟶ B + B | $k_{16} = 0$ | Artificial reaction |

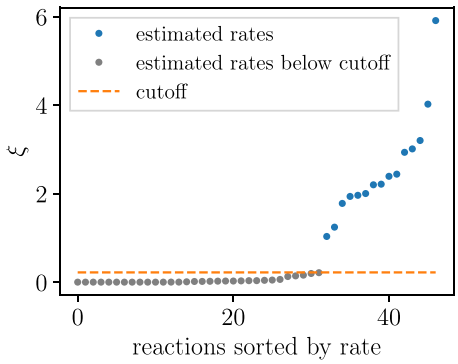


**FIG. 8**. Reaction rates sorted by their magnitude to determine the cutoff $\kappa = 0.22$ of Sec. III A. The rates were estimated using the regularized minimization problem.

## REFERENCES

[1] Abramovitch, R., Tavor, E., Jacob-Hirsch, J., Zeira, E., Amariglio, N., Pappo, O., Rechavi, G., Galun, E., and Honigman, A., "A pivotal role of cyclic AMP-responsive element binding protein in tumor progression," Cancer Res. **64**(4), 1338–1346 (2004).

[2]Andrews, S. S., "Smoldyn: Particle-based simulation with rule-based modeling, improved molecular interaction and a library interface," Bioinformatics **33**(5), 710–717 (2017).
[3]Andrews, S. S., *Particle-Based Stochastic Simulators* (Springer, New York, NY, 2018), pp. 1–5.
[4]Berg, O. G., "A model for the statistical fluctuations of proteins numbers in a microbial population," J. Theor. Biol. **71**(4), 587–603 (1978).
[5]Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Natl. Acad. Sci. U. S. A. **113**, 3932–3937 (2016).
[6]Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Sparse identification of nonlinear dynamics with control (SINDYc)," IFAC-PapersOnLine **49**(18), 710–715 (2016).
[7]Cong, P., Doolen, R. D., Fan, Q., Giaquinta, D. M., Guan, S., McFarland, E. W., Damodara, M. P., Self, K., Turner, H. W., and Henry Weinberg, W., "High-throughput synthesis and screening of combinatorial heterogeneous catalyst libraries," Angew. Chem., Int. Ed. **38**(4), 483–488 (1999).
[8]Donev, A., Yang, C.-Y., and Kim, C., "Efficient reactive brownian dynamics," J. Chem. Phys. **148**(3), 034103 (2018).
[9]Elowitz, M. B., "Stochastic gene expression in a single cell," Science **297**(5584), 1183–1186 (2002).
[10]Fröhner, C. and Noé, F., "Reversible interacting-particle reaction dynamics," J. Phys. Chem. B **122**, 11240 (2018).
[11]Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Moral-Chávez, V. D., Rinaldi, F., and Collado-Vides, J., "RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond," Nucleic Acids Res. **44**(D1), D133–D143 (2016).
[12]Gillespie, D. T., "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," J. Comput. Phys. **22**(4), 403–434 (1976).
[13]Gillespie, D. T., "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem. **81**(25), 2340–2361 (1977).
[14]Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C., "Real-time kinetics of gene activity in individual bacteria," Cell **123**(6), 1025–1036 (2005).
[15]Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer New York, New York, NY, 2009).
[16]Hoerl, A. E. and Kennard, R. W., "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics **12**, 55–67 (1970).
[17]Hoffmann, M., Fröhner, C., and Noé, F., "ReaDDy 2: Fast and flexible software framework for interacting-particle reaction dynamics," preprint bioRxiv:374942 (2018).
[18]J. Iqbal and M. Zaidi, "TNF-induced MAP kinase activation oscillates in time," Biochem. Biophys. Res. Commun. **371**(4), 906–911 (2008).
[19]Isaacson, S. A., "The reaction-diffusion master equation as an asymptotic approximation of diffusion to a small target," SIAM J. Appl. Math. **70**(1), 77–111 (2009).
[20]Jones, E., Oliphant, T., Peterson, P. *et al.*, SciPy: Open Source Scientific Tools for Python, 2001, Online accessed 27 October 2017.
[21]Kiwi-Minsker, L. and Albert, R., "Microstructured reactors for catalytic reactions," Catal. Today **110**(1-2), 2–14 (2005).
[22]Kolch, W., Calder, M., and Gilbert, D., "When kinases meet mathematics: The systems biology of MAPK signalling," FEBS Lett. **579**(8), 1891–1895 (2005).
[23]Kraft, D., "A software package for sequential quadratic programming," Technical Report DFVLR-FB 88-28, Institut für Dynamik der Flugsysteme, Oberpfaffenhofen, 1988.
[24]Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T., "Combinatorial transcriptional control of the lactose operon of Escherichia coli," Proc. Natl. Acad. Sci. U. S. A. **104**(14), 6043–6048 (2007).
[25]Lee, T. I., "Transcriptional regulatory networks in Saccharomyces cerevisiae," Science **298**(5594), 799–804 (2002).
[26]Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., and Lu, P., "Crystal structure of the lactose operon repressor and its complexes with DNA and inducer," Science **271**(5253), 1247–1254 (1996).
[27]Mangan, N. M., Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Inferring biological networks by sparse identification of nonlinear dynamics," IEEE Trans. Mol., Biol. Multi-Scale Commun. **2**(1), 52–63 (2016).
[28]Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A., "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," BMC Bioinf. **7**(Suppl 1), S7 (2006).
[29]Orton, J. R., Sturm, O. E., Vyshemirsky, V., Calder, M., Gilbert, D. R., and Walter, K., "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway," Biochem. J. **392**(2), 249–261 (2005).
[30]Pan, W., Yuan, Y., Goncalves, J., and Stan, G.-b., "Reconstruction of arbitrary biochemical reaction networks: A compressive sensing approach," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* (IEEE, 2012), pp. 2334–2339.
[31]Pantazis, Y. and Tsamardinos, I., "A unified approach for sparse dynamical system inference from temporal measurements," preprint arXiv:1710.00718 (2017).
[32]Quade, M., Abel, M., Kutz, J. N., and Brunton, S. L., "Sparse identification of nonlinear dynamics for rapid model recovery," Chaos **28**(6), 063116 (2018).
[33]Roa, R., Kim, W. K., Kanduč, M., Dzubiella, J., and Angioletti-Uberti, S., "Catalyzed bimolecular reactions in responsive nanoreactors," ACS Catal. **7**(9), 5604–5611 (2017).
[34]Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Data-driven discovery of partial differential equations," Sci. Adv. **3**(4), e1602614 (2017).
[35]Ryu, H., Chung, M., Song, J., Lee, S. S., Pertz, O., and Jeon, N. L., "Integrated platform for monitoring single-cell MAPK kinetics in computer-controlled temporal stimulations," Sci. Rep. **8**(1), 11126 (2018).
[36]Schöneberg, J., Heck, M., Hofmann, K. P., and Frank, N., "Explicit spatiotemporal simulation of receptor-g protein coupling in rod cell disk membranes," Biophys. J. **107**(5), 1042–1053 (2014).
[37]Schöneberg, J. and Noé, F., "ReaDDy-a software for particle-based reaction-diffusion dynamics in crowded cellular environments," PLoS One **8**(9), e74261 (2013).
[38]Schöneberg, J., Ullrich, A., and Noé, F., "Simulation tools for particle-based reaction-diffusion dynamics in continuous space," BMC Biophys. **7**(1), 11 (2014).
[39]Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U., "Network motifs in the transcriptional regulation network of *Escherichia coli*," Nat. Genet. **31**(1), 64–68 (2002).
[40]Shin, S.-Y., Rath, O., Choo, S.-M., Fee, F., McFerran, B., Kolch, W., and Cho, K.-H., "Positive-and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway," J. Cell Sci. **122**(3), 425–435 (2009).
[41]Thattai, M. and van Oudenaarden, A., "Intrinsic noise in gene regulatory networks," Proc. Natl. Acad. Sci. U. S. A. **98**(15), 8614–8619 (2001).
[42]Tian, T., Xu, S., Gao, J., and Burrage, K., "Simulated maximum likelihood method for estimating kinetic rates in gene expression," Bioinformatics **23**(1), 84–91 (2007).
[43]Tibshirani, R., "Regression selection and shrinkage via the lasso," J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996).
[44]van Zon, J. S. and Ten Wolde, P. R., "Green's-function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space," J. Chem. Phys. **123**(23), 234910 (2005).
[45]van Zon, J. S. and Ten Wolde, P. R., "Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics," Phys. Rev. Lett. **94**(12), 128103 (2005).

[46]Winkelmann, S. and Schütte, C., "The spatiotemporal master equation: Approximation of reaction-diffusion dynamics via markov state modeling," J. Chem. Phys. **145**(21), 214107 (2016).
[47]Winkelmann, S. and Schütte, C., "Hybrid models for chemical reaction networks: Multiscale theory and application to gene regulatory systems," J. Chem. Phys. **147**(11), 114115 (2017).
[48]Xing, J., Ginty, D. D., and Greenberg, M. E., "Coupling of the RAS-MAPK pathway to gene activation by RSK2, a growth factor-regulated CREB kinase," Science **273**(5277), 959–963 (1996).
[49]Yildirim, N. and Mackey, M. C., "Feedback regulation in the lactose operon: A mathematical modeling study and comparison with experimental data," Biophys. J. **84**(5), 2841–2851 (2003).
[50]Zhang, L. and Schaeffer, H., "On the convergence of the SINDy algorithm," preprint arXiv:1805.06445 (2018).
[51]Zhang, W. and Tu Liu, H., "MAPK signal pathways in the regulation of cell proliferation in mammalian cells," Cell Res. **12**(1), 9 (2002).
[52]Zou, H. and Hastie, T., "Regularization and variable selection via the elastic-net," J. R. Stat. Soc. **67**(2), 301–320 (2005).